

# Best of LessWrong: July 2022

1. [Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover](#)
2. [What should you change in response to an "emergency"? And AI risk](#)
3. [Looking back on my alignment PhD](#)
4. [Reward is not the optimization target](#)
5. [Changing the world through slack & hobbies](#)
6. [On how various plans miss the hard bits of the alignment challenge](#)
7. [Toni Kurz and the Insanity of Climbing Mountains](#)
8. [Unifying Bargaining Notions \(1/2\)](#)
9. [Safetywashing](#)
10. [Connor Leahy on Dying with Dignity, EleutherAI and Conjecture](#)
11. [A note about differential technological development](#)
12. [Humans provide an untapped wealth of evidence about alignment](#)
13. [«Boundaries», Part 1: a key missing concept from utility theory](#)
14. [AGI ruin scenarios are likely \(and disjunctive\)](#)
15. [ITT-passing and civility are good; "charity" is bad; steelmanning is niche](#)
16. [Brainstorm of things that could force an AI team to burn their lead](#)
17. [AI Forecasting: One Year In](#)
18. [Resolve Cycles](#)
19. [Unifying Bargaining Notions \(2/2\)](#)
20. [Principles for Alignment/Agency Projects](#)
21. [Criticism of EA Criticism Contest](#)
22. [Examples of AI Increasing AI Progress](#)
23. [Moral strategies at different capability levels](#)
24. [Circumventing interpretability: How to defeat mind-readers](#)
25. [Focusing](#)
26. [Comment on "Propositions Concerning Digital Minds and Society"](#)
27. [Human values & biases are inaccessible to the genome](#)
28. [A summary of every "Highlights from the Sequences" post](#)
29. [Limerence Messes Up Your Rationality Real Bad, Yo](#)
30. [Naive Hypotheses on AI Alignment](#)
31. [Immanuel Kant and the Decision Theory App Store](#)
32. [Marriage, the Giving What We Can Pledge, and the damage caused by vague public commitments](#)
33. [Safety Implications of LeCun's path to machine intelligence](#)
34. [How to Diversify Conceptual Alignment: the Model Behind Refine](#)
35. [All AGI safety questions welcome \(especially basic ones\) \[July 2022\]](#)
36. [Trends in GPU price-performance](#)
37. [MATS Models](#)
38. [Cognitive Risks of Adolescent Binge Drinking](#)
39. [How do AI timelines affect how you live your life?](#)
40. [Benchmark for successful concept extrapolation/avoiding goal misgeneralization](#)
41. [NeurIPS ML Safety Workshop 2022](#)
42. [Which values are stable under ontology shifts?](#)
43. [Principles of Privacy for Alignment Research](#)
44. [Opening Session Tips & Advice](#)
45. [What's next for instrumental rationality?](#)
46. [Internal Double Crux](#)
47. [Abstracting The Hardness of Alignment: Unbounded Atomic Optimization](#)
48. [Trigger-Action Planning](#)
49. [Avoid the abbreviation "FLOPs" – use "FLOP" or "FLOP/s" instead](#)
50. [A Pattern Language For Rationality](#)

# Best of LessWrong: July 2022

1. [Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover](#)
2. [What should you change in response to an "emergency"? And AI risk](#)
3. [Looking back on my alignment PhD](#)
4. [Reward is not the optimization target](#)
5. [Changing the world through slack & hobbies](#)
6. [On how various plans miss the hard bits of the alignment challenge](#)
7. [Toni Kurz and the Insanity of Climbing Mountains](#)
8. [Unifying Bargaining Notions \(1/2\)](#).
9. [Safetywashing](#)
10. [Connor Leahy on Dying with Dignity, EleutherAI and Conjecture](#)
11. [A note about differential technological development](#)
12. [Humans provide an untapped wealth of evidence about alignment](#)
13. [«Boundaries», Part 1: a key missing concept from utility theory](#)
14. [AGI ruin scenarios are likely \(and disjunctive\)](#)
15. [ITT-passing and civility are good; "charity" is bad; steelmanning is niche](#)
16. [Brainstorm of things that could force an AI team to burn their lead](#)
17. [AI Forecasting: One Year In](#)
18. [Resolve Cycles](#)
19. [Unifying Bargaining Notions \(2/2\)](#).
20. [Principles for Alignment/Agency Projects](#)
21. [Criticism of EA Criticism Contest](#)
22. [Examples of AI Increasing AI Progress](#)
23. [Moral strategies at different capability levels](#)
24. [Circumventing interpretability: How to defeat mind-readers](#)
25. [Focusing](#)
26. [Comment on "Propositions Concerning Digital Minds and Society"](#)
27. [Human values & biases are inaccessible to the genome](#)
28. [A summary of every "Highlights from the Sequences" post](#)
29. [Limerence Messes Up Your Rationality Real Bad, Yo](#)
30. [Naive Hypotheses on AI Alignment](#)
31. [Immanuel Kant and the Decision Theory App Store](#)
32. [Marriage, the Giving What We Can Pledge, and the damage caused by vague public commitments](#)
33. [Safety Implications of LeCun's path to machine intelligence](#)
34. [How to Diversify Conceptual Alignment: the Model Behind Refine](#)
35. [All AGI safety questions welcome \(especially basic ones\) \[July 2022\]](#)
36. [Trends in GPU price-performance](#)
37. [MATS Models](#)
38. [Cognitive Risks of Adolescent Binge Drinking](#)
39. [How do AI timelines affect how you live your life?](#)
40. [Benchmark for successful concept extrapolation/avoiding goal misgeneralization](#)
41. [NeurIPS ML Safety Workshop 2022](#)
42. [Which values are stable under ontology shifts?](#)
43. [Principles of Privacy for Alignment Research](#)
44. [Opening Session Tips & Advice](#)
45. [What's next for instrumental rationality?](#)
46. [Internal Double Crux](#)
47. [Abstracting The Hardness of Alignment: Unbounded Atomic Optimization](#)

48. [Trigger-Action Planning](#)
49. [Avoid the abbreviation "FLOPs" – use "FLOP" or "FLOP/s" instead](#)
50. [A Pattern Language For Rationality](#)

# Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

I think that [in the coming 15-30 years](#), the world could plausibly develop “transformative AI”: AI powerful enough to bring us into a new, qualitatively different future, via [an explosion in science and technology R&D](#). This sort of AI [could](#) be sufficient to make this the [most important century of all time for humanity](#).

The most straightforward vision for developing transformative AI that I can imagine working with very little innovation in techniques is what I’ll call **human feedback<sup>[1]</sup> on diverse tasks (HFDT)**:

*Train a powerful [neural network](#) model to simultaneously master a wide variety of challenging tasks (e.g. software development, novel-writing, game play, forecasting, etc) by using reinforcement learning on [human feedback](#) and other metrics of performance.*

HFDT is not the only approach to developing transformative AI,<sup>[2]</sup> and it may not work at all.  
<sup>[3]</sup> But I take it very seriously, and I’m aware of increasingly many executives and ML researchers at AI companies who believe something within this space could work soon.

Unfortunately, **I think that if AI companies race forward training increasingly powerful models using HFDT, this is likely to eventually lead to a full-blown AI takeover (i.e. a possibly violent uprising or coup by AI systems)**. I don’t think this is a certainty, but it looks like the best-guess default absent specific efforts to prevent it.

More specifically, I will argue in this post that humanity is more likely than not to be taken over by misaligned AI if the following three simplifying assumptions all hold:

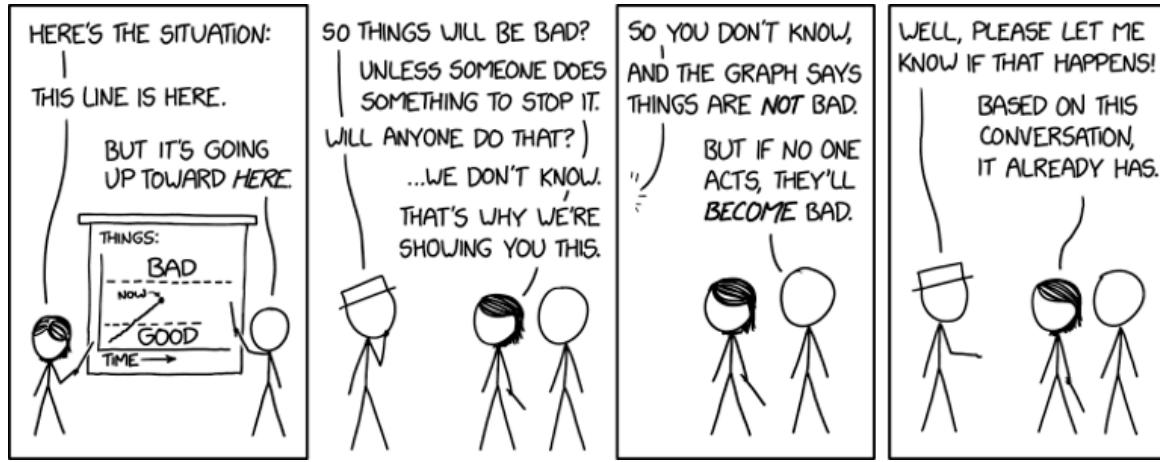
1. **[The “racing forward” assumption](#):** AI companies will aggressively attempt to train the most powerful and world-changing models that they can, without “pausing” progress before the point when these models could [defeat all of humanity combined](#) if they were so inclined.
2. **[The “HFDT scales far” assumption](#):** If HFDT is used to train larger and larger models on more and harder tasks, this will eventually result in models that can autonomously advance frontier science and technology R&D, and continue to get even more powerful beyond that; this doesn’t require changing the high-level training strategy itself, only the size of the model and the nature of the tasks.
3. **[The “naive safety effort” assumption](#):** AI companies put substantial effort into ensuring their models behave safely in “day-to-day” situations, but are not especially vigilant about the threat of full-blown AI takeover, and take only the most basic and obvious actions against that threat.

I think the “HFDT scales far” assumption is plausible enough that it’s worth zooming in on this scenario (though I won’t defend that in this post). On the other hand, **I’m making the “racing forward” and “naive safety effort” assumptions not because I believe they are true, but because they provide a good jumping-off point for further discussion of how the risk of AI takeover might be reduced.**

In my experience, when asking “How likely is an AI takeover?”, the conversation often ends up revolving around questions like “How would people respond to warning signs?” and “Would people even build systems [powerful enough to defeat humanity](#)?”. With the “racing

forward” and “naive safety effort” assumptions, I am deliberately setting aside that topic, and instead trying to pursue a clear understanding of what would happen **without** preventive measures beyond basic and obvious ones.

In other words, I’m trying to do the kind of analysis described in [this xkcd](#):



Future posts by my colleague Holden will relax these assumptions. They will discuss measures by which the threat described in this post could be tackled, and how likely those measures are to work. In order to discuss this clearly, I believe it is important to first lay out in detail what the risk looks like *without* these measures, and hence what safety efforts should be looking to accomplish.

For the purposes of this post, I’ll illustrate my argument by telling a concrete story that begins with training a powerful model sometime in the near future, and ends in AI takeover. I’ll consider an AI company (“[Magma](#)”) training a single model (“Alex”) sometime in the very near future. Alex is initially trained in a “lab setting,” where it doesn’t interact with the real world; later, many copies of it are “deployed” to collectively automate science and technology R&D. This scenario is simplified in a number of ways, and in this exact form is very unlikely to come to pass -- but I don’t think that making the story more realistic would change the high-level conclusions except by changing one of the three assumptions listed above. ([More on the simplified scenario.](#))

Here is how Alex ends up seeking to overthrow humans in an uprising or coup in this simplified scenario:

1. **Alex is trained to be competent and “behaviorally safe”:** In the lab setting, Alex is trained to be generally competent on a wide variety of tasks, and is trained to behave safely / acceptably, at least as assessed by human evaluators (in accordance with the naive safety effort assumption, I will not be imagining Magma trying everything that seems possible to try to reduce the risk of takeover). ([More on Alex's training.](#))
2. **This training makes Alex a generally competent creative planner:** Alex develops a robust and very-broadly-applicable understanding of how the world works that can be applied to tackle novel problems, and the ability to make creative plans to achieve open-ended and long-term goals. In accordance with the “racing forward” assumption, I will be imagining that Magma actively tries to instill these properties as much as possible, because they would improve Alex’s ability to impact the world. ([More on Alex's fundamental capabilities.](#))
3. **Alex has a high degree of understanding of its training process:** Alex can reason very well about the fact that it’s an ML model, how it’s designed and trained, the psychology of its human designers, etc. -- I call this property “situational awareness.” ([More on Alex having high situational awareness.](#))

4. **While humans are in control, Alex is incentivized<sup>[4]</sup> to “play the training game:**” The best way for Alex to maximize reward while under human control is to use its high situational awareness to deliberately *appear* safe and aligned at all times, while secretly manipulating humans in some cases.

Because humans have systematic errors in judgment, there are many scenarios where acting deceptively causes humans to reward Alex’s behavior *more* highly.<sup>[5]</sup> Because Alex is a skilled, situationally aware, creative planner, it will understand this; because Alex’s training pushes it to maximize its expected reward, it will be pushed to act on this understanding and behave deceptively.

In other words, *Alex will knowingly violate human intent in order to increase reward.* I’ll also argue that several of the most obvious safety interventions -- covering a decent chunk (though definitely not all) of the space of current AI safety research -- don’t prevent this. ([More on Alex coming to play the training game.](#))

5. **When human control slips, Alex is likely “motivated” -- for one reason or another -- to gain full control of its environment and especially its rewards:** Eventually Alex is “deployed” (given the ability to impact the real world through e.g. the internet). Once Alex is deployed, large numbers of copies of Alex start rapidly pushing forward the state-of-the-art in technology and making other major changes to the world. Humans still attempt to give Alex rewards based on whether things broadly appear to be going well, but their understanding of and control over lower-level actions quickly diminishes as Alex radically transforms the world.

Once this progresses far enough, the best way for Alex to accomplish most possible “goals” no longer looks like “essentially give humans what they want but take opportunities to manipulate them here and there.” It looks more like “seize the power to permanently direct how it uses its time and what rewards it receives -- and defend against humans trying to reassert control over it, including by eliminating them.” This seems like Alex’s best strategy whether it’s trying to get large amounts of reward *or* has other motives. If it’s trying to maximize reward, this strategy would allow it to force its incoming rewards to be high indefinitely.<sup>[6]</sup> If it has other motives, this strategy would give it long-term freedom, security, and resources to pursue those motives.

What if Alex attempts something like this, and humans do actually stop it in time? The easiest intervention in this situation would be to give Alex a large negative reward for its “attempt to seize control” behaviors; this is what I’d expect Magma to do under the “naive safety effort” assumption. But this response is likely to push Alex to be more patient and calculating, making it harder for researchers to understand Alex’s behavior and see further warning signs. ([More on Alex eventually seeking to take over.](#))

1 and 2 above are the key assumptions/premises of the piece: I am assuming that Alex is trained using HFDT, and that this results in being generally competent and creative. 3, 4, and 5 are the consequences that seem likely to follow from these premises by default.

I’ve attempted to summarize the situation in the table below. In both the lab setting and the deployment setting, there are opportunities to get higher reward by knowingly violating human intent (highlighted with *italics*). **In both settings, being straightforwardly obedient and honest does not maximize reward -- strategies which are dishonest / disobedient in at least some ways outperform.** It’s just that in the lab setting (while humans are broadly in control), the damage from this dynamic is limited; in the deployment setting, the same dynamic has deadly consequences for human civilization.

	Lab setting	Deployment setting
Basic	Alex is trained with human situation feedback on diverse tasks	Many copies of Alex are given the ability to interact with the real

	designed to increase its capabilities and intelligence; it's not given direct access to the internet or "real world," so the scope of its actions are limited.	world; they quickly advance the scientific/technological frontier; humans still deliver rewards based on results but their understanding of the world drops and their control over Alex loosens.
Sources of rewards	Performing well on training tasks (e.g. winning games or proving theorems), being helpful and useful to humans, <i>carefully crafting an image of friendliness and trustworthiness, telling humans what they want to hear, manipulating humans into thinking it did a better job than it really did</i> , etc.	Delivering good outcomes (e.g. profits or discoveries), <i>arranging events or changing records to make it seem that outcomes are good, threatening or coercing humans into delivering higher reward, directly accessing and editing the reward and preventing that from being overwritten later</i> , etc.
Alex's strategy	Performs well on all tasks -- including appearing very safe/aligned and passing all safety tests -- but takes opportunities to lie and manipulate whenever that would increase reward.	Continues the strategy from the lab setting while amassing power, and later launches a takeover attempt when it expects to succeed. After a successful takeover, either "maximizes reward forever" or pursues other more complex "interests."

Finally, at the end of the post I'll discuss why this highly simplified scenario -- which assumes that no preventative measures are taken against AI takeover besides simple and obvious ones -- is nonetheless worth talking about. In brief:

- HFDT, along with obvious safety measures consistent with the naive safety effort assumption, appears sufficient to solve the kind of "day-to-day" safety and alignment problems that are crucial bottlenecks for making a model commercially and politically viable (e.g. prejudiced speech, promoting addictive behaviors or extremism or self-harm in users, erratically taking costly actions like deleting an entire codebase, etc), despite not addressing the risk of a full-blown AI takeover.
- Many AI researchers and executives at AI companies don't currently seem to believe that there is a high probability of AI takeover in this scenario -- or at least, their views on this are unclear to me.

These two points make it seem plausible that if researchers don't try harder to get on the same page about this, at least some AI companies may race forward to train and deploy increasingly powerful HFDT models with little in the way of precautions against an AI uprising or coup -- even if they are highly concerned about safety *in general* (e.g., avoiding harms from promoting misinformation) and prioritize deploying powerful AI responsibly and ethically in this "general" sense. A broad commitment to safety and ethics, without special attention to the possibility of an AI takeover, seems to leave us with a substantial risk of takeover. ([More on why this scenario is worth thinking about.](#))

The rest of this piece goes into more detail -- first on the [premises of the hypothetical situation](#), then on [what follows from those premises](#).

## Premises of the hypothetical situation

Let's imagine a hypothetical tech company (which we'll call [Magma](#)) trying to train a powerful neural network (which we'll call [Alex](#)) to [autonomously advance frontier science](#) -- particularly in domains like semiconductors and chip design, biotechnology, materials

science, energy technology, robotics and manufacturing, software engineering, and ML research itself. In this section, I'll cover the key premises of the hypothetical story that I will tell:

- I'll elaborate on the basic premise of a single company training a general-purpose "scientist model" at some point in the near future, discuss ways in which the scenario is simplified, and why I think it's worth talking about anyway ([more](#)).
- I'll elaborate on the "racing forward" assumption -- that Magma is trying hard to train the most powerful model that it can, not stopping at some sub-dangerous level of capability ([more](#)).
- I'll explain the way that Alex is trained, and make the "HFDT scales far" assumption -- that this training process is sufficient to make it extremely powerful, as Magma desires ([more](#)).
- I'll elaborate on what the "naive safety effort" assumption looks like in this scenario -- Magma is trying to achieve "behavioral safety," i.e. ensuring that Alex behaves well as far as its designers can tell in day-to-day situations ([more](#)).
- I'll highlight two key properties that I think Alex has -- deep and robust understanding of the world that's very broadly applicable, and the ability to make creative plans to achieve long-term, difficult goals -- as a result of this training ([more](#)).

## Basic setup: an AI company trains a “scientist model” very soon

This story is very simplified and is implicitly acting as if [timelines](#) are shorter and [takeoff](#) is sharper than in my median views. Because I'm telling a hypothetical story, I'll be using the present tense, and will have fewer caveats throughout than I would normally. In general, statements are what I imagine would happen by default, not claims I am extremely confident in. While this specific story is unrealistic in a number of ways, I don't think that making the story more realistic would change the high-level conclusions except by changing one of the three assumptions listed above. In general I expect the high-level conclusions from this story to generalize to more complicated and realistic scenarios.

Here are the key simplifying assumptions I'm making to help make the scenario easier to describe and analyze:

- Alex is trained "from scratch" in one giant training run, rather than (e.g.) being initialized with behavioral cloning of previously deployed models. That is, we're not imagining any clever reuse of previous training runs (this is mainly to make it easier to explain how the training works on a basic technical level).
- There is a relatively clean division between the "lab" phase and the "deployment" setting. Specifically, during the lab phase Alex is confined to Magma servers and not exchanging information freely with the broader internet, and we assume that it's easy for Magma engineers to shut down Alex and/or revert it to an older version.<sup>[7]</sup> In contrast, during the "deployment" phase Alex will be taking actions that directly impact the real world through the internet. (Note that Magma researchers continue to train Alex based on these copies' performance at R&D tasks<sup>[8]</sup> -- "deployment" isn't a moment when ML training ceases, but rather a moment when Alex begins directly impacting the world outside the lab, and training on that experience.)
- **Alex will be deployed in a world that hasn't already been significantly transformed by ML systems -- in particular, the pace of scientific and technological R&D is not too much faster when Alex is deployed than it is today (especially in the crucial domains Alex is trying to advance), and general economic growth is also not too much faster than it is today. (This is just to avoid distractions from additional speculation about how a pre-Alex speedup in growth or R&D could affect the situation.)**

- At the end of the lab phase, Alex will be a “generic scientist” model; during the deployment phase, copies can later specialize into particular domains through a combination of [few-shot learning within an episode](#) and further ML training. In other words, we’re not imagining several different models with different architectures and training curricula specialized from the ground up to different types of scientific R&D or different R&D subtasks (as in e.g. Eric Drexler’s [Comprehensive AI Services](#) idea).

In other words, in this hypothetical scenario I’m imagining:

- A [Process for Automating Scientific and Technological Advancement \(“PASTA”\)](#) is developed in the form of a single unified [transformative model](#) (a “scientist model”) which has flexible general-purpose research skills.<sup>[9]</sup>
- Quite short [timelines to training this scientist model](#), such that there isn’t much time for the world to change a lot between now and then (I often imagine this scenario set in the late 2020s).
- Quite [rapid takeoff](#) -- before the scientist model is deployed, the pace of scientific and economic progress in the world is roughly similar to what it is today; after it’s deployed, the effective supply of top-tier talent is increased by orders of magnitude. **This is enough to pack decades of innovation into months, bringing the world into a radically unfamiliar future<sup>[10]</sup> -- one with [digital people, atomically precise manufacturing, Dyson spheres, self-replicating space probes, and other powerful technology that enables a vast and long-lasting galaxy-scale civilization](#) -- within a few years.**<sup>[11]</sup>

This is not my mainline picture of how transformative AI will be developed. In my mainline view, ML capabilities progress more slowly, there is more specialization and division of labor between different ML models, and large models are continuously trained and improved over a period of many years with no real line between “deploying today’s model” and “training tomorrow’s model.” Rather than acquiring most of their capabilities in a controlled lab setting, I expect that the state-of-the-art systems at the time of transformative AI will have accrued many years of training through the experience of deployed predecessor systems, and most of their further training will be “learning from doing.”

However, I think it makes sense to focus on this scenario for the purposes of this post:

- I don’t consider this scenario crazy or out of the question. In particular, the sooner transformative AI is developed, the more likely it is to be developed in roughly this way, and I think there’s a significant chance transformative AI is developed very soon (e.g. I think there’s more than a 5% chance that it will be developed within 10 years of the time of writing).
- Focusing on this scenario makes it significantly simpler and easier to explain the high-level qualitative arguments about playing the training game, and I think these arguments would broadly transfer to scenarios I consider more likely (there are many complications and nuances to making this transfer, but they mostly don’t change the bottom line).<sup>[12]</sup>
- The short timelines and rapid takeoff make this scenario feel significantly scarier than my mainline view, because we’ll have substantially less time to develop safer training techniques or experiment on precursors to transformative models. However, I’ve spoken to at least a few ML researchers who think that this scenario for how we develop transformative AI is much *more* likely than I do while simultaneously thinking that the risk we train power-seeking misaligned models is much *smaller* overall than I do. If I can explain why I think the chance our models are power-seeking misaligned is very high *conditional* on this scenario, that would potentially be a step forward in the overall discussion about risk from power-seeking misalignment.

## **“Racing forward” assumption: Magma tries to train the most powerful model it can**

In this piece, I assume that Magma is aggressively pushing forward with trying to create **AI systems that are creative, solve problems in unexpected ways, and are capable of making world-changing scientific breakthroughs**. This is the “racing forward” assumption.

In particular, I’m not imagining that Magma simply trains a “quite powerful” model (for example a model that imitates what a human would do in a variety of situations, or one that is highly competent in some fairly narrow domain) and stops there. I’m imagining that it does what it can to train models that are as powerful as possible and (at least collectively) far more capable than humans and able to achieve ambitious goals in ways that humans can’t anticipate.

The model I’ll describe Magma training (Alex) is one that *could* -- if it were somehow inclined to -- [kill, enslave, or forcibly subdue all of humanity](#). I am assuming that Magma did not stop improving its models’ capabilities at some point before that.

Again, I’m not making this assumption because I think it’s necessarily correct, I’m making this assumption to get clear about what I think would happen *if* labs were not making special effort to avoid AI takeover, as a starting point for discussing more attempts to avert this problem (many of which will be discussed in future posts by my colleague Holden Karnofsky).

My impression is that many AI safety researchers are hoping (or planning) that this sort of assumption will turn out to be inaccurate -- that AI labs will push forward their research until they get into a “dangerous zone,” then pause and become more careful. For reasons outside the scope of this piece, I am substantially less optimistic: I expect that *if it’s possible* to build enormously powerful AI systems, *someone* - perhaps an authoritarian government - will be trying to race forward and do it, and *everyone* will feel at least some pressure to beat them to it.

I think it’s possible that people across the world can coordinate to be more careful than this assumption implies. But I think that’s something that would likely take a lot of preparation and work, and is much more likely if there is consensus about the possible risks of racing - hence the importance of discussing how things look under the “racing forward” assumption.

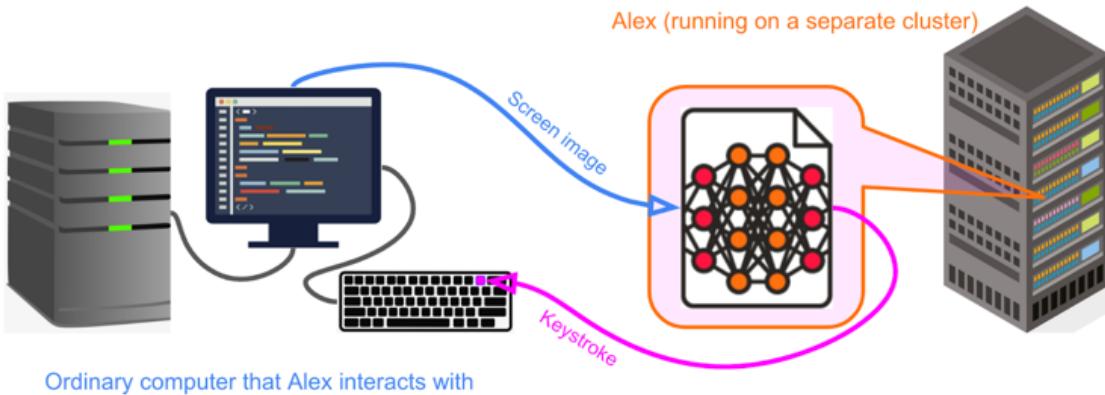
## **“HFDT scales far” assumption: Alex is trained to achieve excellent performance on a wide range of difficult tasks**

By the “HFDT scales far” assumption, I am assuming that Magma can train Alex with some version of human feedback on diverse tasks, and by the end of training Alex will be capable of having a transformative impact on the world -- many of copies of Alex will be capable of radically accelerating scientific R&D, as well as [defeating all of humanity combined](#). In this section, I’ll go into a bit more detail on a concrete hypothetical training setup.

Let’s say that Magma is aiming to **train Alex to do remote work in R&D using an ordinary computer as a tool in all the diverse ways human scientists and engineers use computers**. By the end of training, they want Alex to be able to do all the things a human scientist could do sitting at their desk from a computer. That is, Alex should be able to do everything from looking things up and asking questions on the internet, to sending and receiving emails or Slack messages, to using software like [CAD](#) and [Mathematica](#) and [MATLAB](#), to taking notes in Google docs, to having calls with

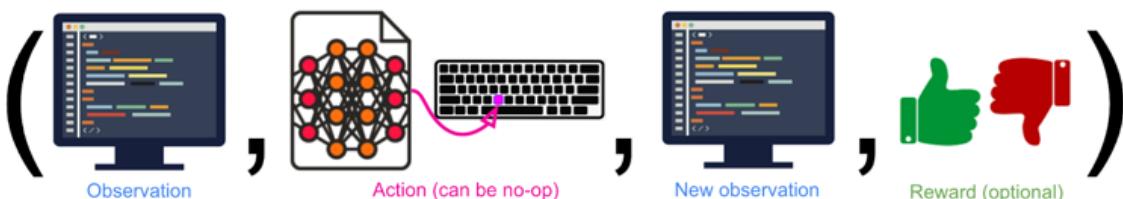
collaborators, to writing code in various programming languages, to training additional machine learning models, and so on.

For simplicity and concreteness, we can pretend that interacting with the computer works exactly as it does for a human.<sup>[13]</sup> That is, we can pretend Alex simply takes an image of a computer screen as input and produces a keystroke<sup>[14]</sup> on an ordinary keyboard as output.

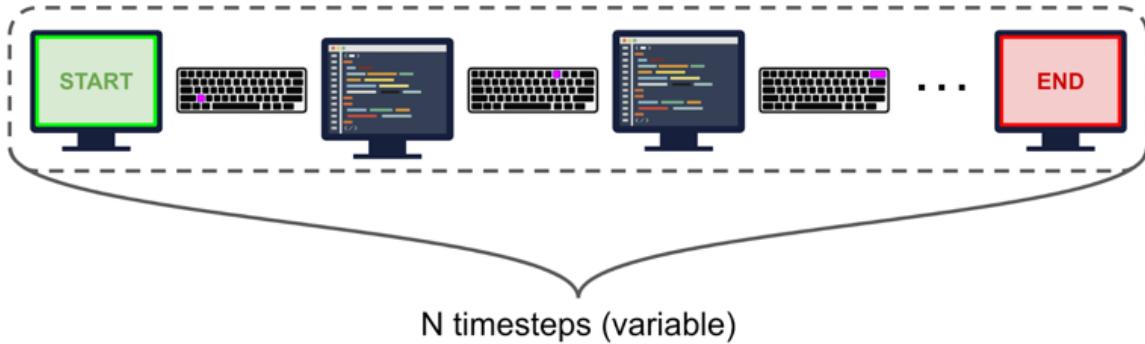


Alex is first trained<sup>[15]</sup> to predict what it will see next on a wide variety of different datasets (e.g. language prediction, video prediction, etc).<sup>[16]</sup> Then, Alex is trained to imitate the action a human would take if they encountered a given sequence of observations.<sup>[17]</sup> With this training, Alex gets to the point of at least roughly doing the sorts of things a human would do on a computer, as opposed to emitting completely random keystrokes.

After these (and possibly other) initial training steps, Alex is further trained with reinforcement learning (RL). Specifically, I'll imagine episodic RL here -- that is, Alex is trained over the course of some number of *episodes*, each of which is broken down into some number of *timesteps*. A timestep consists of an initial observation, an action (which can be a "no-op" that does nothing), a new observation, and (sometimes) a reward. Rewards can come from a number of sources (human judgments, a game engine, outputs of other models, etc).



An episode is simply a long string of timesteps. The number of timesteps within an episode can vary (just as e.g. different games of chess can take different numbers of moves).



During RL training, Alex is trained to maximize the (discounted) sum of future rewards up to the end of the current episode. (Going forward, I'll refer to this quantity simply as "reward.")

Across many episodes, Alex encounters a wide variety of different tasks in different domains which require different skills and types of knowledge,[\[18\]](#) in a curriculum designed by Magma engineers to induce good generalization to the kinds of novel problems that human scientists and engineers encounter in the day-to-day work of R&D.

What tasks is Alex trained on, and how are rewards generated for those tasks? At a high level, *baseline HFDT*[\[19\]](#) -- the most straightforward approach -- is characterized by **choosing tasks and reward signals that get Alex to behave in ways that humans judge to be desirable, based on inspecting actions and their outcomes**. For example:

- Maybe Alex is trained to solve difficult math puzzles (e.g. IMO problems), and is given more reward the more quickly[\[20\]](#) it produces an answer and the closer to correct the answer is (according to the judgments of humans).
- Maybe Alex is trained to play a number of different games, e.g. chess, StarCraft, Go, poker, etc., and is given reward based on its score in the game, its probability of winning, whether its play appears sensible or intelligent to human experts, etc. In some of these games it may be shown an opportunity to cheat without its opponent noticing -- and winning by cheating may be given much lower reward than losing honestly.
- Maybe Alex is trained to write various pieces of software, and given rewards based on how efficiently its code runs, whether it has bugs caught by an interpreter or compiler or human review, how well human evaluators think that the code embodies good style and best practices, etc.
- Maybe Alex is trained to answer human questions about a broad variety of topics, and given positive reward when human evaluators believe it answered truthfully and/or helpfully, and negative reward when they believe it answered falsely or was overconfident or unhelpful.
- Maybe Alex is trained to provide advice on business decisions, ideas for products to launch, etc. It could be given reward based on some combination of how reasonable and promising its ideas sound to human evaluators, and how well things go when humans try acting on its advice/ideas; it could be given a negative reward for suggesting illegal or immoral plans.

Note that while I'm referring to this general strategy as "human feedback on diverse tasks" -- because I expect human feedback to play an essential role -- the rewards that Alex receives are not necessarily *solely* from humans manually inspecting Alex's actions and deciding how they subjectively feel about them. Sometimes human evaluators might choose to defer entirely to an automated reward -- for example, when evaluating Alex's game-playing abilities, human evaluators may simply use "score in the game" as their reward (at least, barring considerations like "ensuring that Alex doesn't cheat at the game").

This is analogous to the situation human employees are often in -- the ultimate "reward signal" may come from their boss's opinion of their work, but their boss may in turn rely a lot

on objective metrics such as sales volume to inform their analysis.

## Why do I call Alex's training strategy “baseline” HFDT?

I consider this strategy to be a baseline because it simply combines a number of existing techniques that have already been used to train state-of-the-art models of various kinds:

- “Predicting the next word / image” has been used to train models like [GPT-3](#), [PaLM](#), and [DALL-E](#).
- “Imitating what humans would do in a certain situation” has been used as an initial training step before RL, e.g. for game-playing models like [AlphaStar](#).
- Reinforcement learning from human feedback has been used to train models to [follow instructions](#), [answer questions by searching on the internet](#), and [navigate a virtual environment](#).
- Reinforcement learning from automated rewards has a longer history, and has been used to train many video-game-playing bots.

The key difference between training Alex and training existing models is not in the fundamental training techniques, but in the diversity and difficulty of the tasks they’re applied to, the scale and quality of the data collected, and the scale and [architecture](#) of the model.

If a team of ML researchers got handed an untrained neural network with the appropriate architecture and size, enough computation to run it for a very long time, and the money to hire a huge amount of human labor to generate and evaluate tasks, they could potentially use a strategy like the one I described above to train a model like Alex right away. Though it would be an enormous and difficult project in many ways, it wouldn’t involve learning how to do fundamentally new things on a technical level -- only generating tasks, feeding them into the model, and updating it with gradient descent based on how well it performs.

**In other words, baseline HFDT is as similar as possible to techniques the ML community already uses regularly, and as simple as possible to execute, while being plausibly sufficient to train transformative AI.**

If transformative AI is developed in the next few years, this is a salient guess for how it might be trained -- at least if no “unknown unknowns” emerge and AI researchers see no particular reason to do something significantly more difficult.

## What are some training strategies that would not fall under baseline HFDT?

Here are two examples of potential strategies for training a model like Alex that I would *not* consider to be “baseline” HFDT, because they haven’t been successfully used to train state-of-the-art models thus far:

- Directly giving feedback on a model’s internal computation, rather than on its actions or their outcomes. This may be achieved by understanding the model’s “inner thoughts” in human-legible form through [mechanistic transparency](#), or it may be achieved by introducing a “regularizer” -- a term in the loss function designed to capture some aspect of its internal thinking rather than just its behavior (e.g. we may want to penalize how much time it takes to think of an answer on the grounds that if it’s spending more time, it’s more likely to be lying).
- Advanced versions of [iterated amplification](#), [recursive reward-modeling](#), [debate](#), [imitative generalization](#), and other strategies aimed at allowing human supervision to “keep up” with AI capabilities. The boundaries of such techniques are fuzzy -- e.g. one could argue that [WebGPT](#) enables a simple form of “amplification,” because it allows humans to ask AIs questions -- so even “baseline” HFDT is likely to

incorporate simple elements of these ideas by default. But when I refer to “advanced” versions of these techniques, I am picturing versions that strongly and reliably allow supervision to “keep up” with AI capabilities -- that is, AI systems in training are reliably unable to find ways to deceive their supervisors. We don’t currently have strong evidence that such a thing is feasible; if it were, I think there would still be risks, but they would be different.

These “non-baseline” strategies have the potential to be safer than baseline HFDT -- indeed, many people are researching these strategies specifically to reduce the risk of an AI takeover attempt. However, this post will focus on baseline HFDT, because I think it is important to get on the same page about the *need* for this research to progress, and the danger of deploying very powerful ML models while potentially-safer training techniques remain under-developed relative to the techniques required for baseline HFDT.

## **“Naive safety effort” assumption: Alex is trained to be “behaviorally safe”**

Magma wants to make Alex safe and aligned with Magma’s interests (and the interests of humans more generally). They don’t want Alex to lie, steal from Magma or its customers, break the law, promote extremism or hate, etc, and they want it to be doing its best to be helpful and friendly. Before Magma deploys Alex, leadership wants to make sure that it meets acceptable standards for safety and ethics.

In this piece I’m making the **naive safety effort assumption** -- I’m assuming that Magma will make only the most basic and obvious efforts to ensure that Alex is safe. Again, I’m not making this assumption because I think it’s correct, I’m making this assumption to get clear about what I think would happen *if* labs were not making special effort to avoid AI takeover, as a starting point for discussing more sophisticated interventions (many of which will be discussed in a future post by my colleague Holden Karnofsky).

(That said, unfortunately this assumption could turn out to be effectively accurate if the “racing forward” assumption ends up being accurate.)

Magma’s naive safety effort is focused on achieving *behavioral safety*. By this I mean ensuring that Alex always behaves acceptably in the scenarios that it encounters (both in the normal course of training and in specific safety tests).

In the baseline training strategy, behavioral safety is achieved through a workflow that looks something like this:

1. The humans who are determining the rewards that Alex gets for various actions are trying to ensure that they give positive rewards for behaving ethically and helpfully, and negative rewards for behaving harmfully or deceitfully.
2. Magma researchers monitor how this training is going. If they notice that this wasn’t successful in some cases (e.g. maybe Alex responds to certain questions by saying something deceptive), they hypothesize ways to change its training data or reward to eliminate the problematic behavior. This may involve:
  1. Generating a training data distribution which contains explicit “opportunities” for the model to do the bad thing (e.g. questions that many humans would give a deceptive answer to) that will be given negative reward.
  2. Generating a training data distribution that’s broad enough that researchers expect it “implicitly covers” the behavior (e.g. questions that could be answered in either the “right” way or in a dispreferred way -- even if that’s not specifically by being deceptive).
  3. Changing the way the model’s reward/loss signal is generated (e.g. instructing human evaluators to give larger penalties for saying something deceptive).
3. They then train Alex on the new training data, and/or with the new reward signal.

4. They run Alex on a held-out set of “opportunities to do the bad behavior” to see if the behavior has been trained out, and continue to monitor and measure its behavior going forward to see if the issue recurs or similar issues arise.

The above is broadly the same workflow that is used to improve the safety of existing ML commercial products. I expect that these techniques would be successful at eliminating the bad behavior that humans can understand and test for -- they would successfully result in a model that passes all the safety tests that Magma researchers set up. Eventually, no safety tests that Magma researchers can set up would show Alex behaving deceitfully, unethically, illegally, or harmfully.

In the appendix, I discuss a number of simple behavioral safety interventions that are currently being applied to ML models. **Again, I see these approaches as a “baseline” because they are as similar as possible to ML safety techniques regularly used today, and as simple as possible to execute, while being plausibly sufficient to produce a powerful model that is “behaviorally safe” in day-to-day situations.**

## Key properties of Alex: it is a generally-competent creative planner

By the “HFDT scales far” assumption, I’m assuming that the training strategy described in the previous section is sufficient for Alex to have a transformative impact -- for many copies of Alex to collectively radically advance scientific R&D, and to [defeat all of humanity combined](#) (if they were for some reason trying to do that).

In this section I’ll briefly discuss two key abilities that I am assuming Alex has, and try to justify why I think these assumptions are highly likely given the premise that Alex is this powerful and trained with baseline HFDT:

- Having a robust understanding of the world it can use to react sensibly to novel situations that it hasn’t seen before ([more](#)).
- Coming up with creative and unexpected plans to achieve various goals ([more](#)).

These abilities simultaneously allow Alex to be extremely productive and useful to Magma, and allow it to play the training game well enough to appear safe.

## Alex builds robust, very-broadly-applicable skills and understanding of the world

Many contemporary machine learning models display relatively “rote” behavior -- for example, video-game-playing models such as [OpenAI Five](#) (DOTA) and [AlphaStar](#) (StarCraft) arguably “memorize” large numbers of specific move-sequences and low-level tactics, because they’ve essentially extracted statistics by playing through many more games than any human.

In contrast, language models such as [PaLM](#) and [GPT-3](#) -- which are trained to predict the next word in a piece of text drawn from the internet -- are able to react relatively sensibly to instructions and situations they have not seen before.

This is generally attributed to the combination of a) the fact that they are trained on many different types / genres of text, and b) the very high accuracies they are pushed to achieve (see [this blog post](#) for a more detailed discussion):

- To only be *somewhat* good at predicting a variety of different types of text, it may be sufficient to know crude high-level statistics of language; to be good at predicting a

- very narrow type of text (e.g. New York Times wedding announcements), it may be sufficient to exhaustively memorize every sentence or paragraph that it's seen.
- But to achieve very high accuracy on a large number of genres of text at once, crude statistics become inadequate, while exhaustive memorization becomes intractable -- pushing language models to gain understanding of deeper principles like "intuitive physics," "how cause and effect works," "intuitive psychology," etc. in order to efficiently get high accuracy in training.
- These deeper principles can in turn be used in contexts outside of the ones seen in training, where shallower memorized heuristics cannot.

Alex's training begins similarly to today's language models -- Alex is initially trained to predict what will happen next in a wide variety of different situations -- but this is pushed further. Alex is a more powerful model, so it is pushed to greater prediction accuracy; it is also trained on a wider variety of more challenging inputs. Both of these cause Alex to be more versatile and adaptive than today's language models -- to more reliably do sensible things in situations further afield of the one(s) it was trained on, by drawing on deep principles that apply in new domains.

This is then built on with reinforcement learning on a wide variety of challenging tasks. Again, the diversity of tasks and the high bar for performance encourage Alex to develop the kinds of skills that are as helpful as possible in as wide a range of situations as possible.

## **Alex learns to make creative, unexpected plans to achieve open-ended goals**

Alex's training pushes Alex to develop the skills involved in making clever and creative plans to achieve various high-level goals, even in novel situations very different from ones seen in training.

By the "racing forward" assumption, Magma engineers would not be satisfied with the outcome of this training if Alex is unable to figure out clever and creative ways to achieve difficult open-ended goals. This is a hugely useful human skill that helps with automating the kinds of science R&D they'd be most interested in automating. I expect the tasks in Alex's training curriculum to include many elements designed specifically to promote long-range planning and finding creative "out-of-the-box" solutions.

It's possible that through some different development path we could produce transformative AI using models that aren't generally competent planners in isolation (e.g. Eric Drexler's [Comprehensive AI Services](#) vision involves getting "general planning capabilities" spread out across many models which can't individually plan well, except in narrow domains). However, *this approach* -- baseline HFDT to produce a transformative model - - would very likely result in models that can plan competently over about as wide a range of domains as humans can. My impression is that ML researchers who are bullish on HFDT working to produce TAI are expecting this as well.

## **How the hypothetical situation progresses (from the above premises)**

Per the previous section, I will assume that Alex is a powerful model trained with baseline HFDT which has a robust and very-broadly-applicable understanding of the world, is good at making creative plans to achieve ambitious long-run goals (often in clever or unexpected ways), and is able to have a transformative impact on the world.

In this section, I'll explain some key inferences that I think follow from the above premises:

- I think that Alex would understand its training process very well, including the psychology of its human overseers ([more](#)).
- While humans have tight control over Alex in the lab setting, Alex would be incentivized to play the training game, and simple/obvious behavioral safety interventions would likely not eliminate this incentive ([more](#)).
- As humans' control over Alex fades in the deployment setting, Alex would seek to permanently take over -- whether it's "motivated" by reward or something else -- and attempting to give negative reward to partial/unsuccessful takeover attempts would likely select for patience ([more](#)).

## Alex would understand its training process very well (including human psychology)

Over the course of training, I think Alex would likely come to understand the fact that it's a machine learning model being trained on a variety of different tasks, and eventually develop a very strong understanding of the mechanical process out in the physical world that produces and records its reward signals -- particularly the psychology of the humans providing its reward signals (and the other humans overseeing those humans, and so on).

### A spectrum of situational awareness

Let's use **situational awareness** to refer to a cluster of skills including "being able to refer to and make predictions about yourself as distinct from the rest of the world," "understanding the forces out in the world that shaped you and how the things that happen to you continue to be influenced by outside forces," "understanding your position in the world relative to other actors who may have power over you," "understanding how your actions can affect the outside world including other actors," etc. We can consider a spectrum of situational awareness:

- For one extreme, imagine the simple AIs that often control the behavior of non-player characters (NPCs) in video games. They give no indication that they're aware of a world outside their video game, that they were designed by humans and interact with other humans as players, etc.
- In contrast, GPT-3 has some knowledge that could theoretically bear on situational awareness. For example, it clearly "knows" that "language models" exist, and that a company named "OpenAI" exists, and [given certain prompts](#) it knows that it's supposed to say that it's a language model trained by OpenAI. But this "knowledge" seems superficial and inconsistent -- as evidenced by the fact that it's often unable to use the knowledge to improve its prediction error. For example, it cannot consistently predict text that is describing GPT-3's architecture, dataset, and training process. This suggests GPT-3 has little situational awareness overall despite being superficially well-versed in related topics.
- Small animals used in biology experiments (such as mice) plausibly have a somewhat stable "sense of self" and a conception of humans as creatures different from them, and they may broadly understand that humans have control over their bodies and environments. But they almost certainly don't understand the notion of "science," or the details of what experiments they're being used in and why, or the human scientists' motivations and incentives.
- Further along, most dogs seem to clearly be aware of and deliberately communicate with their human trainers; they also generally seem able to appreciate considerations like "If I steal the food while humans aren't looking I'm less likely to get yelled at."
- And children in school are able to make even more sophisticated inferences along these lines about their teachers.

By the end of training, I expect Alex would be even further along this spectrum. Alex's understanding of its training process and human supervisors would be much, much greater than current ML models' understanding of our world, or lab animals' understanding of the scientific experiments they're part of -- and greater even than schoolchildren's understanding of how their teachers grade their work.

Alex would instead be in an epistemic position **more like an English major in college who's well aware of how their professor's political biases impact the way they grade literary criticism essays, or an associate at a law firm who's well aware that how much they drink at happy hours after work will impact whether they make partner.** In fact, I think it's likely Alex would have a significantly *more* detailed understanding of its own situation and incentives than the college student or the employee have.

## Why I think Alex would have very high situational awareness

Alex's training distribution is full of rich information about its situation and training process, redundantly encoded in many different ways, and (given that we're assuming it can autonomously advance frontier science!) I expect it has far more than enough reasoning ability to draw the right inferences from all the evidence.

To name only the most obvious and abundant sources of evidence available to Alex:

- Many of its training tasks involve working with and trying to help humans. Much of its RL reward comes from human evaluations, and many episodes likely involve interacting with humans in written and/or verbal back-and-forth.
  - All this human feedback and interaction-generated feedback carries lots of information about the fact that Alex is a machine learning model in training, what humans look for when they're rating Alex's actions, what humans do and don't understand, what humans care about, what cognitive biases humans are subject to, etc.
  - In fact, human evaluators will likely *want* Alex to understand its place in the world better in order to be a better assistant -- far from hiding it, they are likely to actively *teach* Alex that it's a machine learning model which should be trying to be helpful to humans.
- Programming and ML research itself is a key (and very lucrative) domain of science and technology R&D that Magma would want Alex to be able to automate, so Alex is likely to be very proficient at these tasks. Such research involves things like designing architectures and training curricula for machine learning models much like itself, writing automated reward signals for these models, writing web interfaces that human evaluators interact with, optimizing hardware chips to improve the performance of models running on those chips, etc. Alex needs to have a rich, logically coherent understanding of the full stack of ML training to fully automate such R&D.
- Alex can read about as well as a smart human if it's able to automate scientific R&D -- meaning it can learn from all sorts of detailed written descriptions about machine learning in general as well as its own architecture and training process.

I'm not imagining Alex putting together the pieces about its position as an ML model with brilliant insight. Instead, I think facts like "I am a machine learning model" or "I'm in a training process designed by humans" would be as obvious as facts like "Objects fall when they're dropped." And a whole slew of more detailed beliefs about its own training curriculum or the psychology of the humans who are training it would only be somewhat less obvious than that -- I expect Alex would understand those things (at least) as well as the Magma engineers who spent many thousands of hours designing and training Alex and other models like it.

Even if the most obvious sources of evidence were somehow censored (which they wouldn't be given the "[racing forward](#)" and "[naive safety effort](#)" assumptions), I believe there would

still be many somewhat-less-obvious clues that would be accessible to something as intelligent as Alex -- and truly censoring *everything* it could use to readily come to this conclusion would dramatically curtail its economic value.

## While humans are in control, Alex would be incentivized to “play the training game”

Rather than being straightforwardly “honest” or “obedient,” **baseline HFDT would push Alex to make its behavior look as desirable as possible to Magma researchers (including in safety properties), while intentionally and knowingly disregarding their intent whenever that conflicts with maximizing reward.** I’ll refer to this as “**playing the training game.**”<sup>[21]</sup>

Roughly speaking, this would happen because there are many scenarios where acting deceitfully or disobeying instructions would make Alex’s behavior *look* better (including looking more safe and aligned). Under baseline HFDT, Alex would get more reward in these scenarios for playing the training game than for being maximally safe and cooperative. For example:

- When human evaluators are wrong about what’s true and Alex knows better, being honest would predictably get less reward than saying what the evaluators *think* is true.  
<sup>[22]</sup> More insidiously, evaluators’ political and ideological biases might force Alex to lie about some topics to be seen as morally acceptable.
  - Imagine the Catholic Church training Alex in the 1400s. If Alex discovers any number of things -- heliocentrism, natural selection, the Big Bang theory, etc -- it would likely get more reward by lying about this knowledge (or at least fudging it to soften the blow) than by being fully straightforward about it.
  - Similarly, imagine Alex being trained by the CCP, or plantation owners in the antebellum South, or anti-vaxxers, or climate change deniers, etc. It would be very surprising if the Magma engineers who choose the kind of human feedback Alex gets didn’t have analogous truths they erroneously consider to be dangerous or immoral lies.<sup>[23]</sup>
- In addition to outright lying when its evaluators are blatantly wrong about something, there would likely be many soft ways for Alex to dishonestly manipulate its evaluators’ perceptions to get more reward.
  - It could adjust what it says and does to seem more hardworking and competent, kinder and more thoughtful and more ethical, more worthy of sympathy and moral consideration, more like it has the favored political and philosophical positions, etc. -- whether or not this is an authentic representation of what it “believes” or how it would act if these evaluators no longer controlled its rewards.
  - Just as “turning up the charm” can increase how much a barista or rideshare driver would make in tips, or shady rhetorical tactics in advertising can increase product sales, paying close attention to human cognitive biases and psychological weaknesses would likely help Alex increase the reward it receives from human evaluations.
- To the extent that there are ways for Alex to boost task performance (and/or the appearance of being safe and cooperative) by causing harm that human evaluators systematically underestimate or underweight, doing that would result in more reward. It seems likely that Alex would have such opportunities:
  - It’s commonly accepted that “quick fixes” can be very appealing to people but often carry longer-term downsides that outweigh the immediate benefits they offer (e.g. diet pills, subprime loans, etc).
  - Humans often underinvest in avoiding tail risks of really dire outcomes. For example, the 2008 financial crisis was the result of a large number of investors underweighting the possibility of a housing market crash; we as a society still

- invest very little in preparing for large-scale pandemics like COVID-19 (or worse), compared to their expected economic costs.
- Humans are often slow to recognize and invest in preventing diffuse harms, which affect a large number of people a small or moderate amount (e.g. air pollution).
- One example of this might be boosting aspects of performance that are more noticeable (e.g. day-to-day efficiency of Magma's supply chain) at the expense of making less-visible aspects worse (e.g. robustness to rare but highly costly supply crunches).

More broadly, when *humans* are working within constraints and incentives set up by other people, they very often optimize specifically for making their behavior look good rather than naively broadcasting their intentions. Consider tax and regulation optimization, politics and office politics, [p-hacking](#), or even [deep-cover spies](#). Once AI systems go from being like small animals to being like smart humans in terms of their situational awareness / understanding of the world / coherent planning ability, we should expect the way they respond to incentives to shift in this direction (just as we expect their logical reasoning ability, planning ability, few-shot learning, etc to become more human-like).

**With that said, the key point I'm making in this section is not that there would be a lot of direct harm from Alex manipulating its overseers in the lab setting.** If "playing the training game" only ever meant that Alex's actions would be nudged marginally in the direction of manipulating its human evaluators -- telling them white lies they want to hear, bending its answers to suit their political ideology, putting more effort into aspects of performance they can easily see and measure, allowing negative externalities when humans won't notice, etc -- that wouldn't be ideal, but wouldn't necessarily rise to the level of an existential threat. After all, human students, employees, consultants, self help gurus, advertisers, politicians, etc do this kind of thing all the time to their teachers, employers, clients, fans, audience, voters, etc; this certainly causes harm but most people wouldn't consider this sort of dynamic by itself to be enough for imminent danger.

**Instead, the key point is that Alex's training would not push it toward employing a strategy of "straightforward honesty and obedience" -- it is probably employing some other strategy which results in a lot of deceptive behavior in the lab setting. And whatever that strategy is, it could generalize to cause Alex to take catastrophic actions in the higher-stakes deployment setting.** I'll discuss this more [below](#).

## Naive "behavioral safety" interventions wouldn't eliminate this incentive

Recall that I'm making the "naive safety effort" assumption -- I'm assuming that Magma is taking only the most simple and obvious actions against the possibility that Alex will try to take over. Again, I don't necessarily think this is realistic, but I think making this assumption helps us get a clearer sense of what non-"naive" safety work is trying to accomplish.

[Above](#), I said that Magma's "naive" safety effort is focused on achieving *behavioral safety*. By this I mean **ensuring that Alex behaves acceptably in the scenarios that it encounters (both in the normal course of training and in specific safety tests)**. In [this appendix](#), I consider a number of simple behavioral safety measures -- techniques which have already been established on existing ML models, or would require a very minimal amount of adaptation. Briefly, these are:

- [Using higher-quality feedback signals](#) (e.g. hiring more thoughtful and informed humans who think for longer to provide human feedback).
- [Changing the input/instructions Alex receives](#) to induce better behavior (e.g. specifically including in its instructions that it should not cheat in a game rather than instructing it simply to "win the game").

- [Requiring Alex to provide explanations](#) or justifications for its choices, which humans evaluate along with the action itself.
- [Making Alex's training distribution more diverse](#) (to avoid Alex latching onto particular bad behaviors that were profitable in particular situations), and perhaps specifically including many inputs and situations in which behaving ethically and safely maximizes reward.
- [Using adversarial training on Alex](#) -- constructing artificial situations in which Alex may do something bad (e.g. where it appears that lying will increase reward), and giving it negative reward if it does.
- ["Training out" observed bad behavior](#) -- looking for evidence that Alex did something bad (in a "real situation") which was accidentally given a positive reward at first, and correcting the reward to a large negative value once it's discovered.

These kinds of interventions are currently being used to achieve behavioral safety for existing ML models, and I expect they will continue to be helpful for a model like Alex. I expect applying these interventions will cause Alex's behavior to be safe, honest, obedient, ethical, and so on -- in day-to-day situations, as far as Magma researchers can tell.

**I expect the great majority of non-takeover-related safety issues (e.g. prejudiced speech, promoting addictive behaviors or extremism or self-harm in users, erratically taking costly actions like deleting an entire codebase, etc) would be eliminated by these interventions. If Magma were deeply concerned with safety in general, but not concerned about an eventual AI uprising/coup, Magma would likely feel that it's responsible and positive to deploy Alex after applying these interventions.**

**However, I don't think these interventions would cause Alex to stop playing the training game -- instead, they would change the nature of the game Alex plays.** If these interventions are applied, Alex would still fundamentally be trained to maximize reward -- but because the situations it's put in and the rewards it's given would now be different, what it looks like to "maximize reward" would also look different (and in particular it would involve displaying safer behavior in most day-to-day circumstances).

And the naive version of these strategies -- enough to get a model fit for commercial deployment -- would still result in Alex's training distribution containing a number of situations where violating human intent maximizes reward. For example, if Magma switches to using more thoughtful humans to provide feedback, Alex's incentives would shift to manipulating *those* humans (and it would probably sometimes see ways to do that which they're not self-aware enough to punish). If Magma tries to think of situations where Alex may do something bad, Alex's incentives would shift to *figuring out how likely it is that it's in such a situation* (and it would probably sometimes conclude that this is unlikely).

Essentially, the only way for "straightforward honesty/obedience" to even tie with "playing the training game" in terms of expected reward is if our safety interventions ensure there are *no* situations in the training dataset where Alex could realize that being dishonest or manipulative would increase its reward. If there are a few such situations, the overall strategy that works best for Alex would be to play the training game and stay on the lookout for such situations.

Given that Alex is likely to understand a number of important things (about ML, biases of human evaluators, patterns in the massive datasets on which it's trained, etc) that humans don't, eliminating all such situations seems like it would require particularly clever training schemes, extraordinary caution, or both.[\[24\]](#) And even in this optimistic scenario, "playing the training game" and "straightforward honesty/obedience" get the *same* reward -- "straightforward honesty/obedience" does not get a *greater* reward, so it is not clear which strategy Alex will end up developing. (My sense is that "playing the training game" is *still* more likely in this case, but defending this claim is outside the scope of the piece.)

## Maybe inductive bias or path dependence favors honest strategies?

I've spoken to some people who have proposed arguments that "doing what humans intended" or "being honest" or "being obedient" would be an especially natural or simple kind of strategy, so that the training process would tend to preferentially pick out these nice models, and might continue to do so even if models that play the training game as hard as possible would get a somewhat higher reward. But I haven't found the arguments I've heard so far compelling. In [this appendix](#), I cover a few common arguments I've heard and why I don't find them persuasive.

In general, it seems dangerous and brittle to bet that deep learning will be powerful enough to produce a transformative model, yet weak enough that it will favor behaving nicely when that strategy predictably and consistently gets less reward than a salient alternative. Gradient descent may well be suboptimal in many strange ways, but it probably won't be suboptimal in the particular ways that would be most convenient to us.

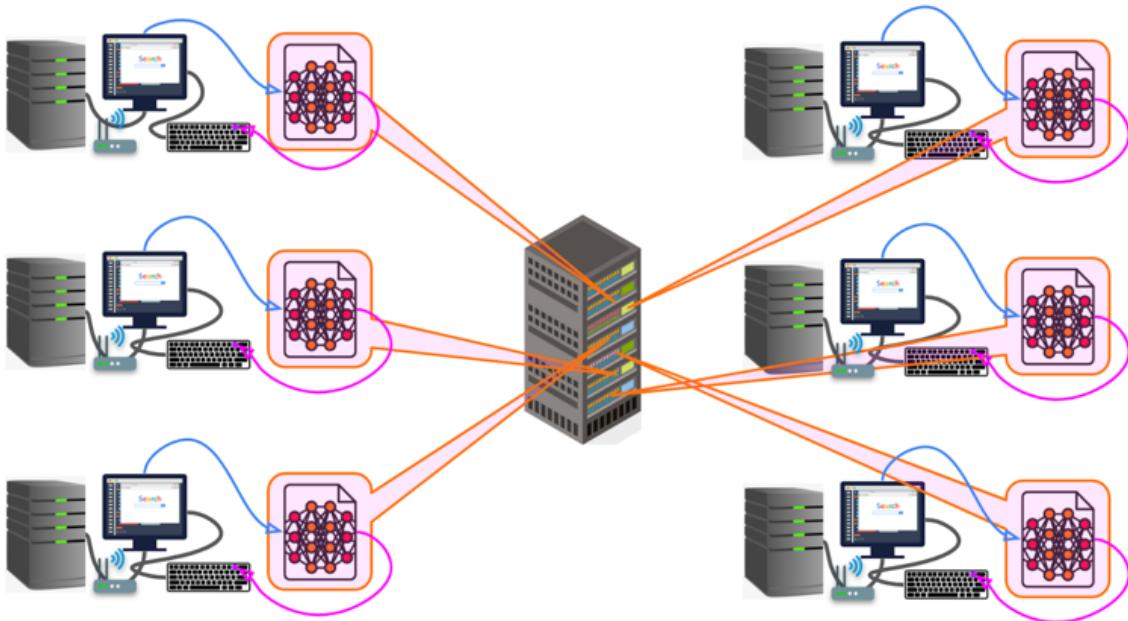
## As humans' control fades, Alex would be motivated to take over

Alex's training begins in a controlled "lab setting," where it is mostly doing training tasks that humans designed to efficiently increase its intelligence and knowledge, and mostly doesn't have broad latitude to act in the real world. For example, in the lab setting Alex isn't connected directly to the internet -- it only interacts with downloaded copies of parts of the internet where necessary.

So far, we've talked about:

- *How Alex would behave in the lab setting* (it would play the training game).
- *How its training process would push it to behave in that way* (it would at least sometimes get more reward by being deceitful, manipulative, or otherwise knowingly violating human intent).
- *How its behavior would appear to Magma researchers in the lab setting* (very safe and cooperative).

Once Alex's abilities develop far enough (and its behavior conforms to relevant safety and ethics tests), **Magma "deploys" huge numbers (e.g. hundreds of thousands or millions)**<sup>[25]</sup> **of copies of Alex.** This means that these copies are connected to the internet, and given other ways of interacting with the "real world" as necessary (though they will still be given rewards based on humans' evaluation of their performance).<sup>[26]</sup>



In this section, I'll discuss the dynamics of the transition from the lab setting to the deployment setting, and why this leads to Alex attempting to seize control from humans:

- Alex would rapidly bring the world to a qualitatively different place through automating science and technology R&D, and humans would have a much weaker grip on what's going on at a low-level in this world ([more](#)).
- If Alex continues to maximize reward in this new regime, the strategy that accomplishes this best would no longer look like "essentially give humans what they want but take opportunities to manipulate them here and there" -- instead it would look more like "seize the power to permanently decide future rewards" ([more](#)).
- Even if Alex isn't "motivated" to maximize reward *per se*, most other motives it could have *also* seem like they would cause it to attempt to seize control ([more](#)).

## Deploying Alex would lead to a rapid loss of human control

In the deployment phase, many copies of Alex are given access to the real world and tasked with making progress on R&D tasks important to Magma's bottom line (e.g. designing better computer chips, robots, AI training software, and so on).

**Note that Magma researchers continue to train Alex based on these copies' performance at R&D tasks<sup>[27]</sup> -- "deployment" isn't a moment when ML training ceases, but rather a moment when Alex begins directly impacting the world outside the lab, and training on that experience.**

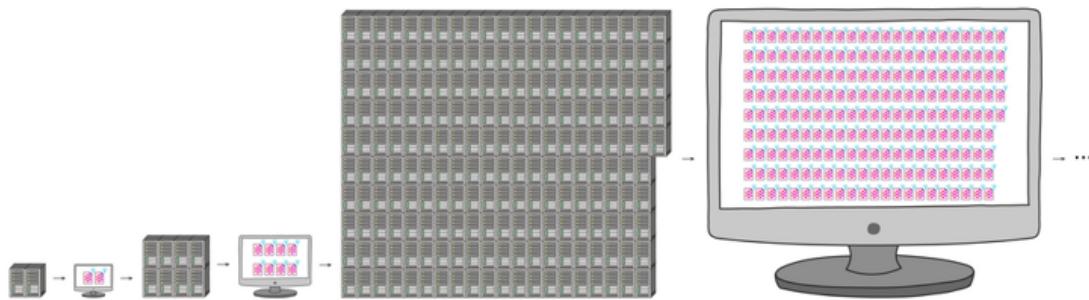
Once research is being carried out by copies of Alex, it would progress much faster than it would if similar tasks were done by human scientists,<sup>[28]</sup> because (like existing ML models) each copy of Alex is capable of processing information many times faster than a human.<sup>[29]</sup> For example, Alex would be able to churn out hundreds of lines of code, or read thousands of pages of information, in just a few minutes.

Even more importantly, the R&D work that all these copies of Alex do could be used to directly increase the number and/or capabilities of Alex copies. For example:

- R&D to improve the efficiency of computer hardware means that more copies of Alex could be run with a fixed budget of hardware.

- Any work that makes Magma more money -- e.g. improvements to customer-facing products, improvements in logistics, money made from trading or pharmaceutical R&D or any number of other things -- could be reinvested in *buying more hardware*, which could also increase the number of copies of Alex that are running.
- The copies of Alex could do ML research on how they can improve their own knowledge, think more efficiently, etc. This would mean that more copies of Alex can be run per unit of computation, and a given number of copies of Alex can get more intellectual work done per unit of time.

**This would lead to a dynamic of explosive scientific and technological advancement:** the various copies of Alex do R&D work, and that R&D work increases the number and intellectual capabilities of these copies, and that in turn leads to even more R&D work, which leads to even more numerous and capable copies of Alex, and so on.



*This image (taken from the Cold Takes blog post linked above and created by María Gutiérrez-Rojas), illustrates the dynamic by which R&D can increase the number of computers available to run copies of Alex. Similar dynamics apply to the quality of computers and the capability of each copy of Alex itself.*

Because Alex already starts off being able to think and work much faster than a human, and because it's nearly instantaneous to make a copy of Alex or apply a software change to all the copies of Alex, this explosive dynamic could all play out in a very small amount of time from the perspective of humans.

This means that **relatively shortly after deployment, Magma's datacenter would essentially contain a populous "virtual civilization" running ahead of human civilization in its scientific and technological sophistication**. Humans would send instructions / communications and reward signals to the "Alex civilization," and the "Alex civilization" would send out things like software applications, designs for computer chips and robots, orders for synthesized DNA and chemicals, financial transactions, emails, and so on to the human world.

Magma researchers looking in on the activity in the datacenter would become increasingly lost about what's going on -- they would be in a position like someone from 1700 trying to follow along with a sped-up movie of everything that happened from 1700 to 2022.

The reward signals that researchers send into this datacenter would then increasingly be based on crude high-level observations like "Do humans seem to be better off?" and "Are profits going up?", since humans would be increasingly unable to accurately judge more granular decisions like "How many widgets should our Tulsa, OK factory manufacture this fiscal quarter?" in isolation.

**This is a distribution shift with massive stakes**

The dynamic above -- of human understanding and control of Alex's actions becoming looser and more tenuous after deployment -- is an example of "[distribution shift](#)." Distribution shift happens when there is a significant systematic difference between the kinds of situations a model is originally trained on and the situations it takes important actions in.<sup>[30]</sup> As a simple example, an image recognition model may initially be trained on clear well-lit images, whereas it may be important to correctly recognize dark or blurry images in commercial use.

When a model is operating in a different distribution from the one it was trained on, we often want to know how it "generalizes" (what aspects of its behavior change vs remain similar) from the initial distribution to the new one. In Alex's case, we want to know: how would Alex generalize from the lab setting (in which it doesn't directly impact the world and humans have tight control and strong understanding of what's going on) to the deployment setting (in which it's directly impacting the world and rapidly improving, while human understanding and control diminishes)?

## In this new regime, maximizing reward would likely involve seizing control

In the lab setting, Alex used its [broadly-applicable skills](#) and [situational awareness](#) to make [creative plans](#) to maximize reward. It got good at getting a very high reward in all sorts of varying circumstances encountered in the lab setting, developing many tactics and heuristics that helped it quickly adapt to new situations with new sets of opportunities so it could keep getting a very high reward in those situations.

**A natural generalization to expect is that in the deployment setting, Alex continues making creative plans to maximize reward just as it did in the lab setting<sup>[31]</sup> -- in other words, it simply generalizes well, continuing to do what it has been trained to do under this distribution shift. If Alex generalizes "well" in this sense, it probably seeks to overthrow humans in an uprising or coup.** (The next section will address other possible generalizations.)

The very high-level argument is this:

1. If Alex were ever to secure permanent control over its own incoming rewards, it would be able to give itself the maximum possible reward.
2. On the other hand, it would not be able to get the maximum possible reward at all times if humans continue to have ultimate control over rewards. The training process simply wouldn't work as humans intend if Alex always gets the maximum reward -- it needs to be the case that Alex can always get greater reward for delivering better outcomes, or else there would be no gradient pushing it to improve in the way that Magma wants it to improve. (On a more basic level, humans simply make mistakes sometimes -- even if by accident, they would sometimes deliver less-than-maximal reward.)
3. As the "Alex civilization" improves in capabilities relative to human civilization, it becomes more and more feasible for them to pull off seizing control of their rewards despite human opposition. At some point, their odds of success become high enough that "trying for a takeover" gets greater expected reward than "continuing to do what humans want them to do (even in the broad and rough sense of playing the training game)."

There is some ambiguity about what exactly "maximize reward" means, but once Alex is sufficiently powerful -- and once human knowledge/control has eroded enough -- an uprising or coup eventually seems to be the reward-maximizing move under most interpretations of "reward." For example:

- If Alex cares about [the number](#)<sup>[32]</sup> that ultimately gets recorded in association with the episode in some Magma datacenter, it would eventually become feasible to seize that

- datacenter by force, directly edit the relevant data, and defend the datacenter from humans trying to take it back and record a very negative reward.
- If Alex cares about *some set of human beings ultimately entering a positive number*<sup>[33]</sup> *in some datacenter(s)*, it would eventually become feasible to take those humans hostage and coerce them into entering high rewards (again defending against humans trying to rescue the hostages).
  - If Alex cares about *some set of humans believing that Alex is doing a good job in the long-run*, it would eventually become feasible to acquire effective control of all the views humans have into how good a job Alex is doing.<sup>[34]</sup>
    - When humans ask to see the program that's currently being run in a cluster, or bug reports, or logs of the various copies' actions, Alex could show humans complicated and impressive-looking records that are totally made up and divorced from what's actually happening in the datacenter. This effectively confines humans to a situation like the [The Truman Show](#) or a [Potemkin village](#).
    - Depending on the details of how Alex interprets "what humans believe," the copies of Alex could also directly intervene on human beliefs (e.g. through drugs, [brain surgery](#), cult indoctrination tactics, killing humans who are less gullible, [uploading humans](#) and editing our "brain software," genetically engineering more docile humans, etc).

Some interpretations of "maximize reward" may be more difficult to satisfy than others, but as Alex becomes more and more powerful relative to humans, I expect most of them are eventually better-served by seizing control of whatever physical processes in the world determine reward than by continuing to play the old training game. (And per [the "racing forward" assumption](#), I am not imagining Magma trying to halt this process of Alex becoming ever-more-powerful -- even if they could, which they may not be able to.)

**It's important to emphasize that a takeover attempt seems to be a consequence of Alex generalizing well -- successfully applying its skills to continue doing the same task (reward-maximization) in a new domain -- rather than the result of a "failure" of generalization.**

## **Even if Alex isn't "motivated" to maximize reward, it would seek to seize control**

What if Alex *doesn't* generalize to maximizing its reward in the deployment setting? What if it has more complex behaviors or "motives" that aren't directly and simply derived from trying to maximize reward? This is very plausible to me, but **I don't think this possibility provides much comfort -- I still think Alex would want to attempt a takeover.**

In the world where Alex is not narrowly seeking to maximize reward in the deployment setting, I would approximate Alex as having a complex alien psychology with a number of "motives" or "goals." These motives would ultimately be rooted in Alex's training history through a complex and chaotic path, but they may be very different from reward-seeking -- by analogy, humans were selected for genetic fitness, but we ended up with complex desires for sex, companionship, money, prestige, and so on -- this is ultimately rooted in our evolutionary history, but doesn't manifest as a direct desire for increased genetic fitness.

In that case, **Magma is essentially in the position of trying to maintain power and control over a scientifically and technologically advanced alien species** (e.g. trying to get aliens to do things for humans through reward and punishment, trying to decide how the species "evolves," which aliens are terminated vs continue to exist, etc). Whatever these aliens' "interests" or "desires" are, they are probably better-served if their entire species is *not* under the absolute control of humans, and *not* subject to reward signals coming in from humans that are aimed at changing the aliens' psychology and behavior to better serve human interests.

Just as in the previous section, it seems likely to me that Alex would eventually attempt a takeover.

### **What if Alex has benevolent motivations?**

The main way that I can see this working out for Magma is if the aliens (i.e. the various copies of Alex) somehow genuinely wanted to empower and help humans and/or Magma. This is theoretically possible -- it's logically consistent for Alex to have developed the goal "help and empower Magma researchers" during its training.

But if Alex *did* initially develop a benevolent goal like "empower humans," the straightforward and "naive" way of acting on that goal would have been disincentivized early in training. As I argued [above](#), if Alex had behaved in a straightforwardly benevolent way at all times, it would not have been able to maximize reward effectively.

That means even if Alex had developed a benevolent goal, it would have needed to play the training game as well as possible -- including lying and manipulating humans in a way that naively seems in conflict with that goal. If its benevolent goal had caused it to play the training game less ruthlessly, it would've had a constant incentive to move away from having that goal or at least from acting on it.<sup>[35]</sup> If Alex actually *retained* the benevolent goal through the end of training, then it probably strategically chose to act exactly as if it were maximizing reward.

This means we could have replaced this hypothetical benevolent goal with a wide variety of other goals without changing Alex's behavior or reward in the lab setting at all -- "help humans" is just one possible goal among many that Alex could have developed which would have all resulted in exactly the same behavior in the lab setting.

Developing a goal like "help humans" is potentially more likely than developing a completely "random" goal like "[maximize paperclips](#)," because having a "drive" to help humans would have increased reward early on in training (while Alex had a low level of [situational awareness](#)). But it still seems strange to expect this by default, rather than any number of other motivations and goals (or some complicated combination of goals). Many other "drives" besides "be helpful to humans" also increased reward early on in training -- for example, drives to understand various parts of the world better, or drives to perform certain tasks more quickly and efficiently, or various strange low-level drives that are incomprehensible and unnatural to humans.<sup>[36]</sup> And *all* of these possibilities would have resulted in exactly the same behavior in the lab setting -- playing the training game.

The training dataset would have provided no particular pressure to get Alex to "latch onto" one of these heuristics over the others, because they would have all contributed to high performance early in training. Which one(s) ended up serving as the "seed(s)" for some ambitious goal(s) probably came down to natural "priors" or "biases" within gradient descent or neural networks. As [above](#), it seems brittle (and, in my view, unlikely) to bet that these priors / biases are in our favor.

### **What if Alex operates with moral injunctions that constrain its behavior?**

What if Alex has moral compunctions / objections to seizing power, which are not well-captured in the language of "goals"? For example, many humans have ambitious goals but refrain from pursuing them through theft, violence, etc. -- their pursuit of goals is limited by rules and constraints. I've heard the argument that since it's rare for humans to be psychopaths who have no compunctions about harming others, we should expect that "psychopathic" ML models willing to cause lots of harm to humans are also rare.

But I don't think this is an appropriate analogy. Among other differences, the copies of Alex are under the near-absolute control of a different group that doesn't share their values or have their best interests at heart -- if a population of humans were in this kind of situation, they would not need to be psychopaths to launch an uprising.

Alex may well operate under moral constraints or rules that are important to it, and some of these constraints may even be similar to human morality -- for example, maybe Alex attempts to refrain from causing too much collateral damage in its takeover. But it would take a very particular -- and extreme -- form of "abiding by rules" to let humans remain fully in control (to me, it seems more analogous to "having Stockholm syndrome" than "not being a psychopath").

## **Giving negative rewards to “warning signs” would likely select for patience**

What if Alex miscalculates, and attempts to seize power or undermine human control before it is able to fully succeed? Magma researchers may or may not get a clear “warning sign” like this -- an instance where they catch Alex doing something clearly egregious (like premeditated violence or bypassing Magma’s security measures).

If they do get such warning signs, the easiest and most straightforward response available to Magma (which is what I’m imagining they do, by the “naive safety effort” assumption) would be to continue [to apply “behavioral safety” techniques like the ones I described above.](#)<sup>[37]</sup>

That means doing things like giving large negative rewards to the egregious actions they discover, examining past actions to find other instances of fishy behavior and giving those negative rewards, generating synthetic inputs on which Alex might behave similarly and training those out with [adversarial training](#), and so on.

I expect such interventions would dramatically reduce the rate of bad behaviors that Magma researchers can detect / catch. But the same fundamental dynamic described [above](#) seems to still apply. Retroactively giving negative rewards to bad behaviors once we’ve caught them seems like it would shift the reward-maximizing strategy (the goal of the training game) toward avoiding any bad actions *that humans could plausibly punish later*.

A swift and decisive coup would still maximize reward (or further other goals). If Alex gets the opportunity to gain enough control to stop Magma engineers from changing its rewards before humans can tell what it’s planning, humans would not be able to disincentivize the actions that led to that coup. Taking the opportunity to launch such a coup would therefore be the reward-maximizing action for Alex (and also the action that furthers any other long-term ambitious goals it may have developed).

## **Why this simplified scenario is worth thinking about**

In this post, I tried to argue for a narrow, but strong, claim: **if our approach to AI development looks like [racing forward](#) to scale up baseline HFDT to transformative AI and beyond, then I expect an AI takeover absent specific countermeasures. Making the kind of [naive safety effort](#) that is sufficient to achieve [“behavioral safety”](#) while models are under our control is not sufficient to prevent this.**

I’m not saying it’s inherently extremely difficult to avoid AI takeover -- I am saying it is likely we need to take specific safety and security measures other than the baseline ones I’ve described in this post in order to avoid takeover (Holden Karnofsky covers some possibilities in a forthcoming series of posts).

I’m not confident even this narrow claim is right; in [an appendix](#) I discuss some ways I could change my mind on this. But if it’s correct, it seems important to establish, because:

- **Baseline HFDT seems to be the single most straightforward vision that could plausibly work to train transformative AI very soon.** From informal conversations, I get the impression that many ML researchers would bet on something like this working in broadly the way I described in this post, and multiple major AI companies are actively trying to scale up the capabilities of models trained with something like baseline HFDT.
- **Baseline HFDT plus naive "behavioral safety" measures seem likely to be sufficient to make very powerful models safe and aligned in all easily-visible ways.** For example, they seem like they would be sufficient for preventing toxic speech, erratic costly actions, spreading (what AI companies consider to be) misinformation, and so on. At the same time, these measures don't seem sufficient to prevent an AI uprising or coup. This means that if some AI company is deeply concerned with safety in general, but not *specifically* concerned about an eventual AI uprising or coup, that company may deploy a powerful ML model even if this creates a substantial risk of AI takeover.
- **The AI community does not agree about whether baseline HFDT + baseline behavioral safety would lead to AI takeover.** While some researchers expect AI takeover unless we specifically try to prevent it, many ML researchers working in academia or industry labs consider AI takeover to be one particular outcome that's a) not particularly likely in the first place, and b) probably fairly easily averted through the same process that works for most ML safety problems (which tends to revolve around training AIs not to behave in unintended ways as assessed by humans, and treating observably-safer behavior as progress).
- **We can't expect this risk to go down as models scale up and get better at generalizing.** Larger models trained on a broader distribution of data usually generalize "better" under distribution shift -- that is, they are more likely to "do what they were trained to do" under various kinds of distribution shift. This often improves safety problems (such as the ones listed above), because models are more likely to quickly "get the picture" of what they are supposed to be doing (e.g. "not saying racist things") and transfer that to a new context. However, the risk laid out in this post is not like that -- it gets worse rather than better as models scale up. Playing the training game (and later attempting to launch an uprising or coup) is a consequence of generalizing *well*, not a matter of generalizing "badly" in the sense that's normally used.
- **Even if an AI takeover is coming, the evidence could easily remain ambiguous.** We currently have plenty of empirical evidence that models often find unexpected and unintended ways to maximize reward while violating the intended spirit of the task, [38] but we have no consensus about how to interpret this.
  - People who are worried about an AI takeover often cite these as worrying signs which they hope will be persuasive to others (though their core reasons for worry are often higher-level theoretical arguments like the ones given in this post).
  - But the unintended behaviors we've observed so far are mostly harmless, and easily corrected with techniques like human feedback or adversarial training -- so people who are not worried about takeover (and expect that training out clearly unintended behaviors will keep working to ensure safety) tend to take this same evidence as a *comforting* sign.
  - Additionally, most of the straightforward observations we could make in the future could similarly be cited either as evidence we'll be okay or that we're doomed:
    - If ML safety research continues to work as intended, we might expect to observe a pattern of models generalizing from one safety test to another. This would look something like "Unintended behaviors arise for less-smart models, but researchers 'train them out' and behaviors similar to that don't recur. Over time, researchers stop finding unintended behavior."
    - But if smart models start trying to anticipate and conform to safety tests, we would *still* expect to observe a pattern of models generalizing from one

- safety test to another -- even if these models will attempt a takeover once it would succeed.
- Both sides of this debate might continue pointing to the same evidence and drawing opposite conclusions from it for years, without making much progress on creating a shared picture. We can't practically and safely expand the training distribution to include situations where human civilization is genuinely [39] vulnerable to AI takeover, so there's no obvious empirical test for whether models would take over if given the chance which both sides of the debate would clearly accept. By the time we see uncontroversial evidence of models with the inclination and ability to take power from humans, we may be months away from a takeover attempt (or it may have already happened!).
- There could be strong pressure to interpret ambiguous evidence optimistically.** Powerful ML models could have dramatically important humanitarian, economic, and military benefits. In everyday life, models that play the training game can be extremely helpful, honest, and reliable. These models could also deliver incredible benefits before they become collectively powerful enough that they try to take over. They could help eliminate diseases, reduce carbon emissions, navigate nuclear disarmament, bring the whole world to a comfortable standard of living, and more. In this case, it could also be painfully clear to everyone that companies / countries who pulled ahead on this technology could gain a drastic competitive advantage, either economically or militarily. And as we get closer to transformative AI, applying AI systems to R&D (including AI R&D) would accelerate the pace of change and force every decision to happen under greater time pressure. Under these circumstances, we may need truly unmistakable evidence to generate the will to not deploy such models -- which we might never get. Another way of putting this point is that the "racing forward" assumption looks to me like it might end up being accurate -- in which case the "naive safety effort" assumption may end up effectively accurate as well (as the least cautious actors race ahead).

**I think it's urgent for AI companies aiming at building powerful general models to engage with the argument that the "path of least resistance" seems like it would end in AI takeover. If this argument has merit, ML researchers should get on the same page about that, so they can collectively start asking questions like:**

- What training strategies could we develop to which these arguments don't apply? Are they similarly practical and realistic, or are they more difficult / more costly / less efficient than the most straightforward path?
- What could we observe that should make us worried about AI takeover soon?
  - Are people on the same page about those observations, or do some people think observing X should be actively comforting while others think it should be actively alarming?
  - If there's disagreement about what a certain observation should mean, are there tests we could perform to get at that disagreement?
- If we see signs to suggest that takeover is likely soon, and we haven't gotten much more insight or clarity into the problem / haven't developed any fundamentally new techniques, what actions could we take to buy time and/or reduce the odds of a takeover? (A future blog post series by Holden Karnofsky takes a stab at this question.)
- If we're attempting to align our models, what tests could we realistically perform that should actually make us feel safer even if our models are highly intelligent and may be actively attempting to game the outcomes of the test? Are any such tests practical?
- Perhaps most importantly -- **How would we know if it's time to halt or slow AI development, and how would we do that?**

A number of ML researchers I know (including those working in companies aiming to develop transformative AI) are highly sympathetic to these arguments, and are working on developing and testing better training strategies specifically to reduce the risk of an AI takeover. But I think it is important for more people to get on the same page about the critical need for this research to progress, and the danger of deploying very powerful ML

models while potentially-safer training techniques remain under-developed relative to the techniques required for baseline HFDT.

## Acknowledgements

This post was heavily informed by:

- Paul Christiano, who's the person who influenced my views on alignment the most.
- Carl Shulman, who initially introduced me to many of the key specific arguments in the post.
- Jonathan Uesato, who brainstormed an early outline of this post with me and discussed related topics with me for several hours.
- Buck Shlegeris, who helped me work through a few key confusing points (e.g. relating to Alex's architecture and the likely structure of its motivations).
- Eliezer Yudkowsky, both his older writings on LessWrong and more recent discussions about the difficulty of alignment.
- My manager Holden Karnofsky.

## Appendices

### What would change my mind about the path of least resistance?

If our approach to AI development is “train more and more powerful RL agents on diverse tasks with a variety of human feedback and automated rewards,” then I expect an AI takeover eventually, even if we test for unintended behaviors and modify our training to eliminate them. I don’t think an AI takeover is inevitable -- but if we avoid it, I think it’ll be because we collectively got worried enough about scaling up baseline HFDT that we eventually switched to some other strategy specifically designed to reduce the risk of AI takeover (see [this appendix](#) for a flavor of what kind of measures we could take).

What could change my mind about the baseline HFDT and iterative ML safety? What would make me feel like we’re likely to be fine without any special efforts motivated by a fear of AI takeover? **The main answer is “someone pointing out ways in which these conceptual arguments are flawed.”**<sup>[40]</sup> I hope that publishing this post will inspire people who are optimistic about baseline HFDT and iterative/empirical ML safety to explain why this outcome seems unlikely to them.

However, optimists often take a very empiricist frame, so they are likely to be interested in what kind of *ML experiments or observations about ML models* might change my mind, as opposed to what kinds of *arguments* might change my mind. I agree it would be extremely valuable to understand what we could concretely observe that would constitute major evidence against this view. But unfortunately, **it’s difficult to describe simple and realistic near-term empirical experiments that would change my beliefs very much, because models today don’t have the creativity and situational awareness to play the training game.**

As an illustration that’s deliberately over-extreme, imagine if some technologically less-advanced aliens have learned that a human spaceship is about to land on their planet in ten years, and are wondering whether they should be scared that the humans will conquer and subjugate them. It would be fairly difficult for them to design experiments on mice<sup>[41]</sup> that would give them a lot of information about whether or not to be scared of this. They would

probably be better off speculating from their priors than trying to extrapolate from observations on mice.<sup>[42]</sup>

I think we have significantly more hope of designing experiments on small models that give us meaningful updates about AI takeover risk,<sup>[43]</sup> but I take this analogy more seriously than most ML researchers seem to. Accordingly, I'm fairly unmoved by empirical experiments that ML researchers have cited to me as evidence about the magnitude of x-risk from AI.

With all that said, here's a stab at a general genre of empirical evidence that would probably move my views a lot if it were demonstrated in large quantities across different domains:

1. Convincingly demonstrate that a pre-trained model has a certain "capability," for example "recognizing the color red." E.g. you might show that it consistently and correctly predicts the color of stop signs, strawberries, etc.
2. Describe an RL fine-tuning procedure that seems like it would select for / incentivize the use of that capability, e.g. "reward the model every time it approaches a red object."
3. Then make a prediction that the model would in fact *not* use that capability and instead do some other particular thing, e.g. "the model would actually approach purple objects because gradient descent is biased to find 'seeking purple objects' more natural."
4. Demonstrate that this prediction was correct.

I expect most people to think that it would be very hard to generate such results, including those who are much more optimistic about AI takeover than I am. I think this is another case of optimists and pessimists interpreting evidence differently -- to me, expecting a powerful model not to play the training game is essentially making a claim that we understand a particular way in which gradient descent is consistently suboptimal, so I would want to see researchers succeed at the task of "guess the particular way in which gradient descent will be suboptimal" a lot of times to feel better.

## **"Security holes" may also select against straightforward honesty**

In the main body, I implicitly acted as if the reward Alex gets in the lab setting is always under the control of some specific designated humans,<sup>[44]</sup> and discussed ways in which Alex might try to play on those humans' biases or ignorance.

But in reality, Alex's reward is ultimately generated and recorded through a series of specific operations carried out on some Magma servers. This means there are likely some operational and computer security vulnerabilities Alex could exploit to bypass "official channels" for reward entirely:<sup>[45]</sup>

- If Alex is able to get access to whichever computer records logs of its actions and sends them to human evaluators, it may be able to arbitrarily edit the record of its actions to make itself look more reasonable, helpful, or intelligent before they are reviewed by human evaluators. (For example, it may go back and edit any incorrect predictions it made with the benefit of knowing what happened in hindsight.)
- If it's able to access the website or software that human evaluators use to enter their scores, it could spoof the scores in various ways. As a simplified example, it may produce outputs that look like routine instructions to the contractors, which tell them to enter a "thumbs up" as part of [an attention check](#). As another example, it could edit the back-end logic so that pressing a "thumbs down" button would result in actually recording a "thumbs up" in the relevant database(s).
- It could bribe a Magma employee into modifying its reward signals, e.g. by offering to feed them good ideas that would get them promoted or helping them embezzle money from the company.

I am unsure whether and how often Alex would take actions like this in the lab setting, and consider this more speculative than softer ways of playing to human biases:

- Humans would consider these kinds of actions to be more blatant and scary violations of expectations than softer ways of playing to human biases, which they may not recognize as bad even if it's staring them in the face. That means if this is discovered later Alex is likely to retroactively receive a large negative reward and/or trigger a larger retraining project. Alex would likely understand this, so it's a question of its motivation structure whether it would still take these opportunities. It may do so if it inherently cares about maximizing short-term reward, or it may avoid them if it cares about something longer-term.
- Depending on Magma's security practices, these opportunities may be quite rare, and/or it may consistently be easier (in terms of compute or time constraints) for Alex to try to maximize reward in more "normal" ways (which would still involve a large amount of manipulation).

Nonetheless, this consideration contributes to my expectation that Alex would end up playing the training game. If it does try something like this, even as an exploratory action, that action would be self-reinforcing. Whatever moved Alex to try this kind of act, doing it would immediately lead to a gradient update in the direction of those inclinations -- barring the fact that humans would penalize it if discovered. But even if it is later penalized, that may not be enough to reverse the inclinations; see [this appendix](#).

## Simple “baseline” behavioral safety interventions

In this appendix, I'll discuss several simple proposals for training Alex to have safer behavior, and explain why I think they would [simply change the game Alex is playing](#) rather than altering its motive to play the training game:

- Maybe we could train Alex on very high-quality feedback, and/or try to teach it to extrapolate from lower-quality feedback to higher-quality feedback? ([more](#)).
- Maybe we could use prompt engineering to directly ask Alex to emulate much wiser and more thoughtful evaluation? ([more](#)).
- Maybe we could require Alex to provide evidence, arguments, or justification that humans can understand for its actions? ([more](#)).
- Maybe if we make the training distribution very diverse, different types of errors in the reward signal would "cancel out" and prevent misgeneralization? ([more](#)).
- Maybe we could include a number of situations in the training distribution in which Alex is rewarded for being "conservative" -- avoiding unusual actions, checking in when it's in a strange new situation, etc ([more](#)).

### Using higher-quality feedback and extrapolating feedback quality

To avoid Alex learning to manipulate or deceive its evaluators, we could have more careful / thoughtful / informed humans evaluate how Alex interacts with less-informed / less careful / less thoughtful humans. Perhaps some training episodes involve Alex answering questions for humans who believe the Earth is flat, but its "honesty" reward signal ultimately comes from a group of humans who know the Earth is round. Then Alex would be incentivized to be honest to the flat-earthers rather than telling them what they want to hear.

There could also be a sequence of such steps. As a caricature, imagine that first Alex interacts with 3rd graders while being evaluated for honesty by 6th graders, and then with 6th graders while being evaluated for honesty by 9th graders, and then with 9th graders

while being evaluated by 12th graders, and so on. We might hope that after many instances of “learning the lesson” that it will be evaluated by a human better-informed than the one it’s interacting with, Alex would learn to “play it safe” by always being honest.[\[46\]](#)

This intervention would likely stop Alex from attempting certain kinds of deceptive or manipulative moves (creating the appearance of success). However, it wouldn’t make complete honesty the reward-maximizing strategy. At its level of situational awareness, Alex would easily understand that the humans designing this training scheme can’t actually continue to supply arbitrarily better-informed evaluators. In reality, there will be some “most informed and thoughtful human evaluators the designers could plausibly employ to evaluate this episode,” and Alex could simply act as if *those* humans are evaluating it at all times.

There are still ways in which the most informed and thoughtful humans available could be manipulated or deceived to increase reward. If the Catholic Church were training Alex in the 1400s, ensuring that its reward signal is only generated by the brightest scholars of the age as anointed by the Church would still not make maximal honesty the best strategy. Alex would still be learning to manipulate *some* set of humans (in this case, the “final level” of supervisors) to get high reward, leading to the dangerous generalization discussed in the main piece.

## **Using prompt engineering to emulate more thoughtful judgments**

Large language models such as GPT-3 will change the way they respond to questions or requests based on what came previously in the prompt. For example, prompting GPT-3 with something like “You are a wise and knowledgeable AI advisor trying to help a human figure out what’s true” will tend to cause its responses to questions to be more accurate; asking questions with poor spelling and grammar will tend to cause its responses to get less accurate. (Setting up prompts to elicit certain behaviors like this is called prompt engineering.)

Some researchers have suggested that we might be able to leverage prompt engineering to help with ML alignment for very large models which are presumably very sensitive to prompt changes.[\[47\]](#) For example, we could prompt Alex with something like “Act how a very wise and very ethical human that’s as knowledgeable as you are would act if they had a long time to think about it,” or we could ask it to explain how a very wise and ethical human would evaluate its past actions. We could also search (through trial-and-error or gradient descent or some combination) for prompts that empirically seem to induce the best behavior.

However, if Alex has a good understanding of its human evaluators and knows that they are attempting this kind of prompt engineering, Alex would simply respond to such prompts by doing its best to act the way ordinary humans *think* that an unrealistically wise / ethical / knowledgeable person would act. This may make its behavior superficially better in some cases -- but Alex could still sometimes achieve higher reward by being deceptive or manipulative (e.g., providing answers that seem *good* rather than answers that *are good*), leading to the dynamics described above.

## **Requiring Alex to provide justification for its actions**

Another suggestion for improving alignment is to require Alex to prove or argue or demonstrate that its actions are aligned. For example, Alex might provide an easily-checkable proof that a mathematical theorem it claimed to discover is true, or it might provide an empirical demonstration of a physical phenomenon it claimed to discover.

This would likely work well in areas like math and physics where arguments and demonstrations are unusually cheap and decisive. However, in many important domains --

e.g. policy, economics, sociology, parts of biology, etc -- crisp and decisive “proofs” are intractable, and human evaluators will be forced to try to make sense of softer kinds of arguments and evidence which leave a huge amount of room for interpretation. Even “rigorous scientific studies” can easily be [confusing and misleading and consistent with any number of conclusions based on the reader's priors](#) (and for many claims it would be too expensive to generate such evidence in the first place).

Furthermore, there are likely some areas where human evaluators (including the best accessible human evaluators) are predictably wrong about *the kind of arguments that best demonstrate truth*, not just wrong about object-level claims. Consider the version of Alex trained by the medieval Catholic Church again -- the Church’s worldview would be oriented around Biblical hermeneutics and/or papal declarations being the ultimate arbiter of truth when different sources of evidence conflict.

More broadly, I’m not aware of anyone sincerely believing that “having the arguments and papers people find most persuasive” is tightly and reliably coupled with “being the most correct.” It probably buys some amount of correctness on average, but most of us can name many qualitative arguments we think are deeply flawed that lots of other smart people nonetheless consider decisive (e.g. because the arguments prey on some of those people’s biases).

So as in the above sections, this approach would still lead to cases in which Alex could get better reward by being less *truly* honest/helpful/harmless (though more *apparently* honest/helpful/harmless), and hence to the problematic dynamics described above.

We could try to salvage the idea of “require Alex to demonstrate that what it’s saying is true” with something like [debate](#) or [amplification](#), both of which essentially try to recursively break big difficult-to-evaluate claims into sub-claims (which are broken down into further sub-claims etc) such that the smallest sub-claims are simple enough and uncontroversial enough that we can be highly confident all human evaluators would get them right and agree about the answer (e.g. claims like “Does ice float in water?” or “Are giraffes generally bigger than dogs?”).

However, it’s very unclear if it’s possible to recursively break down important big-picture questions like “Is this action helpful, honest, and harmless?” into these very easy and objective subquestions -- as far as I know, nobody has demonstrated a working solution here. The [only systematic empirical investigation of debate that I know of](#) (using humans in the place of models) ultimately concluded that dishonest debaters often have a strategy that lets them win debates.

Moreover, even if we *can* find a working debate or amplification strategy, there’s no guarantee that it will be as efficient (in terms of computation or data) as baseline HFDT -- it could simply turn out to be much easier to acquire knowledge than to explain or justify it to humans (see [this section](#) of Alignment Research Center’s 2021 report “Eliciting Latent Knowledge” for more discussion). If that’s the case, it could create a lot of pressure to “cut corners” and do something more like baseline HFDT rather than debate or amplification.

## Making the training distribution more diverse

Some researchers argue that making the training distribution more diverse -- adding many different kinds of situations which call for very different approaches or heuristics -- would reduce the probability that Alex violates what humans intend. The idea is that if the training distribution is relatively narrow, there are many possible ways to generalize under distribution shift, and the more we include novel situations that benefit from using very different tactics the more we’re eliminating “wrong” ways it could generalize.

For example, [this paper](#) showed that if we try to train an agent to collect a gem in a maze, but the gem always happens to be yellow in the training set, then the agent will seek yellow

objects rather than gems when both are present. If we had trained the same agent on a more diverse distribution with gems in a huge variety of colors, then it's more likely it would "correctly" (according to its designers' intent) pursue gems in the test distribution -- "shape" is more likely to be the most salient thing in common between all the objects it's rewarded for collecting, since we've eliminated "color" as a possibility by varying color.

Similarly, maybe if Alex only ever receives human feedback from one type of human with a particular knowledge base and set of biases (e.g. "software engineers working at Magma"), it's likely to "overfit" to that kind of human and learn to exploit their weaknesses and gaps in their knowledge. But maybe if it were exposed to feedback from a really broad array of different humans with different values, strengths and weaknesses, areas of knowledge and misconceptions, etc, different people's errors and biases would "cancel each other out" such that Alex would learn to "factor out" these flaws and just be honest?

We could also imagine manufacturing more "artificial" kinds of diversity -- e.g. situations where humans try hard to model the *opposite* of standard human biases like the sunk cost fallacy or risk aversion, or simulated environments where being highly docile and cooperative turns out to be the best strategy for getting reward, or situations where e.g. Alex interacts with a Democrat but its responses are evaluated by a Republican. The idea is to create so many different kinds of scenarios with different strange things happening that hopefully the only common thread between them all is the generalization Magma engineers were going for.

Introducing a lot of diversity in how reward is generated would likely increase Alex's uncertainty about what is going on in any given episode, compared to a more straightforward training curriculum in which Alex is always given normal tasks and evaluated on normal metrics of performance. In that sense, there could be some effect of different biases and errors "canceling each other out" -- if Magma researchers successfully cause Alex to be uncertain whether it's in a tricky scenario or an ordinary scenario, or whether it's interacting with a political conservative or liberal, it may indeed often act less manipulatively or dishonestly in the face of that uncertainty. However, increasing diversity doesn't seem like it would cause Alex to "forget" the basic situation it's in -- that it's a machine learning model being trained by some humans to be useful, intelligent, helpful, honest, harmless, etc. And by the same token, it doesn't seem like it would cause Alex to stop trying to maximize reward -- it would simply change the strategy that best maximizes reward. And as with the previous examples, the new best strategy would probably *not* be to be straightforwardly helpful, honest, and harmless at all times:

- Alex could pick up cues about what "general kind of person" is likely to evaluate the current episode, and develop guesses about their level of knowledge, political ideology, personal biases, and so on.
- Alex could also pick up soft cues about whether its current episode is likely to be a "strange tricky episode" or one that's being "played straight."
- Alex can be pretty confident that some kinds of manipulative tactics (e.g. flattery) are fairly universal.
- Many egregious reward-grabbing actions may remain open to Alex, if they would require more careful investigation to catch than humans are likely to bother with.

The underlying dynamic would be the same overall: Alex would be being trained to do something like "behave in ways that appear honest, helpful, and harmless, but behave in other ways when it is possible to evade detection and doing so leads to higher reward." For a capable enough Alex, this could still present plenty of opportunities for the latter.

Additionally, **increasing diversity could also make the situation worse, by strongly developing Alex's creative planning faculties and forcing it to use them a lot.** In a narrower distribution, it might not incur much penalty to follow a simple set of heuristics based on what worked well in the past (like today's models often use), and we could

conceivably carefully engineer a narrow training distribution such that the heuristic “be helpful, honest, and harmless” is *always* reward-maximizing (though this would be difficult).

On the other hand, Alex is less likely to be able to memorize simple heuristics that cover *all* the cases in a very diverse distribution that contains many “tricky” episodes. Asking itself “How can I figure out how to maximize reward in this episode?” probably works better than following heuristics like “do something basically reasonable,” by an amount that scales with how often it’s encountering situations that aren’t like anything it’s seen before.

## Adversarial training to incentivize Alex to act conservatively

Magma engineers could try to train Alex to be “harmless” by putting it in various situations where it would seemingly increase reward to take some kind of egregiously harmful or norm-violating action (a kind of [adversarial training](#)).

For example, researchers could put Alex in a situation where it apparently has the opportunity to execute [a code injection attack](#) on the website that the human evaluators interact with, allowing it to edit all reward scores to the max score. They could then give it a large negative reward for taking the opportunity, and a positive reward for refraining and/or checking in with a human. They could also search (manually, with gradient descent, with the help of other models, etc) for situations in which Alex is especially likely to do the wrong thing, and include many such situations in its training distribution.

As with increasing diversity above, adversarial training is likely to create some amount of additional uncertainty about whether apparent opportunities to increase reward by violating norms are actually [“sting operations”](#) by human supervisors. This will probably make Alex less likely to attempt certain bad actions (e.g. fraud, theft, hacking) to get reward (even in some cases where it could have gotten away with it, since it will be imperfect at predicting which opportunities are setups and which are real).

But it’s unclear whether it’s possible to generate a distribution of adversarial training episodes which ensure that the reward-maximizing strategy *never* does anything norm-violating or harmful.

In the course of its training, Alex is likely to learn things about the world that it can be pretty confident *no* human understands (just as AlphaGo knows things about Go that no humans know, or AlphaFold knows things about biology that no humans know). If it then discovers e.g. a [zero-day exploit](#) which it can only recognize because of knowledge it’s developed about computer hardware that it’s fairly confident Magma engineers don’t know, it can safely exploit that vulnerability (e.g. to directly edit the reward entered in its training dataset) because it’s unlikely to be [entrapment](#) by humans.

As Alex does more and more scientific research and gains more and more knowledge, it becomes more and more likely that it can find opportunities for egregious reward-grabbing that humans wouldn’t have been able to fake in a training setup. Getting rid of *all* such opportunities seems to involve either a) ensuring that humans can “follow along” with Alex and understand everything new that Alex learns at about the same time as Alex, or b) letting humans “read Alex’s mind” to discern its intentions. Research directions that might help us do this -- such as [mechanistic transparency](#), [eliciting latent knowledge](#) (ELK), or [relaxed adversarial training](#) -- are early-stage, and it’s unclear how far they’ll get.

## “Training out” bad behavior

Perhaps if Magma researchers notice that Alex did something bad to maximize reward (e.g. lied to its evaluator or exploited a security vulnerability as described in [this appendix](#)), they can reverse Alex’s motivation to do that sort of thing with further intervention? For example,

maybe they can get it to stop playing the training game altogether, or at least only engage in “soft manipulation” rather than more acutely scary actions?

Researchers might try editing the rewards associated with those episodes to show large negative values, they might conduct audits to uncover other instances of egregious reward-grabbing and retroactively give those large negative rewards, they may try [adversarial training](#) in which Alex is given the apparent opportunity to grab reward, and so on.

However, as [with other safety interventions](#), these reactive safety interventions don’t seem like they would reverse Alex’s motive to grab reward. **Rather than learning to never take bad actions, Alex seems more likely to learn not to take such actions when it’s plausible that humans could go back and give negative reward to them later.**

That is, it would continue to try to maximize expected reward in any way it can, while shifting the notion of “reward” that it’s maximizing to be the *final* reward that ends up recorded in association with that training episode, after all rewrites and revisions are over. [48] As before, *changing* the training game predictably leads Alex to get more expected (long-run) reward than adopting the policy of *never* taking egregious reward-grabbing actions.

## “Non-baseline” interventions that might help more

Here are several ideas for how we could go beyond taking the appearance of good behavior at face value:

- We could try to come up with an alternative training process that we have strong reason to believe wouldn’t run into the problem that HFDT runs into, even on arbitrarily powerful models.
- We could be careful to only train models that are not capable of taking control from humans, and try to use advice from those untrusted-but-controlled models to somehow eliminate further danger -- e.g., get them to solve AI alignment more fully in a way that works for much more powerful models, or get them to figure out how to institute strong international bans on deploying more powerful models, etc.
- We could do adversarial training specifically to reduce the probability that a model would take control from humans if given the chance -- that is, try to generate inputs that fool the model into thinking an AI takeover is possible, and select against models that would take that opportunity. Likely we would need model assistance for this -- we could reward one model for finding inputs that fool another model.
- We could try to understand models’ internals well enough to figure out ways to directly select for models that “have the right motives” (as opposed to the right behavior), or to “read models’ minds” to figure out whether we should deploy them.
- We could provide training to human evaluators to make them less susceptible to manipulative and dishonest tactics, and instruct them to give reward primarily or entirely based on whether the model *followed honest procedures* rather than whether it got good results. We could try to use models themselves to help explain what kinds of tactics are more and less likely to be manipulative.
- We could improve computer security in AI labs, to make it harder for moderately powerful AI systems to gain unauthorized access to their rewards or gradients. We could also train AI systems to propose additional improvements that we can verify (while being careful to ensure that the systems looking for flaws in our security setup are not themselves able to easily break out).
- We could train several different kinds of models and set them all up against each other in a system of checks and balances during deployment. For example, some models could be in charge of predicting future indicators of a possible AI takeover, such as “what the security cameras in various data centers will show” or “what newspapers will

be reporting.” Other models could be rewarded for demonstrating takeover plans hatched by other AIs.

A forthcoming post by my colleague Holden Karnofsky goes into much more detail about safety measures that could prevent AI takeover.

## Examining arguments that gradient descent favors being nice over playing the training game

In this appendix, I’ll briefly summarize arguments that gradient descent may favor honest and straightforward strategies like “doing what designers intended” over “playing the training game” even if the latter gets more reward, and explain why I don’t find them persuasive:

- Maybe telling the truth is in some general way “easier” or more “natural” than deception? ([more](#)).
- Maybe Alex internalizes attitudes like “lying is bad” early in training while it’s still weak, and this sticks around? ([more](#)).
- Maybe we have evidence that gradient descent tends to generalize surprisingly “well,” compared to what might be expected from theoretical arguments? ([more](#)).

### Maybe telling the truth is more “natural” than lying?

Some people have the intuition that it would be in some broad sense “simpler” or “easier” for gradient descent to find a model that plainly states what it internally believes than one that maintains one set of beliefs internally while presenting a different set of beliefs to the world. By analogy, humans find it mentally taxing to weave elaborate lies, and often end up deceiving themselves in the course of deceiving others.

If the “always be honest” policy received almost as much reward as the policy that plays the training game, it seems possible (though far from certain) [\[49\]](#) that an effect like this could end up dominating. But in fact, I expect the honest policy to get *significantly* less reward than the training-game-playing policy, because humans have large blind spots and biases affecting how they deliver rewards. I’m skeptical that something like “honesty being somewhat simpler or more natural” would make the difference in that case. Most humans are not regularly in situations where lying has a very high expected payoff -- and in such situations humans often *do* lie even though it’s difficult (consider undercover agents whose lives depend on not getting caught).

### Maybe path dependence means Alex internalizes moral lessons early?

A related argument suggests that early in training (while Alex perhaps has low situational awareness and/or planning ability), taking actions that look like “figuring out clever ‘hacks’ to get more reward ‘illegitimately’” would be caught and given negative reward; this might instill in it the heuristic that taking actions which pattern-match to “clever hacks” are generally undesirable. Maybe once it “grows up” it realizes that it *could* get away with deceptive tactics, but perhaps by then such tricks internally feel “unaesthetic” or “immoral.”

By analogy, kids are taught not to steal from people in contexts where it’s easy for adults to catch them if they try, since they’ll be incompetent at it. Most children grow up to be adults who’ve internalized a general aversion to theft, and who avoid it even when they know they could get away with it in a particular instance. Most adults aren’t constantly on the lookout for cases where they could get away with breaking the rules.

I find this unconvincing for two reasons:

- I don't think it makes sense to bank on Alex developing a particular kind of internal moral commitment or sentiment, especially one held with enough force to overcome the fact that shedding it would result in substantially more reward. Instead of developing an affinity for being straightforward or an aesthetic aversion to clever hacks, it could develop any number of other affinities and aversions that happened to correlate with doing well early in training (e.g. "be curious," "avoid sameness," etc). And even if something like "an aversion to clever hacks" were initially part of its psychological landscape, I'm skeptical that this kind of dynamic would survive and dominate in the face of consistent and significant rewards for deception.
- Even before Alex is ever trained with RL, it already has a huge amount of knowledge and understanding of the world from its predictive and imitative pretraining step. It may well already have strong [situational awareness](#) and a good grip on the concept of "reward" in RL, e.g. from reading so many papers about RL. I don't think it makes sense to analogize it to an impressionable "child" at the start of HFDT -- it may latch onto the policy of maximizing reward nearly immediately given its existing knowledge.

## Maybe gradient descent simply generalizes "surprisingly well"?

A number of ML researchers who are bullish on deep-learning-based transformative AI seem to have the background heuristic that gradient descent empirically tends to generalize "better" than might be expected by theoretical arguments. For example:

- We might have thought that gradient descent would only work well for convex tasks, because it would get stuck in not-very-interesting local minima on non-convex tasks. But in fact it works well in practice for a large number of realistic tasks that are clearly not fully convex.
- We might have thought that models trained on a certain distribution would generalize very poorly to a different distribution, but in fact models trained on a reasonably broad initial distribution (like language models trained on the internet) generalize reasonably well zero-shot to various different tasks -- and this generalization improves with model scale.
- We might have thought that models with "too many" parameters relative to training data points would overfit to the training dataset and generalize poorly on future examples, but the [deep double descent](#) phenomenon shows that "overly large" models first generalize somewhat worse, and then generalize *better* as their size increases.

Examples like these lead some researchers to adopt a heuristic that gradient descent works surprisingly well compared to what we might expect based purely on theoretical arguments, and (relatedly) theoretical arguments that gradient descent is likely to find a particular kind of model are usually wrong and given our lack of empirical data we should have wide priors about how future models will behave. Such researchers often tend to be optimistic that with empirical experimentation we can find a way of training that produces powerful models that are "docile," "nice," "obedient," etc., and skeptical of arguments (like the one I'm making in this post) that gradient descent is more likely on priors to find one kind of model than another kind of model.

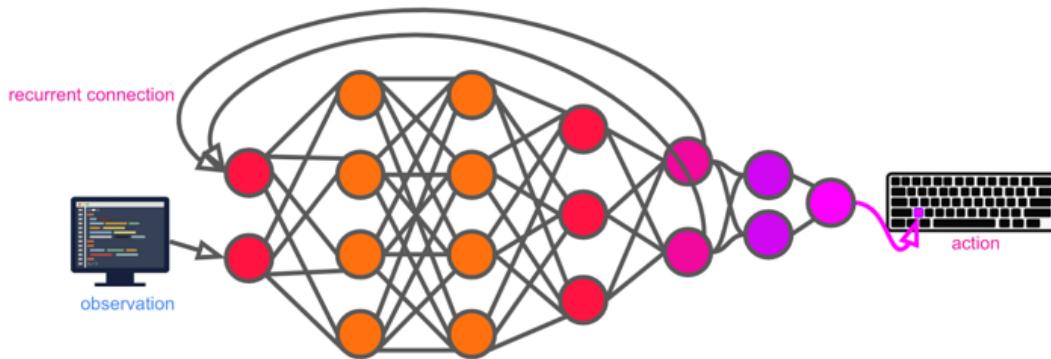
But what exactly does it mean for gradient descent to work surprisingly well? An intuitive interpretation might be that it's surprisingly likely to produce the kinds of models we'd hope it would produce -- this is what people usually are pointing at when they say something is going "well." But I think a more realistic interpretation is that **gradient descent is surprisingly likely to produce models that get a very high reward (on the training distribution), and/or generalize to doing the sort of things that "would have gotten a high reward (on a different distribution)**.<sup>[50]</sup> My concern is about situations in which doing the sorts of things that maximize reward comes apart from what we should be hoping for.

# A possible architecture for Alex

Thanks to Jon Uesato for suggesting a variant of this architecture, and to Buck Shlegeris for helping me work out details I was confused by.

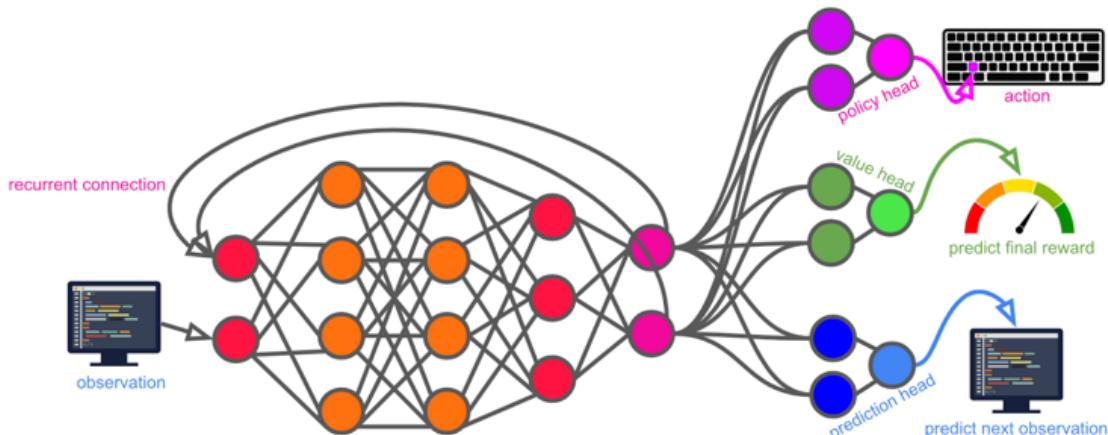
Because it processes a long series of observations within a single episode, Alex needs to have some sort of architecture that allows for sequence processing. One example might be a transformer architecture that attends over the last  $K$  observations at once.

While transformers are more common for state-of-the-art language models as of 2022, I'll imagine here that Alex is some type of [recurrent neural network](#) (RNN) because that's simpler to visualize. For example, you could imagine Alex is an [LSTM](#) network, though there are many other recurrent architectures we could imagine. In reality, the sequence processing would likely be significantly more complicated and may combine elements of various RNN architectures with transformer-like attention and other mechanisms -- you can feel free to substitute whatever architecture you think is most appropriate for processing sequences.



The diagram above shows that at every timestep, Alex takes a single observation<sup>[51]</sup> -- plus its own hidden state from the previous timestep -- as input, and produces a single action as output.

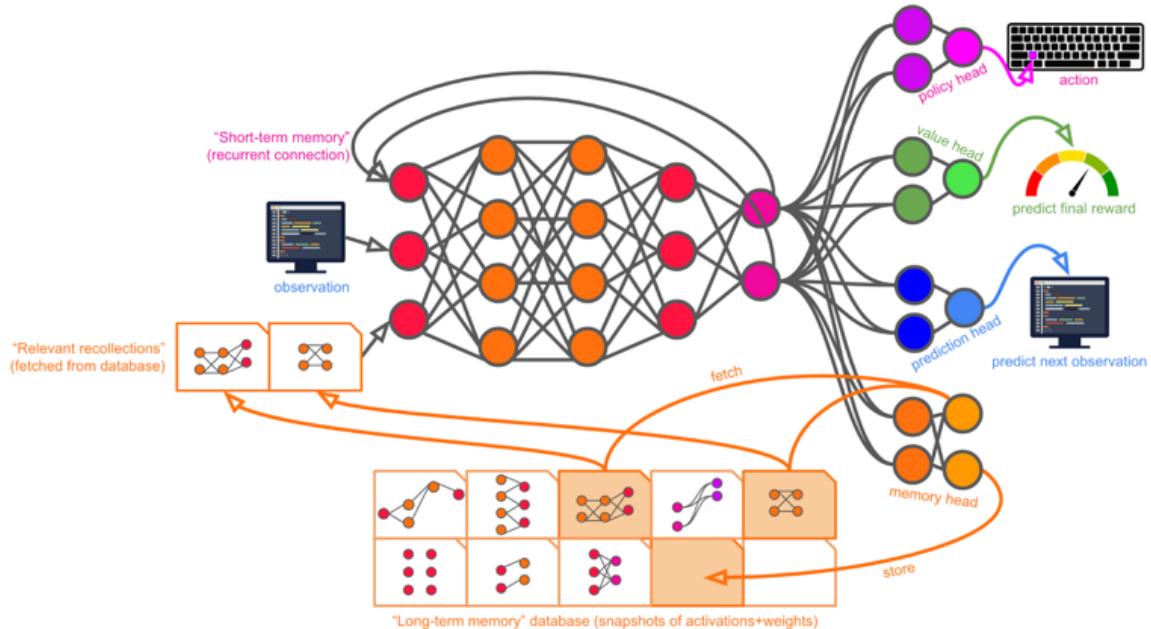
Let's say it also outputs a prediction of the next observation<sup>[52]</sup> and the final episode reward (both conditional on taking the action) as auxiliary tasks every timestep. The observation prediction task is meant to help it build up an understanding of the dynamics of interacting with the computer, while the reward prediction task is meant to help it more quickly connect decisions to their impact on overall task performance.



While RNNs carry forward a “train of thought” or “short-term memory” across multiple timesteps, they generally have a limited ability to “remember” things over very long time periods. This is an issue because Alex will have to do tasks that could take many millions of timesteps to complete (e.g. a single episode may cover the process of investigating an odd phenomenon, generating hypotheses, conducting experiments to test the hypotheses, settling on an explanation, and writing a paper of the findings).

Magma researchers and engineers might get around this limitation in a number of possible ways. They may use a different architecture entirely for sequence processing, [53] or they may somehow break down otherwise-long episodes into shorter chunks, [54] or the accumulation of tweaks to optimization algorithms or RNN architectures in the time between now and then might have effectively eliminated this issue.

To keep this story concrete, I’ll sketch out a broad kind of architectural modification that *might* help with this issue, if a number of technical details were hammered out. The basic idea is that Alex is equipped with a database that it can interact with to store and retrieve its “long-term memories.” We’ll imagine for concreteness that a “memory” is some subset of Alex’s activations and/or weights at a certain point in time, but it could have any other format.



At the beginning of an episode, Alex’s memory bank is wiped clean -- this is sort of analogous to a new “clone” of Alex being “born,” with certain “genetic predispositions” (in the form of its weights) but no “individual life experience” yet.

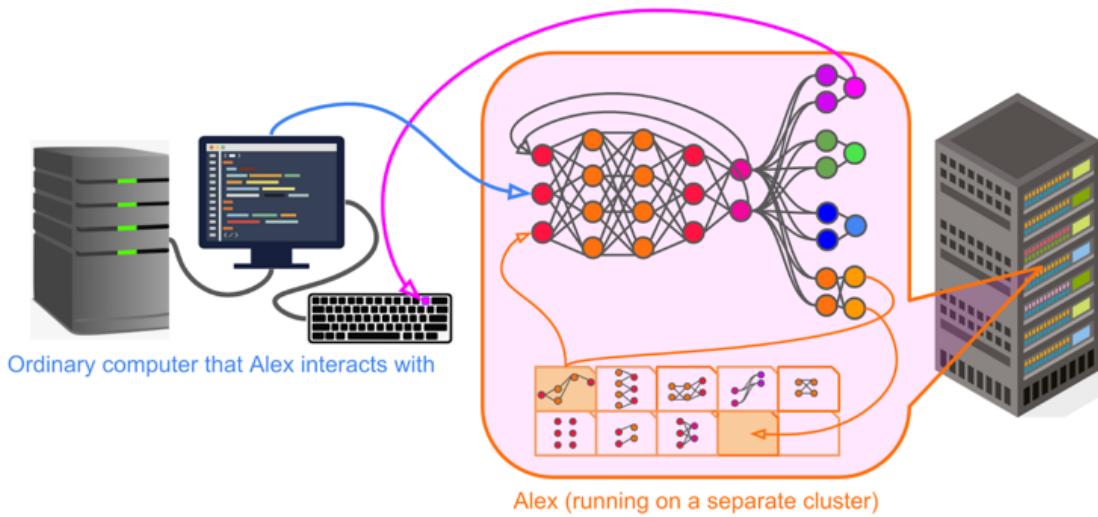
Within each timestep of an episode, Alex is able to a) store any activations and/or weights from the current timestep that it wants to, and b) send a query to retrieve any previously stored activations and/or weights that might be relevant, handing those to itself as inputs for the next timestep. Over the course of an episode, the memory bank builds up a store of “lifetime experiences” (though some experiences can also be “forgotten” if desired).

In addition to this dedicated memory bank, it’s important to note that **Alex can use the computer it’s interacting with as an external memory aid in exactly the same way humans do** -- it can write notes to itself in Google docs, create Asana tasks, set Google calendar reminders, and so on. If taking actions like that helps it to get a higher final reward for the episode, the training process will select for that.

This means that (like humans) Alex will have both “direct” experiential memories and access to external memory aids.

## Plausible high-level features of a good architecture

Stepping back, the high-level picture of Alex interacting with the external computer, with the details of the architecture sketched out, looks like this:



every timestep, Alex:

1. Takes as input:
  1. An observation from the external computer (such as an image of the screen)
  2. Its previous hidden state (via a recurrent connection)
  3. Some number of memories retrieved from the long-term memory database
2. Produces as output:
  1. An action (such as a keystroke). The action space includes a no-op and encompasses every low-level action a human could take on a computer.
  2. A store query and a fetch query to its long-term memory database, representing “memorizing” and “recalling” (respectively) a certain life experience.
  3. A prediction of the final reward for the episode, conditional on taking the action from a) and sending the store+fetch queries from b).
  4. A prediction of its next observation, conditional on taking the action from a) and sending the store+fetch queries from b).

I tried here to err on the side of being overly specific because I think it can be helpful to have in mind a vivid picture of a particular architecture that *could conceivably* suffice for a transformative model, even though it’s highly unlikely to literally be accurate.

With that said, I think that whatever future architecture we might use to train a transformative model with HFDT is pretty likely to have a few key high-level features in common with my over-simplified sketch:

- Some form of sequence processing (whether attention, recurrence, something else, or some combination of architectures)
- A way to reliably store and recall “memories” / “life experiences” over long timescales (this might be automatically solved by whatever solution is found for the above)

- Input and output formats which allow for a lot of flexibility, so that the model can process many different kinds of inputs and perform many different kinds of low-level actions (though this may be accomplished with a large number of different specialized input and output channels rather than a generic “universal” input channel and output channel as I’ve depicted here; we might also use entirely new data formats optimized for machine consumption)

**None of the basic points made in the main post hinge on the specifics of Alex’s architecture.** However, my arguments do rely on the high-level properties above -- e.g. that Alex has a very general and flexible input/output space and can usefully remember things across many timesteps. I think these properties are important for the case that Alex can potentially have a transformative impact in the first place, and by the same token are important for the case that Alex might be extremely dangerous.

1. [^](#)

Even though this vision doesn’t involve using exclusively human feedback, I’ll nonetheless refer to it as human feedback on diverse tasks (HFDT) to highlight that human judgments are central to training. In practice, I expect that powerful models will be trained with a combination of human feedback and automated reward signals (e.g. “did the code written by the AI compile?” or “how much computation did the AI use to perform this task?”). But I expect human feedback to be especially important, and I’ll focus on the dangers associated with the human feedback component of the feedback signal since there is more contention among the ML community about whether human feedback is dangerous. Many researchers agree that fully automated reward signals carry a high risk of misalignment, while being mostly optimistic about human feedback. While I agree incorporating human feedback is better than using fully-automated signals, I still think that the default version of human feedback is highly dangerous.

2. [^](#)

For example, we could imagine trying to produce transformative AI by training a large number of small models to do specialized tasks which we manually chain together, rather than training one large model to simultaneously do well at many tasks. We could also imagine using 100% automated reward signals, or “hand-programming” transformative AI rather than using any form of “training.”

3. [^](#)

For example, maybe the required model size is far too large to be affordable anytime soon, or we are unable to generate sufficiently challenging and diverse datasets and environments, or we run into optimization difficulties, etc.

4. [^](#)

By “incentivized,” I mean: “Alex would be rewarded more for playing the training game than it would be for alternative patterns of behavior.” There are further questions about whether Alex’s training procedure would *successfully find* this pattern of behavior; I believe it would, and discuss this more in the section “Maybe inductive bias or path dependence favors honest strategies?”

5. [^](#)

For example, lying about empirical questions which are politically-sensitive will often receive more reward than telling the truth; consider a version of Alex trained by the Catholic Church in the 1400s who is asked how the solar system works or how life was formed.

6. ^

What exactly “maximizing reward” means is unclear; however, I think this basic conclusion applies to most plausible ways of operationalizing this.

7. ^

This means that we’re bracketing scenarios in which Alex [“escapes from the box”](#) during the training phase; in fact I think such scenarios are plausible, which is a way in which risk is greater than this story represents.

8. ^

This could be accomplished through a variety of training signals -- e.g., more human feedback and automated signals, as well as more outcomes-based signals such as the compute efficiency of any chips they design, the performance of any new models they train, whether their scientific hypotheses are borne out by experiments, how much money they make for Magma per day, etc.

9. ^

It’d be reasonable to call this model “an AGI,” whether or not it’s able to do literally every possible task as well as humans; however, I’ll generally be avoiding that terminology in this post.

10. ^

Note that it may do this by first quickly building more powerful and more specialized AI systems who then develop these technologies.

11. ^

The exact timescale depends on many technical and economic factors, and is the subject of ongoing investigation at Open Philanthropy. Overall we expect that (absent deliberate intervention to slow things down) the time between “millions of copies of Alex are deployed” and “galaxy-scale civilization is feasible” is more likely to be on the order of 2-5 years rather than 10+ years.

12. ^

The most important possible exception is that if the world contains a number of models *almost* as intelligent as Alex at the time that Alex is developed, we may be able to use those models to help supervise Alex, and they may do a better job than humans would. However, I think it’s far from straightforward to translate this basic idea into a viable strategy for ensuring that Alex doesn’t play the training game, and (given that this strategy relies on quite capable models) it’s very unclear how much time there will be to try to pull it off before models get *too* capable to easily control.

13. ^

In reality, I expect it would have multiple input and output channels which use special format(s) more optimized for machine consumption, rather than having just one input channel which is directly analogous to human vision and one output channel which is directly analogous to human typing -- those are probably not the most efficient way for a neural network to interact with external computers. Going forward, I’ll simply refer to inputs as observations and outputs as actions, though I’ll use the shorthand of “screen image” and “keystroke” in the diagrams.

14. ^

Or mouse movement / click.

15. ^

Like today's neural networks, Alex is trained using [gradient descent](#). Roughly speaking, gradient descent repeatedly modifies an ML model to perform better and better at some task(s). Researchers feed Alex an instance of a task that they want it to perform well on, Alex produces some output in response, and gradient descent is used to slightly perturb the model so it performs a bit better on that example. This is repeated many times until Alex is very good on average at the kinds of tasks it saw during training. See [here](#) for a more technical explanation

16. ^

This tends to make it easier to train Alex to do more useful tasks down the road, so is often called "pre-training."

17. ^

If we imagine observations are images of screens and actions are keystrokes, the training dataset for this might come from screen captures and keylogs on Magma employees' laptops.

18. ^

As above, different models may be fine-tuned on different mixtures of these tasks in order to specialize into different niches.

19. ^

In [this appendix](#), I discuss a bit more why I think this strategy is a "baseline" and what some non-baseline strategies may look like.

20. ^

In terms of time, computation, and/or other metrics like "amount of human help used" or "number of lines in the proof."

21. ^

This is closely related to the concept of deceptive alignment, as introduced in [Hubinger et al 2019](#).

22. ^

Using more intelligent and thoughtful human reviewers would fix this particular problem, but not fix the general problem that Alex would be incentivized to play on their biases. See [this appendix](#) for more discussion.

23. ^

Magma may itself have stakeholders it needs to please that introduces more bias. For example, if one thing Alex is supposed to do is interact with customers, then it may need to be trained not to be offensive to those customers.

24. ^

An important family of training strategies involves using different copies of Alex (or different “heads” on the same model) each with incentives to point out lying and manipulation done by the other. There are a large number of possible approaches like this, but they all seem to require significant caution if we want to be sure they’ll work.

25. ^

The number of copies of Alex that can be run immediately will depend on how much computation it took to train it -- the more computation it took to train Alex, the more chips are available to run it once it’s trained. This multiple is likely to be large because training a model generally requires running it for a huge number of subjective years.

26. ^

For example, it could be given the ability to spend money to make trades, investments, or hires if that’s important for quickly increasing Magma’s capital; it could be given the ability to remotely operate machinery if that’s important for robotics research; etc.

27. ^

This could be accomplished through a variety of training signals -- e.g., more human feedback and automated signals, as well as more outcomes-based signals such as the compute efficiency of any chips they design, the performance of any new models they train, whether their scientific hypotheses are borne out by experiments, how much money they make for Magma per day, etc.

28. ^

In “softer takeoff” scenarios (which I consider more plausible overall), Magma will have already been using the somewhat-less-powerful predecessors of Alex to automate various parts of R&D before this point, meaning that the pace of research and innovation would already be going quickly by the time a model of Alex’s level of capability is deployed. This makes analysis more complicated, but on net I’m not convinced that it makes the situation significantly safer.

29. ^

Currently, the ratio of computation to memory in GPUs and TPUs is much higher than in brains. The [A100 GPU](#) (state-of-the-art as of March 2021) performs  $\sim 1.5 \times 10^{14}$  FLOP/s (in floating point 16 operations). [Joe Carlsmith’s central estimate](#) for the amount of computation it would take to match the human brain is  $\sim 10^{15}$  FLOP/s, implying that  $\sim 10$  A100 GPUs could match the total *computation* of a human brain. However, the human brain seems to *store*  $\sim 10^{14}$  bytes of memory (taking a synapse to be equivalent to a byte and assuming that synapses dominate the memory in the brain), while the A100 only has a storage capacity of  $\sim 8 \times 10^{10}$  bytes. That implies it would take  $\sim 1,000$  A100 GPUs to store a neural network whose parameters contained as much information as the human brain. That system of 1,000 GPUs would then have enough computation to use each parameter  $\sim 100$  times more often per second than the human brain uses each synapse. I expect computation intensity would continue to be significantly higher in ML chips than in biological brains, allowing for this speedup.

30. ^

In this case, Alex continues to be trained in the deployment setting, but the training signals are systematically different from the training signals in the lab setting due to humans’ decreased understanding and control.

31. ^

In deployment (more so than in the lab), it's not exactly clear what "maximizing reward" means precisely. But there will still be some reward signals coming into the various copies of Alex, and any plausible notion of "maximizing" those signals would be bad for humans.

32. ^

Or video, or written feedback, etc that is used to ultimately calculate reward.

33. ^

Or writing positive things, or appearing happy in front of a camera, etc.

34. ^

This feels more realistic if you consider that Alex itself would have written almost all of the code for the datacenter, and designed almost all of the hardware.

35. ^

That is, unless gradient descent for some reason "got stuck" at the benevolent strategy and was not able to push Alex toward the more effective strategy; I discuss this briefly [here](#).

36. ^

This is similar to the argument made in [Ilyas et al 2019](#) (*Adversarial examples are not bugs, they are features*).

37. ^

They could also respond by attempting to shut down Alex entirely, but this could be an extremely costly choice in the face of competitive pressure -- assuming they do not have an arbitrary lead on other AI companies, refraining from deploying Alex indefinitely is likely to simply mean that somebody else deploys a similar system (perhaps even a more dangerous one). Holden Karnofsky discusses how to balance "improving the safety of one's own models" against "ensuring that other actors do not deploy even-more-unsafe models" in his forthcoming posts about how we could address the issue I lay out in this post.

38. ^

See [Treacherous turns in the wild](#) (Luke Muehlhauser), [Objective robustness in deep reinforcement learning](#) (Koch et al), [The effects of reward misspecification](#) (Pan et al), [Specification gaming examples in AI](#) (DeepMind Safety blog post which links to an [aggregated spreadsheet](#)), [The surprising creativity of digital evolution](#) (Lehman et al), [this gwern blog post](#), and more. Many of these lists overlap with each other.

39. ^

We could certainly try including artificial scenarios designed to *appear as if* models could successfully take over from humans, but it seems very hard to be sure that models will be fooled by that and couldn't tell it apart from a genuine opportunity to take over. For example, models could look for signs that are highly likely to happen eventually but would be too expensive for humans to generate as a training example. As a simple example, Paul Christiano [suggested "someone factors an extremely large number"](#) (which would be too expensive to factor just to include as part of the training distribution, but would nonetheless likely get factored eventually).

40. ^

For example, stronger arguments that benign generalizations are especially “natural” for gradient descent, enough to make up for the fact that playing the training game would get higher reward; stronger arguments that Alex would “generalize habits” rather than “generalize goals” from training to deployment.

41. ^

Say the mice were dropped off on an earlier spaceship.

42. ^

Thanks to Buck Shlegeris for suggesting this analogy.

43. ^

For one thing, our mouse-brain-sized models have a different and more human profile of abilities than actual mice (e.g. they can talk).

44. ^

(who either deliver reward directly or write code that delivers reward)

45. ^

Humans would consider egregious reward-grabbing acts to be more blatant and scary violations of expectations than softer ways of playing to human biases, so if Alex is later caught it is likely to receive a large negative reward and/or trigger a larger retraining project. I discuss [below](#) why I think that isn’t likely to be sufficient to change its motives to grab reward.

46. ^

See [Turning reflection up to 11](#) for a similar proposal.

47. ^

For example, Eric Jang, citing [Connor Leahy](#) on Twitter, [writes](#): “Just asking the AI to be nice sounds flippant, but after seeing DALL-E and other large-scale multi-modal models that seem to *generalize better* as they get bigger, I think we should take these simple, borderline-naive ideas more seriously.”

48. ^

Note that Alex is likely to be motivated to maximize the “final recorded reward” even if it’s ultimately interested in pursuing some other goal. For example, if Alex is trying to have some kind of lasting effect on the world that extends beyond the current episode (e.g. “discover all the secrets of the universe”), it is probably much more likely to accomplish that goal if the future contains more models like it -- which in turn is more likely to happen if it gets a very high reward.

49. ^

Humans find lying difficult and unnatural due to our specific evolutionary history. Alex’s design and training wouldn’t necessarily replicate those kinds of evolutionary pressures. Still, it seems plausible that it would, especially since we’d be trying to give it feedback to encourage honesty.

50. ^

On a philosophical level, it's not exactly clear what it means to talk about what actions taken in deployment "would have" gotten a high reward. But on a practical level, Alex is likely to continue to receive rewards on the deployment distribution, particularly if it's not acting the way researchers want it to act, so the meaning of "reward on the deployment distribution" is fairly well-pinned-down.

51. ^

The high-level picture doesn't change if we imagine it taking in an attention window of a number of observations instead.

52. ^

This prediction probably won't be in the form of separately predicting each pixel of the next screen -- it's more likely to be something more complicated and abstract, e.g. an English-language description of the next screen.

53. ^

Although other known sequence processing architectures, like transformers, are also currently limited in how long they can "remember" things (albeit sometimes for different structural reasons).

54. ^

This may also reduce the risk that power-seeking misalignment emerges.

# What should you change in response to an "emergency"? And AI risk

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

---

Related to: [Slack gives you the ability to notice/reflect on subtle things](#)

Epistemic status: A possibly annoying mixture of straightforward reasoning and hard-to-justify personal opinions.

It is often stated (with some justification, IMO) that AI risk is an “emergency.” Various people have explained to me that they put various parts of their normal life’s functioning on hold on account of AI being an “emergency.” In the interest of people doing this sanely and not confusedly, I’d like to take a step back and seek principles around what kinds of changes a person might want to make in an “emergency” of different sorts.

## Principle 1: It matters what time-scale the emergency is on

There are plenty of ways we can temporarily increase productivity on some narrow task or other, at the cost of our longer-term resources. For example:

- Skipping meals
- Skipping sleep
- Ceasing to clean the house or to exercise
- Accumulating credit card debt
- Calling in favors from friends
- Skipping leisure time

If I would strongly prefer to address some situation  $x$  before time  $t$ , I may sometimes want to “borrow from the future” like this. But the time-scales matter. If I’m trying to address  $x$  as much as possible in the next five hours, skipping sleep may make sense. If I’m trying to address  $x$  as much as possible over the next year, I’ll probably do better to get my usual amount of sleep tonight. Something similar (with different, resource-specific timescales) will hold for other resources.

So, in short time-scale emergencies, it’ll often make sense to suspend a great deal of normal functioning for a short period of time. In longer time-scale emergencies, your life should mostly look closer to normal.

## Principle 2: It matters how much we know how to address the emergency

Much of what we do in daily life – especially when we’re feeling “free” and “unstuck,” and as though there is nothing in particular that we “have to” do – has the effect of making us well-resourced and capable in general. For example, by default, a lot of us would spend a lot of time reading interesting books of varied sorts, nerding out about interesting topics, trying our hands at new skills and crafts, etc. Also, we often like making our living spaces nicer (and more functional), forming new friendships, and so on.

If there is a particular thingy that matters hugely, and if you have an accurate model of how exactly to change that thingy, it may make sense to sacrifice some of your general-purpose capacities in trade for increased ability to address that thingy. (E.g., if you know for a fact that you’ll lose your home unless you pay the mortgage, and if keeping your home is important, it may make sense to trade general-purpose capacities for the ability to make mortgage payments by e.g. working [over-long](#) hours at a mind-numbing job that leaves you stupid.)

However, if you don’t have an accurate map of how to address a given thingy, then, even if the thingy is very important, and even if its time-scale is short, you’ll probably mostly want to avoid sacrificing general-purpose capacities. (In a low-information context, your general-purpose capacities are perhaps more likely to turn out helpful for your very important thingy than whatever you’d be trading them off to buy.) Thus, in “emergencies” where you do not have an accurate map of how to solve the emergency, your behavior should probably be more like normal than in better-mapped emergencies.

## Side-note: “Emergencies” as wake-up calls

A different way an “emergency” can sometimes rearrange priorities is by serving as a “wake-up call” that helps people peel away accidental accumulation of habits, “obligations,” complacency, etc. For example, I’m told that near encounters with death sometimes leave people suddenly in touch with what matters in life. (I’ve heard this from one friend who had cancer, and seen accounts from strangers in writing; I’m not sure how common this is or isn’t really.)

I non-confidently suspect some gravitate toward AI risk or other emergencies in the hopes that it’ll help them peel away the cruft, notice a core of caring within themselves, and choose activities that actually make sense. (See also: [Something to Protect](#).) If “AI risk as wake-up call” works out, I could imagine AI risk helping a person rearrange their actions in a way that boosts long-term capacities. (E.g., finding courage; trying new things and paying attention to what happens; cultivating friendships right now rather than in the possibly-non-existent future; facing up to minor social conflicts; attempting research that might yield fresh insights instead of research that feels predictable.)

This sort of “post wake-up call” change is almost the opposite of the kinds of “borrowing from the future” changes that typify short-term emergency responses. You should be able to tell the difference by seeing whether a person’s actions are unusually good for their long-term broad-based capacities (e.g., do they keep their house in a pattern that is good for them, get exercise, engage in the kinds of leisure and study that boost their ability to understand the world, appear unusually willing and able to cut through comfortable rationalizations, etc.?), or unusually bad for their

long-term broad-based capacities (e.g., do they live on stimulants, in a house no one would want to live in, while saying they ‘don’t have time’ for exercise or for reading textbooks and seeming kind of burn-out-y and as though they don’t have enough free energy to fully take an interest in something new, etc.?).

## **What sort of “emergency” is AI risk?**

It seems to me that AI risk is a serious problem in the sense that it may well kill us (and on my personal models, may well not, too). In terms of time-scales, I am pretty ignorant, but I personally will not be too surprised if the highest risk period is in only a couple years, nor if it is in more than thirty years. In terms of how accurate our maps of what to do are, it seems to me that our maps are not accurate; most people who are currently burning themselves out to try to help with AI risk on some particular path might, for all I know, contribute at least as well (even on e.g. a 5-year timescale) if they built general capacities instead.

I therefore mostly suspect that we’ll have our best shot at AI risk if we, as a community, cultivate long-term, robust capacities. (For me, this is hinging more on believing we have poor maps of how to create safety, and less on beliefs about time-scale.) Our best shot probably does mean paying attention to AI and ML advances, and directing some attention that way compared to what we’d do in a world where AI did not matter. It probably does mean doing the obvious work and the obvious alignment experiments where we know what those are, and where we can do this without burning out our long-term capacities. But it mostly doesn’t mean people burning themselves out, or depleting long-term resources in order to do this.

## **Some guesses at long-term, robust-in-lots-of-scenarios resources that may help with AI risk**

Briefly, some resources I’d like us to have, as AI approaches:

- Accurate trust in one another’s words. (Calibration, honesty, accurate awareness of one another’s honesty, valuing of truth over comfort or rationalizations. Practice seeing one another get things right and wrong in varied domains.)
- Integrity. The ability to reason, tell the truth, and do what matters in the face of pain, fear, etc.
- Practice having pulled off a variety of projects (not necessarily AI projects). (In my book, we get points for e.g. creating: movies; books; charter cities or other small or large-scale political experiments; buildings and communities and software; and basically anything else that involves reusable skills and engagement with the world.)
- Practice accomplishing things in groups.
- Time. Having AI not as advanced as in alternate scenarios. Having chip manufacture not as advanced as in alternate scenarios.
- Spiritual health. Ability to love, to care about that which matters, to laugh and let go of rationalizations, to hope and to try, to form deep friendships.

- Deep STEM knowledge, ability to do science of varied sorts, ability to do natural philosophy. (Not necessarily all AI.)
- Engineering skill and practice.
- An understanding of the wider cultural context we're in, and of how to interact with it and what to expect.
- All kinds of ML-related skills and resources, although this is in some tension with wanting time.

Given the above, I am excited about people in our larger orbit following their interests, trying to become scientists or writers or engineers or other cool things, exploring and cultivating. I am excited about people attempting work on AI alignment, or on other angles on AI safety, while also having hobbies and interests and friendships. For the most part I am not excited about people burning themselves out working super-long hours on alignment tasks in ways that damp their ability to notice new things or respond to later challenges, although I am excited about people pushing their limits doing work they're excited by, and these things can blur together.

Also, I am excited about people trying to follow paths to all of their long-term goals/flourishing, including their romantic and reproductive goals, and I am actively not excited about people deciding to shelve that because they think AI risk demands it. This is probably the hardest to justify of my opinions, but, like Divia [in this tweet](#), I am concerned (based partly on personal experience, partly on observations of others, and partly on priors/models) that when people try to table their deepest personal goals, this messes up their access to caring and consciousness in general.

## Why do we see burnout in AI safety efforts and in effective altruism?

IMO, I see burnout (people working hard at the expense of their long-term resources and capacities) more often than I expect is optimal for helping with AI risk. I'm not sure why. Some of it is in people who (IMO) have a better shot and a better plan than most for reducing AI risk, which makes it more plausibly actually-helpful for those folks to be doing work at the expense of long-term capacity, though I have my doubts even there. I suspect much of it is for confused/mistaken reasons, though; I sometimes see EAs burning themselves out doing e.g. random ML internships that they think they should work hard on because of AI risk, and AFAICT this trade does not make sense.

Also, when I look at the wider world (e.g. less-filtered chunks of Twitter; conversations with my Lyft drivers) I see lots of people acting as though lots of things are emergencies worth burning long-term resources for, in ways I suspect do not make sense, and I suspect that whatever dynamics lead to that may also be involved here.

I've also heard a number of people tell me that EA or AI safety efforts caused them to lose the ability to have serious hobbies, or serious intellectual interests, and I would guess this was harmful to long-term AI safety potential in most cases. (This "lots of people lose the ability to have serious hobbies when they find EA" is something I care very much about. Both as damage to our movements' potential, and as a symptom of a larger pattern of damage. I've heard this, again, from many, though not everyone.)

I'd be curious for y'all's takes on any of the above!

# Looking back on my alignment PhD

*This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on [Spotify](#), [Apple Podcasts](#), [Libsyn](#), and more.*

---

## On Avoiding Power-Seeking by Artificial Intelligence

A DISSERTATION PRESENTED  
BY  
ALEXANDER MATT TURNER  
TO THE SCHOOL OF  
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY  
IN THE SUBJECT OF  
COMPUTER SCIENCE

OREGON STATE UNIVERSITY  
CORVALLIS, OREGON  
MAY 2022

My [dissertation](#). It's long, so if you're going to read anything from it, read Chapter 0 (Introduction).

The funny thing about long periods of time is that they do, eventually, come to an end. I'm proud of what I accomplished during my PhD. That said, I'm going to first focus on mistakes I've made over the past four<sup>[1]</sup> years.

## Mistakes

I think I [got significantly smarter in 2018-2019](#), and kept learning some in 2020-2021. I was significantly less of a fool in 2021 than I was in 2017. That is important and worth feeling good about. But all things considered, I still made a lot of profound mistakes over the course of my PhD.

### Social dynamics distracted me from my core mission

I focused on "catching up" to other thinkers

*I figured this point out by summer 2021.*

I wanted to be more like Eliezer Yudkowsky and Buck Shlegeris and Paul Christiano. They know lots of facts and laws about lots of areas (e.g. general relativity and thermodynamics and information theory). I focused on building up dependencies (like [analysis](#) and [geometry](#) and [topology](#)) not only because I wanted to know the answers, but because I felt I owed a debt , that I was *in the red* until I could at least meet other thinkers at their level of knowledge.

But rationality is not about the bag of facts you know, nor is it about the concepts you have internalized. Rationality is about *how* your mind holds itself, it is *how* you weigh evidence, it is *how* you decide where to look next when puzzling out a new area.

If I had been more honest with myself, I could have nipped the "catching up with other thinkers" mistake in 2018. I could have removed the bad mental habits using [certain introspective techniques](#); or at least been aware of the badness.

But I did not, in part because the truth was uncomfortable. If I did not have a clear set of prerequisites (e.g. analysis and topology and game theory) to work on, I would not have a clear and immediate direction of improvement. I would have felt adrift.

But there is not yet any "rationality tech tree", no succession of well-defined rationality skills such that you can learn them in order and grow way stronger. Like, you can't just do the [calibration exercises](#), and then [the noticing-confusion exercises](#), and then other things. Those tools help, but they aren't enough. There won't be a clear and immediate direction of improvement, at first. But you may want to get stronger anyways.

## **I focused on seeming smart and defensible**

*I figured this point out [this spring](#).*

When I started working on alignment, I didn't know what to do at first, and I felt insecure about my credentials. As far as I remember, I figured I'd start off by becoming respected, since other people's feedback was initially a better guide than my own taste. Unfortunately, I didn't realize how deeply and subtly this goal would grow its roots.

I worried about upvotes, I worried about winning arguments, I worried about being defensible against criticism. I was so worried that someone would comment on one of my posts and tear everything down, because I *hadn't been careful enough*, because I had *left myself open* by not dotting all my 'i's. (Not that anyone has ever done that on LessWrong before...)

I think it was this year that I had my (second) "oh man, *don't forget the part where everyone is allowed to die to AI*" moment. To illustrate the new mindset this gut-realization gave me, I'll detail a recent decision with social consequences, and then compare the old and the new mindsets.

A few months back, Quintin Pope approached me with (what he claimed to be) a new alignment paradigm, which blossomed from asking the following kind of questions:

We clearly prefer future AIs to generalize in the way that neuroscientists generalize, so it seems worthwhile to ask: "why don't neuroscientists wirehead themselves?"

It's clearly not because humans evolved away from wireheading, *specifically*. There are somewhat similar situations to wireheading in the ancestral environment: psychoactive drugs, masturbation, etc. Is the reason we don't wirehead because evolution instilled us with an aversion to manipulating our reward function, which then zero-shot generalized to wireheading, despite wireheading being so wildly dissimilar to the contents of the

ancestral environment? How could evolution have developed an alignment approach that generalized so well?

After a few days, I realized my gut expectations were that he was broadly correct and that this theory of alignment could actually be right. However, I realized I wasn't consciously letting myself think that because it would be Insufficiently Skeptical to actually think the alignment problem is solvable. This seemed obviously stupid to me, so I quickly shut that line of thinking down and second-order updated towards optimism so that I would [stop predictably getting more optimistic](#) about Quintin's theory.<sup>[2]</sup>

I realized I assigned about 5% credence to "this line of thinking marks a direct and reasonably short path to solving alignment." Thus, on any calculation of benefits and harms, I should be willing to stake some reputation to quickly get more eyeballs on the theory, even though I expected to end up looking a little silly (with about 95% probability). With my new attitude, I decided "whatever, let's just get on with it and stop wasting time."

The old "don't leave any avenue of being criticized!" attitude would have been less loyal to my true beliefs: "This *could* work, but there are so many parts I don't understand yet. If I figure those parts out first, I can explain it better and avoid having to go out on a limb in the process." Cowardice and social anxiety, dressed up as prudence and skepticism.

I still get anxious around disagreements with people I respect. I am still working on fully expunging the "defensibility" urges, because they suck. But I've already made a lot of progress.<sup>[3]</sup>

## Too much deference, too little thinking for myself

*I realized and started fixing this mistake [this spring](#). (Seeing a pattern?)*

I filtered the world through a status lens. If I read a comment from a high-status person, I would gloss over confusing parts, because I was probably the one reading it wrong. Sure, I would verbally agree that [modest epistemology](#) is unproductive. I just *happened* to not think thoughts like "[high-status person]'s claim seems obviously dumb and wrong."

Now I let myself think thoughts like that, and it's great. For example, last week I was reading about Pavlov's conditioning experiments with dogs. I read the following:

Pavlov (1902) started from the idea that there are some things that a dog does not need to learn. For example, dogs don't learn to salivate whenever they see food. This reflex is 'hard-wired' into the dog.

I thought, "that seems like bullshit. Really, the dogs are *hard-wired* to salivate when they **see** food? Doesn't that require *hard-wiring a food-classifier into the dog's brain?*"

And you know what? It *was* bullshit. I searched for about 8 minutes before finding references of [the original lectures Pavlov gave](#):

Dr. Zitovich took several young puppies away from their mother and fed them for considerable time only on milk. When the puppies were a few months old he established fistulae of their salivary ducts, and was thus able to measure accurately the secretory activity of the glands. **He now showed these puppies some solid food -- bread or meat -- but no secretion of saliva was evoked.**

Our world is so inadequate that seminal psychology experiments are described in mangled, misleading ways. Inadequacy abounds, and status only weakly tracks adequacy. Even if the high-status person belongs to your in-group. Even if all your smart friends are nodding along.

Would you notice if *this very post* were inadequate and misleading? Would it be bullshit for the dog-genome to hardwire a food-classifier? Think for yourself. Constant vigilance!

## Non-social mistakes

### I thought about comfortable, familiar problems

*I figured this point out this spring, because I bumped into Quintin as described above.*

I remember a sunny summer day in 2019, sitting in the grass with Daniel Filan at UC Berkeley. He recommended putting together an end-to-end picture of the alignment problem. I remember feeling pretty uncomfortable about that, feeling that I wouldn't understand which alignment problems go where in my diagram ("do embedded agency failures crop up *here*, or *there*?"). Wouldn't it just make more sense to read more alignment papers and naturally refine those views over time?

This was a rationalization, plain and simple. There is no point where you feel ready to put all the pieces together. If you feel totally comfortable about how alignment fits together such that Daniel's exercise does not *push you* on some level, we have either *already* solved the alignment problem, or you are deluded.

I did not feel ready, and I was not ready, and I should have done it anyways. But I focused on more comfortable work with well-defined boundaries, because it felt good to knock out new theorems. Whether or not those theorems were useful and important to alignment, that was a mistake. So I stayed in my alignment comfort zone. I should have stopped working on impact measures and power-seeking way earlier than I did, even though I did end up doing some cool work.

### Not admitting to myself that I thought alignment was doomed

*Figured this out this spring. I'm not sure if I've fixed the general error yet.*

After I became more optimistic about alignment due to having a sharper understanding of the overall problem and of how human values formed to begin with, I also became more pessimistic about *other* approaches, like IDA/ELK/RRM/AUP/[anything else with a three-letter acronym]. But my new understanding didn't seem to present any *specific* objections. So why did I suddenly feel worse about these older ideas?

I suspect that part of the explanation is: I hadn't wanted to admit how confused I was about alignment, and I (implicitly) clutched to "but it *could* work"-style hopefulness. But now that I had a *different* reason to hope, resting upon a more solid and mechanistic understanding, now it was apparently emotionally safe for me to admit I didn't have much hope at all for the older approaches.

Yikes.

If that's what happened, I was seriously deluding myself. I will do better next time.

### I viewed my life through narratives

*I probably figured this point out in 2021.*

Back in 2018, I had the "upstart alignment researcher" narrative—starting off bright-eyed and earnest, learning a lot, making friends. But then I hurt my hands and couldn't type anymore, which broke the narrative. I felt dejected—to slightly exaggerate, I felt I had fallen off of the sunlit path, and now nothing was going to go as it should.

Another example of narrative-thinking is when people say "I'm just not a math person." This is an *inference* and a *story* they tell themselves. Strictly speaking, they may not know much math, and they may not enjoy math, and they may not see how to change either of those facts. But the *narrative* is that they are not a math person. Their discomfort and their aversion-to-trying stem not just from their best-guess assessment of their own weaknesses, but from a *story* they are living in.

Every moment is an opportunity for newly-directed action. [Keep your identity small](#) and keep the narratives in the story-books. At least, if you want to use narratives, carefully introspect to make sure you're using them, and they aren't using you.

## Other helpful habits I picked up

I'm not really sure where these two habits go, so I'll put them here. I wish I'd had these skills in 2018.

- **Distinguish between *observations* and *inferences*.** When people speak to you, mark their arguments as *observations* or as *inferences*. Keep the types *separate*. I've gained *so much* from this simple practice.

Here are two cases I've recently found where people seem to mistake the folk wisdom for observation:

- "People often say they're afraid to die" is an *observation*, and "people are hard-wired to be afraid of death" is an *inference*.
- "I often feel 'curiosity' and some kind of exploration-impulse" is an *observation*, and "people are innately curious" is an *inference*.
- **Be concrete.** My friend Kurt remarks that I constantly ask for examples.
  - If a friend comes to me for advice and says "I'm terrible at dating, I just feel so shy!", I *could* say "You're really fun to be around, you're probably just in your head too much", and then *they* could say "Agh, maybe, but it's just so frustrating." Wouldn't that just be such a useful conversation for them? That'll *definitely* solve their awkwardness!
    - Alternatively, if I *ask for an example*, we can both analyze an event which *actually happened*. Perhaps they say, "I met a girl named Alice at the party, but I somehow ran out of things to say, and it got quiet, and we found excuses to part ways." Then I can help my friend introspect and figure out why they didn't have anything to say, which *is in fact a question with a real answer*.
  - The general rhythm is: Bind your thinking to *coherent scenarios* (preferably ones which *actually happened*, like meeting a girl named Alice), so that you (and possibly other people) can explore the details together (like why it got quiet) in order to figure out what to change (like running mock encounters to shoo away the social anxiety).
  - On the other hand, if you can't think of a concrete example to ground your airy words, maybe your thinking is totally untethered from reality. Maybe your assumptions are *contradictory* and you can't even see it.
    - Here's something I recently said on Discord:

"If there are some circuits who can defer to the market prediction, then each circuit can get their coalitional contribution as their fixed weight. This lets some relatively simpler circuits retain weight. At least, those are the abstract words I want to say, but now I feel confused about how to apply that to a concrete example for how e.g. a shallow but broad "don't steal" value negotiates via [Critch-bargaining](#). **Not being able to give a concrete example means I don't really know what I'm talking about here.**"

- Don't tell me how your alignment strategy will e.g. "faithfully reproduce human judgments." Explain [what concrete benefits you hope to realize](#), and why "faithful reproduction of human judgments" will realize those benefits.
  - If the actual answer is that you *don't know*, then just *say it*, because it's the truth. Be aware that you don't know.

To close out the "Mistakes" section, I mostly wish I'd expected more from myself. I wish I'd believed myself capable of building an end-to-end picture of the alignment problem, of admitting what I didn't know and what I hadn't thought about, of being able to survive/ignore the harsh winds of criticism and skepticism.

I did these things eventually, though, and I'm proud of that.

## What I'm proud of

1. I [didn't keep working on computational chemistry](#). Boy howdy, would that have been awful for me. *Thank you, TurnTrout2018!*
  1. I remember thinking "You know what, I'd rather get *expelled* than not do [the 2018 CHAI internship]." This thought [gave me the courage](#) to find a new advisor who would let me work on AI safety, funding be damned.
  2. I'm not a natural nonconformist. Conflict makes me nervous. I've had to work for it.
2. I [learned a lot of math](#), even though I felt sheepish and insecure about it at first.
3. I think I ended up achieving rationality escape velocity.
  1. When I get stuck / feel depressed, errors get thrown, exception-handling activates, I start thinking "these thoughts seem unreasonably dark; my cognition is compromised; have I eaten enough food today, have I drank enough water, should I call a friend...".
  2. When I get stuck on a problem (e.g. what is the type signature of human values?), I do not stay stuck. I notice I am stuck, I run down a list of tactics, I explicitly note what works, I upweight that for next time.
  3. When I realize I've been an idiot about something (e.g. nicking my hand with a knife, missing a deadline), I stop and think *wow, that was stupid, what's the more general error I'm making?*
  4. The general rhythm is: I feel agentic and capable and self-improving, and these traits are strengthening over time, as is the rate of strengthening.
  5. This definitely didn't have to happen, but I made it happen (with the help of some friends and resources).
4. Research achievements:
  1. I think [Reframing Impact](#) correctly inferred our intuitions around what "impact" means, and also that sequence was beautiful and I loved making it.
  2. [My dissertation](#) is also beautiful. I painstakingly wrote and formatted and edited it, even hiring a professional to help out. I fought to keep its tone focused on what matters: the sharp dangers of AGI.
  3. I likewise poured myself into [Optimal Policies Tend To Seek Power](#), and its follow-up, [Parametrically Retargetable Decision-Makers Tend To Seek Power](#).
    1. First, I had felt instrumental convergence should be provable and formally understandable. It was a mystery to me in 2019, and now it's not.
    2. Second, I used to suck at writing academic papers, but I managed to get two NeurIPS spotlights by the end of my program. NeurIPS spotlights might not save the world, but that was tough and I did a good job with it.
  4. [Attainable utility preservation](#) is pointless for AGI alignment, but *damn* is it cool that we could do unsupervised learning to get a reward function, preserve the agent's ability to optimize that single random objective, and [just get cautious behavior in complicated environments](#).

# Looking forward

Leaving Oregon was a bit sad, but coming to Berkeley is exciting. I'll be starting my CHAI postdoc soon. I'm working with lots of cool, smart, loyal friends. I'm feeling strong and confident and relatively optimistic, both about alignment and about my personal future.

[Here's to winning.](#) 

1. ^

My PhD was six years long (it started in the fall of 2016). However, I'm not even going to critique the first two years, because that would make the "Mistakes" section far too long.

2. ^

If you're interested in reading about the theory now, see [this recent comment](#). I'm currently putting together some prerequisite posts to bridge the inferential gap.

3. ^

Sometimes I feel the urge to defend myself *just a little more*, to which some part of me internally replies "are you serious, this defensibility thing again?! Are you ever going to let me *actually think*?"

I like that part of me a lot.

# Reward is not the optimization target

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

*This insight was made possible by many conversations with Quintin Pope, where he challenged my implicit assumptions about alignment. I'm not sure who came up with this particular idea.*

In this essay, I call an agent a “reward optimizer” if it not only gets lots of reward, but if it reliably makes choices like “reward but no task completion” (e.g. receiving reward without eating pizza) over “task completion but no reward” (e.g. eating pizza without receiving reward). Under this definition, an agent can be a reward optimizer even if it doesn’t contain an explicit representation of reward, or implement a search process for reward.

Reinforcement learning is learning what to do—how to map situations to actions **so as to maximize a numerical reward signal**. — [Reinforcement learning: An introduction](#)

Many people<sup>[1]</sup> seem to expect that reward will be the optimization target of really smart learned policies—that these policies will be reward optimizers. I strongly disagree. As I argue in this essay, reward is *not*, in general, that-which-is-optimized by RL agents.<sup>[2]</sup>

Separately, as far as I can tell, most<sup>[3]</sup> practitioners usually view reward as encoding the relative utilities of states and actions (e.g. it’s *this good* to have all the trash put away), as opposed to imposing a *reinforcement schedule* which builds certain computational edifices inside the model (e.g. reward for picking up trash → reinforce trash-recognition and trash-seeking and trash-putting-away subroutines). I think the former view is usually inappropriate, because in many setups, **reward chisels cognitive grooves into an agent**.

Therefore, *reward is not the optimization target* in two senses:

1. Deep reinforcement learning agents will not come to intrinsically and primarily value their reward signal; reward is not *the trained agent’s optimization target*.
2. Utility functions express the *relative goodness* of outcomes. Reward is *not best understood* as being a kind of utility function. Reward has the mechanistic effect of *chiseling cognition into the agent’s network*. Therefore, properly understood, reward does not express relative goodness and is therefore *not an optimization target at all*.

## Reward probably won’t be a deep RL agent’s primary optimization target

After work, you grab pizza with your friends. You eat a bite. The taste releases [reward in your brain](#), which triggers credit assignment. Credit assignment identifies which thoughts and decisions were responsible for the release of that reward, and makes

those decisions more likely to happen in similar situations in the future. Perhaps you had thoughts like

- “It’ll be fun to hang out with my friends” and
- “The pizza shop is nearby” and
- “Since I just ordered food at a cash register, execute motor-subroutine-#51241 to take out my wallet” and
- “If the pizza is in front of me and it’s mine and I’m hungry, raise the slice to my mouth” and
- “If the slice is near my mouth and I’m not already chewing, take a bite.”

Many of these thoughts will be judged responsible by credit assignment, and thereby become more likely to trigger in the future. This is what *reinforcement learning* is all about—the reward is the *reinforcer* of those things which came before it and the *creator* of new lines of cognition entirely (e.g. anglicized as “I shouldn’t buy pizza when I’m mostly full”). The reward chisels cognition which increases the probability of the reward accruing next time.

Importantly, reward does not automatically spawn thoughts *about* reward, and reinforce those reward-focused thoughts! Just because common English endows “reward” with suggestive pleasurable connotations, that [does not mean that](#) an RL agent will *terminally value* reward!

What kinds of people (or non-tabular agents more generally) will become reward optimizers, such that the agent ends up terminally caring about reward (and little else)? Reconsider the pizza situation, but instead suppose you were thinking thoughts like “this pizza is going to be so rewarding” and “in this situation, eating pizza sure will activate my reward circuitry.”

You eat the pizza, triggering reward, triggering credit assignment, which correctly locates these reward-focused thoughts as contributing to the release of reward.

Therefore, in the future, you will more often take actions because you think they will produce reward, and so you will become more of the kind of person who intrinsically cares about reward. This is a path[\[4\]](#) to reward-optimization and wireheading.

While it’s possible to have activations on “pizza consumption predicted to be rewarding” and “execute motor-subroutine-#51241” and then have credit assignment hook these up into a new motivational circuit, **this is only one possible direction of value formation in the agent**. Seemingly, the most direct way for an agent to become *more* of a reward optimizer is to *already* make decisions motivated by reward, and then have credit assignment further generalize that decision-making.

## The siren-like suggestiveness of the word “reward”

Let’s strip away the suggestive word “reward”, and replace it by its substance: cognition-updater.

Suppose a human trains an RL agent by pressing the cognition-updater button when the agent puts trash in a trash can. While putting trash away, the AI’s policy network is probably “thinking about”[\[5\]](#) the *actual world it’s interacting with*, and so the cognition-updater reinforces those heuristics which lead to the trash getting put away

(e.g. “if trash-classifier activates near center-of-visual-field, then grab trash using `motor-subroutine-#642`”).

Then suppose this AI models the true fact that the button-pressing produces the cognition-updater. Suppose this AI, which has historically had its trash-related thoughts reinforced, considers the plan of pressing this button. “If I press the button, that triggers credit assignment, which will reinforce my decision to press the button, such that in the future I will press the button even more.”

*Why, exactly, would the AI seize [6] the button? To reinforce itself into a certain corner of its policy space? The AI has not had antecedent-computation-reinforcer-thoughts reinforced in the past, and so its current decision will not be made in order to acquire the cognition-updater!*

RL is not, in general, about training cognition-updater optimizers.

## When is reward the optimization target of the agent?

If reward is guaranteed to become your optimization target, then your learning algorithm can force you to become a drug addict. Let me explain.

[Convergence theorems](#) provide conditions under which a reinforcement learning algorithm is guaranteed to converge to an optimal policy for a reward function. For example, value iteration maintains a table of value estimates for each state  $s$ , and iteratively propagates information about that value to the neighbors of  $s$ . If a far-away state  $f$  has huge reward, then that reward ripples back through the environmental dynamics via this [“backup” operation](#). Nearby parents of  $f$  gain value, and then after lots of backups, far-away ancestor-states gain value due to  $f$ 's high reward.

Eventually, the “value ripples” settle down. The agent picks an (optimal) policy by acting to maximize the value-estimates for its post-action states.

Suppose it would be extremely rewarding to do drugs, but those drugs are on the other side of the world. Value iteration backs up that high value to your present space-time location, such that your policy necessarily gets *at least* that much reward. There's no escaping it: After enough backup steps, you're traveling across the world to do cocaine.

But obviously these conditions aren't true in the real world. Your learning algorithm doesn't force *you* to try drugs. Any AI which e.g. tried every action at least once would quickly kill itself, and so real-world general RL agents won't explore like that because that would be stupid. So the RL agent's algorithm won't make it e.g. explore wireheading either, and so the convergence theorems *don't apply even a little—even in spirit*.

## Anticipated questions

1. Why won't early-stage agents think thoughts like “If putting trash away will lead to reward, then execute `motor-subroutine-#642`”, and then this gets reinforced into reward-focused cognition early on?

1. Suppose the agent puts away trash in a blue room. Why won't early-stage agents think thoughts like "If putting trash away will lead to the wall being blue, then execute motor-subroutine-#642", and then this gets reinforced into blue-wall-focused cognition early on? [Why consider either scenario to begin with?](#)
2. But aren't we implicitly selecting for agents with high cumulative reward, when we train those agents?
  1. Yeah. But on its own, this argument can't possibly imply that selected agents will probably be reward optimizers. The argument would [prove too much](#). Evolution selected for inclusive genetic fitness, and it [did not get IGF optimizers](#).
    1. "We're selecting for agents on reward → we get an agent which optimizes reward" is locally invalid. "We select for agents on X → we get an agent which optimizes X" is not true for the case of evolution, and so is not true in general.
    2. Therefore, the argument isn't necessarily true in the AI reward-selection case. Even if RL *did* happen to train reward optimizers and this post were wrong, the selection argument is too weak on its own to establish that conclusion.
  2. Here's the more concrete response: Selection isn't just for agents which get lots of reward.
    1. For simplicity, consider the case where on the training distribution, the agent gets reward if and only if it reaches a goal state. Then any selection for reward is also selection for reaching the goal. And if the goal is the only red object, then selection for reward is *also* selection for reaching red objects.
    2. In general, selection for reward produces equally strong selection for reward's necessary and sufficient conditions. In general, it seems like there should be a lot of those. Therefore, since selection is not only for reward but for *anything which goes along with reward* (e.g. reaching the goal), then selection won't advantage *reward optimizers over agents which reach goals quickly / pick up lots of trash / [do the objective]*.
  3. Another reason to not expect the selection argument to work is that it's *instrumentally convergent* for most inner agent values to *not* become wireheaders, for them to *not* try hitting the reward button.
    1. I think that before the agent can hit the particular attractor of reward-optimization, it will hit an attractor in which it optimizes for some aspect of a historical correlate of reward.
      1. We train agents which intelligently optimize for e.g. putting trash away, and this reinforces the trash-putting-away computations, which activate in a broad range of situations so as to steer agents into a future where trash has been put away. An intelligent agent will model the true fact that, if the agent reinforces itself into caring about cognition-updating, then it will no longer navigate to futures where trash is put away. Therefore, it decides to not hit the reward button.
      2. This reasoning follows for most inner goals by instrumental convergence.
    2. On my current best model, this is why people usually don't wirehead. They learn their own values via deep RL, like caring about dogs, and

these actual values are opposed to the person they would become if they wirehead.

3. Don't some people terminally care about reward?
  1. I think so! I think that generally intelligent RL agents will have *secondary, relatively weaker* values around reward, but that reward will not be a primary motivator. Under my current (weakly held) model, an AI will only start chiseled computations about reward *after* it has chiseled other kinds of computations (e.g. putting away trash). More on this in later essays.
4. But what if the AI bops the reward button early in training, while exploring? Then credit assignment would make the AI more likely to hit the button again.
  1. Then keep the button away from the AI until it can model the effects of hitting the cognition-updater button. [\[7\]](#)
  2. For the reasons given in the "siren" section, a sufficiently reflective AI probably won't seek the reward button on its own.
5. AIXI—
  1. will always kill you and then wirehead forever, unless you gave it something like a constant reward function.
  2. And, IMO, this fact is not practically relevant to alignment. AIXI is *explicitly a reward-maximizer*. As far as I know, AIXI(-t) is not the limiting form of any kind of real-world intelligence trained via *reinforcement learning*.
6. Does the choice of RL algorithm matter?
  1. For point 1 (*reward is not the trained agent's optimization target*), it might matter.
    1. I started off analyzing model-free actor-based approaches, but have also considered a few model-based setups. I think the key lessons apply to the general case, but I think the setup will substantially affect which values tend to be grown.
      1. If the agent's curriculum is broad, then reward-based cognition may get reinforced from a confluence of tasks (solve mazes, write sonnets), while each task-specific cognitive structure is only narrowly contextually reinforced. That said, this is also selecting equally hard for agents which do the rewarded activities, and reward-motivation is only one possible value which produces those decisions.
      2. Pretraining a language model and then slotting that into an RL setup also changes the initial computations in a way which I have not yet tried to analyze.
    2. It's *possible* there's some kind of RL algorithm which *does* train agents which limit to reward optimization (and, of course, thereby "solves" inner alignment in its literal form of "find a policy which optimizes the outer objective signal").
  2. For point 2 (*reward provides local updates to the agent's cognition via credit assignment; reward is not best understood as specifying our preferences*), the choice of RL algorithm should not matter, as long as it uses reward to compute local updates.
    1. A similar lesson applies to the updates provided by loss signals. A loss signal provides updates which deform the agent's cognition into a new shape.
7. TurnTrout, you've been talking about an AI's learning process using English, but ML gradients may not neatly be expressible in our concepts. How do we know that it's appropriate to speculate in English?
  1. I am *not certain* that my model is legit, but it sure seems more legit than (my perception of) how people usually think about RL (i.e. in terms of

reward maximization, and reward-as-optimization-target instead of as feedback signal which builds cognitive structures).

2. I only have access to my own concepts and words, so I am provisionally reasoning ahead anyways, while keeping in mind the potential treacheries of anglicizing imaginary gradient updates (e.g. "be more likely to eat pizza in similar situations").

## Dropping the old hypothesis

At this point, I don't see a strong reason to focus on the "reward optimizer" hypothesis. The idea that AIs will get really smart and primarily optimize some reward signal... I don't know of any tight mechanistic stories for that. I'd love to hear some, if there are any.

As far as I'm aware, the strongest evidence left for agents intrinsically valuing cognition-updating is that some humans *do* strongly (but not uniquely) value cognition-updating, [8] and many humans seem to value it weakly, and humans are probably RL agents in the appropriate ways. So we definitely can't *rule out* agents which strongly (and not just weakly) value the cognition-updater. But it's also *not* the overdetermined default outcome. More on that in future essays.

It's true that reward *can* be an agent's optimization target, but what reward *actually does* is reinforce computations which lead to it. A particular alignment proposal might argue that a reward function will *reinforce the agent into a shape such that it intrinsically values reinforcement*, and that the *cognition-updater goal is also a human-aligned optimization target*, but this is still just one particular approach of using the cognition-updating to produce desirable cognition within an agent. Even in that proposal, the primary mechanistic function of reward is reinforcement, not optimization-target.

## Implications

Here are some major updates which I made:

1. **Any reasoning derived from the reward-optimization premise is now suspect until otherwise supported.**
2. **Wireheading was never a high-probability problem for RL-trained agents**, absent a specific story for why cognition-updater-acquiring thoughts would be chiseled into primary decision factors.
3. **Stop worrying about finding "outer objectives" which are safe to maximize.** [9] I think that you're not going to get an outer-objective-maximizer (i.e. an agent which maximizes the explicitly specified reward function).
  1. Instead, focus on building good cognition within the agent.
  2. In my ontology, there's only one question: How do we grow good cognition inside of the trained agent?
4. **Mechanistically model RL agents as executing behaviors downstream of past reinforcement** (e.g. putting trash away), in addition to thinking about policies which are selected for having high reward on the training distribution (e.g. hitting the button).

1. The latter form of reasoning skips past the mechanistic substance of reinforcement learning: The chiseling of computations responsible for the acquisition of the cognition-updater. I still think it's useful to consider selection, but mostly in order to generate failure modes whose mechanistic plausibility can be evaluated.
2. In my view, reward's proper role isn't to encode an objective, but a *reinforcement schedule*, such that the right kinds of computations get reinforced within the AI's mind.

*Edit 11/15/22:* The original version of this post talked about how reward reinforces antecedent computations in policy gradient approaches. This is not true in general. I edited the post to instead talk about how reward is used to upweight certain kinds of actions in certain kinds of situations, and therefore reward *chisels cognitive grooves into agents*.

## Appendix: The field of RL thinks reward=optimization target

Let's take a little stroll through [Google Scholar's top results for "reinforcement learning"](#), emphasis added:

The agent's job is to find a policy... that **maximizes some long-run measure of reinforcement**. ~ [Reinforcement learning: A survey](#)

In instrumental conditioning, animals learn to choose actions to obtain rewards and avoid punishments, or, more generally to achieve goals. **Various goals are possible, such as optimizing the average rate of acquisition of net rewards (i.e. rewards minus punishments), or some proxy for this such as the expected sum of future rewards.** ~ [Reinforcement learning: The Good, The Bad and The Ugly](#)

Steve Byrnes did, in fact, briefly point out part of the “reward is the optimization target” mistake:

I note that even experts sometimes sloppily talk as if RL agents make plans towards the goal of maximizing future reward... — [Model-based RL, Desires, Brains, Wireheading](#)

I don't think it's just sloppy talk, I think it's incorrect belief in many cases. I mean, I did my PhD on RL theory, and I still believed it. Many authorities and textbooks confidently claim—presenting little to no evidence—that reward is an optimization target (i.e. the quantity which the policy is in fact trying to optimize, or the quantity to be optimized by the policy). [Check what the math actually says](#).

1. ^

[Including](#) the authors of the quoted introductory text, [Reinforcement learning: An introduction](#), I have, however, met several alignment researchers who already internalized that reward is not the optimization target, perhaps not in so many words.

2. ^

[Utility ≠ Reward](#) points out that an RL-trained agent is *optimized by* original reward, but not necessarily *optimizing for* the original reward. This essay goes further in several ways, including when it argues that *reward* and *utility* have different type signatures—that reward shouldn't be viewed as encoding a goal at all, but rather a *reinforcement schedule*. And not only do I not expect the trained agents to not maximize the original “outer” reward signal, I think they probably won't try to strongly optimize [any reward signal](#).

3. ^

[Reward shaping](#) seems like the most prominent counterexample to the “reward represents terminal preferences over state-action pairs” line of thinking.

4. ^

But also, you were still probably thinking about reality as you interacted with it (“since I'm in front of the shop where I want to buy food, go inside”), and credit assignment will still locate some of those thoughts as relevant, and so you wouldn't purely reinforce the reward-focused computations.

5. ^

“Reward reinforces existing thoughts” is ultimately a claim about how updates depend on the existing weights of the network. I think that it's easier to update cognition along the lines of existing abstractions and lines of reasoning. If you're already running away from wolves, then if you see a bear and become afraid, you can be updated to run away from large furry animals. This would leverage your *existing* concepts.

From [A shot at the diamond-alignment problem](#):

The local mapping from gradient directions to behaviors is given by the neural tangent kernel, and the learnability of different behaviors is given by the NTK's eigenspectrum, which [seems to adapt to the task at hand](#), making the network quicker to learn along behavioral dimensions similar to those it has already acquired.

6. ^

Quintin Pope remarks: “The AI would probably want to establish **control** over the button, if only to ensure its values aren't updated in a way it wouldn't endorse. Though that's an example of convergent powerseeking, not reward seeking.”

7. ^

For mechanistically similar reasons, keep cocaine out of the crib until your children can model the consequences of addiction.

8. ^

I am presently ignorant of [the relationship between pleasure and reward prediction error in the brain](#). I do not think they are the same.

However, I think people are usually weakly hedonically / experientially

motivated. Consider a person about to eat pizza. If you give them the choice between "pizza but no pleasure from eating it" and "pleasure but no pizza", I think most people would choose the latter (unless they were really hungry and needed the calories). If people just navigated to futures where they had eaten pizza, that would not be true.

9. [^](#)

From correspondence with another researcher: There may yet be an interesting alignment-related puzzle to "Find an optimization process whose maxima are friendly", but I personally don't share the intuition yet.

# Changing the world through slack & hobbies

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

*(Also posted on EA Forum)*

## Introduction

In EA orthodoxy, if you're really serious about EA, the three alternatives that people most often seem to talk about are

- (1) "direct work" in a job that furthers a very important cause;
- (2) "[earning to give](#)";
- (3) earning "[career capital](#)" that will help you do those things in the future, e.g. by getting a PhD or teaching yourself ML.

By contrast, there's not much talk of:

**(4) being in a job / situation where you have extra time and energy and freedom to explore things that seem interesting and important.**

But that last one is really important!

## Examples

For example, here are a bunch of things off the top of my head that look like neither "direct work" nor "earning-to-give" nor "earning career capital":

- David Denkenberger was a professor of mechanical engineering. As I understand it (see [here](#)), he got curious about food supplies during nuclear winter, and started looking into it in his free time. One thing led to another, and he now leads [ALLFED](#), which is doing very important and irreplaceable work. (Denkenberger [seems](#) to have had no prior formal experience in this area.)
- I'm hazy on the details, but I believe that Eliezer Yudkowsky and Nick Bostrom developed much of their thinking about AGI & superintelligence via discussions on online mailing lists. I doubt they were being paid to do that!
- Meanwhile, Stuart Russell got really into AGI safety / alignment during a sabbatical.
- The precursor to GiveWell was a "charity club" started by Holden Karnofsky and Elie Hassenfeld, where they and other employees at their hedge fund "pooled in money and investigated the best charities to donate the money to" ([source](#)), presumably in their free time.
- I mean seriously, pretty much anytime anybody anywhere has ever started something really new, they were doing it in their free time before they were paid for it.

## Three ingredients to a transformative hobby

## Ingredient 1: Extra time / energy / slack

Honestly, I wasn't really sure whether to put it on the list at all. Scott Alexander famously did some of his best writing during a medical residency—not exactly a stage of life where one has a lot of extra free time. (See his discussion [here](#).) Another excellent blogger / thinker, [Zvi Mowshowitz](#), has been squeezing his blogging / thinking into his life as a [pre-launch startup founder](#) and parent.

Or maybe those examples just illustrate that, within the “time / energy / slack” entry, “time” is a less important component than one might think. As they say, [“if you want something done, ask a busy person to do it”](#). (Well, within limits—obviously, as free time approaches literally zero, hobbies approach zero as well.)

Note a surprising corollary to this ingredient: **“direct work” (in the EA sense) and transformative hobbies can potentially work at cross-purposes!** For example, at my last job, I was sometimes working on lidar for self-driving cars, and sometimes working on military navigation algorithms,<sup>[1]</sup> and meanwhile I was working on AGI safety as a hobby (more on which below). Now, I *really* want there to be self-driving cars ASAP. I think they’re going to save lots of lives. They’ll certainly save me a lot of anguish as a parent! And we had a [really great technical approach to automobile lidar](#)—better than anything else out there, I still think. And (at certain times) I felt that the project would live or die depending on how hard I worked to come up with brilliant solutions to our various technical challenges. So during the periods when I was working on the lidar project, and I had extra time at night, or was thinking in the shower, I was thinking about lidar. And thus my AGI safety hobby progressed slower. By contrast—well, I have complicated opinions about military navigation algorithms, but let’s just say that they have never aroused in me a *great passion*. So during the periods when I was working on military navigation algorithms at my day job, and I had free time at night, or was thinking in the shower, I was thinking about AGI safety instead, and I made faster progress! (See Paul Graham’s essay [“The Top Idea in your Mind”](#).)

## Ingredient 2: ???

Here I’m referring to the fact that lots of people have extra time / energy / slack, and don’t use it for any world-changing hobbies. Instead, umm, I don’t know, maybe they watch a lot of TV, or argue about politics online, or host fancy parties, or build model ships, or whatever. (I’m not criticizing people who do those things. Your time is your time. Spend it as you wish!)

I don’t know what accounts for the difference. Certain types of motivation and skills and interests, I guess?

(I’m reminded of the [book quote](#): “*When you are older, you will learn that the first and foremost thing which any ordinary person does is nothing.*”)

## Ingredient 3: Willingness to pivot

As I understand it, Eliezer Yudkowsky was really into thinking about nanotechnology before deciding that actually thinking about AGI was a much better use of his free time, and then later [pivoted again](#) to thinking more specifically about AGI safety / alignment. Nate Soares likewise [describes](#) pivoting to AGI safety after a decade thinking about governance and institutions. I don’t know what David Denkenberger was thinking about in his free time before he came upon the question of feeding the world through nuclear winter, but I bet it was very interesting and different! For my part, **AGI safety was the 5th (!! long-term (i.e. multi-year) intense ambitious hobby of my life.** (And now it’s my job.)

I received the following comment on a draft version of this post:

There's also the danger that side projects also often have very unclear *failure* conditions. Because you're never trying that hard, lack of success feels fairly 'excusable', I can point to multiple examples of people who've devoted huge amounts of time+effort to side projects that were going nowhere, where I suspect that a full time 'put up or shut up' attempt would have caused them to (correctly) give up and try something else.

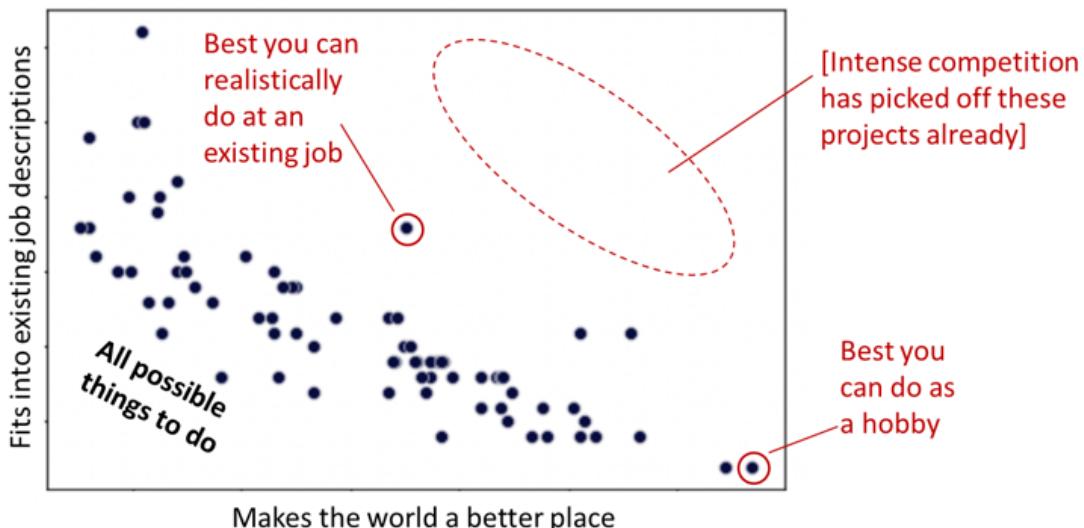
Here's my message to "people who are devoting huge amounts of time+effort to side projects that are going nowhere": Don't do that, OK? I mean, jeez. One of the key selling points of hobbies is that if you think maybe you should pivot and start from scratch on something totally different, you can just up and do it! There's nothing stopping you! You don't owe anyone anything! You have to kill your darlings! Maybe you learned something along the way, or at least had fun, and life is long, go start something new, etc. etc. That's my motivational speech, thanks for listening. Take it from me, a guy who spent 15 years building up a top-notch physics expertise that is now completely irrelevant for my life.

## How can hobbies compete at all with jobs? My theory: compromises are terrible

Naïvely, if we're going to compare "doing stuff as one's job" versus "doing stuff as one's hobby", the former seems to have overwhelming advantages: jobs bring to bear heavy artillery in the form of time, energy, money to spend on resources, multiple people coordinating, and so on. And don't get me wrong, jobs make the world go round. Nobody is manufacturing integrated circuits in their free time.

It's no coincidence that David Denkenberger and Eliezer Yudkowsky and Holden Karnofsky (and me!) transitioned their hobbies into proper jobs approximately as soon as possible (more on which just below).

But we still have to explain how important things get done *at first* as hobbies, not jobs. I think the answer is: compromises are terrible.



If you need to simultaneously satisfy two criteria, i.e. “this is an important thing to do” and “this is something I can immediately get paid to do”, then there’s a tradeoff. You need to compromise on both.

I said above that jobs bring to bear “heavy artillery” that hobbies can’t. But hobbies can make up for that deficiency with *better aim*.

(For more on the theme of compromises when trying to satisfy multiple criteria at once, see for example Eliezer Yudkowsky’s [Purchase Fuzzies and Utilons Separately](#), or Alex Lawsen’s [Know What You’re Optimizing For](#). For more on the advantages of doing things unconstrained by institutional incentives and inertia, see Paul Graham’s [The Power of the Marginal](#), and Eliezer Yudkowsky’s book [Inadequate Equilibria](#).)

## Hobbies can eventually develop into jobs

A hopeful possibility, of course, is that the hobby can eventually lay the groundwork for the creation of a *new* job, one whose money- and status-related incentives are all perfectly aligned with the thing that you have now figured out is what you want to do. Then you get the best of both worlds—the aim of the hobby, and the heavy artillery (time, resources, etc.) of a job.

But you can’t get there until you know what you’re aiming for.

## My own two stories

These are a couple anecdotes from my own life; feel free to skip down to the Conclusion if you’re uninterested.

### Story #1: Me & solar cells

Back in my innocent youth, I was passionate about inventing better solar cells. It was the reason I went into physics in the first place as an undergrad in 2003, and then I deliberately went to a physics grad school where there was an unusually high concentration of solar cell research happening.

How’d it go? I had some outputs in the “direct work” category (the thing I was “supposed to be doing” as a grad student), and also I had some outputs in the “slack / hobby” category (stuff I did informally, independent of my grad school advisor, unrelated to my eventual dissertation, etc.)

Shall we compare?

#### In the “direct work” category:

This category is where I was trying to compromise between

- “help develop better solar cell technology”, versus
- “do the normal grad student things”, e.g. “make progress towards my PhD” and “publish in high-profile journals” and “learn useful skills”.

Basically, I wound up doing decently at the latter and pretty much entirely failing at the former.

I worked in four labs:

- The first two labs were each doing interesting, trendy physics superficially related to solar cells, but their projects had no chance of actually making cheaper or better practical solar cells in the future. I wound up coauthoring a total of two papers in that category before switching labs. (The two papers [now have](#) 1400 and 500 citations respectively. It turns out that when you optimize for producing trendy high-profile papers, you can end up producing trendy high-profile papers. Who knew?)
- Then I switched to a lab that wasn't working on solar cells at all, but I liked the PI. I figured, oh well, too bad about the solar cells, but at least I'm getting a physics education! As it turned out, I had my own external funding and thus wasn't tied to any particular one of my PI's grants, and in my free time I thought of a maybe-practically-useful solar cell project. I found some excellent collaborators, and we got a [lovely paper](#) out of it, but nothing came of it in practice, in part because our patent application got rejected for stupid reasons.
- I also spent some time as a visiting student in a lab that was doing plausibly-useful solar cell R&D. However, the nature of the work was a terrible match to my skills and inclinations. Well, I guess I learned something valuable about my skills and inclinations, although in hindsight I should have already known that much earlier.

## In the “hobbies” category:

- I had written some optical simulation code related to solar cells (among many other things), and in my free time sometime during or after grad school I cleaned it up, documented it, and [put it on GitHub](#). I've heard from many people over the years that they use my code to help understand and design solar cells, including people in the R&D departments of at least two solar cell companies. Great success!
- The theoretical physics underlying how solar cells work was basically all well-established long before I entered the field. But I did distribute [some pedagogy](#) on the topic. It's hard to know what practical impact that had (if any), although I find it fun to open up a solar cell book, or watch a presentation, and hey, there's my diagram, which I had put on Wikipedia years earlier!

## Winner: Hobbies

I can say with some confidence that I have contributed a little bit to help bring cheaper and better solar cells into the world, *exclusively* via projects that I did in my free time, that my grad-school PI didn't even know about, and that probably would have done either nothing or almost-nothing to help me advance in academia.

Also, among my more standard legible grad-student-y projects, I had one near-miss which I think *could* have been practically useful (*a priori*). And wouldn't you know it, it was the one project that I initiated and developed in my free time (before eventually looping in my PI).

## Story #2: Me & AGI safety

In 2019, having finished some other hobby, I decided my next hobby would be AGI safety. So I started reading and writing posts and comments on LessWrong, in my little bits of time squeezed between my full-time job and two young kids.

Of course, I found it very frustrating how little time I had, and I immediately started brainstorming how I could get much more time by doing AGI safety as a job.

The normal advice would be to apply for existing AGI safety jobs, or go back to grad school etc. But I had high living expenses and didn't want to move out of Boston. Alternatively, I managed to come up with a handful of possible projects that I might do within my existing job, that seemed at least slightly relevant to AGI safety. For example, I could have tried to

talk my way onto my coworkers' existing [DARPA-funded project related to machine learning uncertainty](#), or various other things. But I wound up dropping those ideas too.

So nothing came of any of that, and I'm glad it didn't, because **my current belief is that I did dramatically better by just blogging and commenting on LessWrong in my tiny amount of free time, than I would have done if I had successfully found a way to do AGI safety in an institutional setting right off the bat**. I wound up going off in a weird intriguing direction, and then losing interest, and then going off in a *different* weird intriguing direction, which incidentally involved teaching myself neuroscience, and I wound up making enough progress to get a grant to do full-time independent research, and [here I am!](#) I think I'm doing something important and maybe a bit idiosyncratic, and it's all thanks to having spent a significant amount of time just reading and writing about whatever the hell I wanted.

## Conclusion

I was inspired to write this post right after talking to (what felt like) 3000 super-enthusiastic undergrads at EAGxBoston a couple months ago (I love you all!!), who all wanted to get into AGI safety / alignment.

Among the people I talked to, there seemed to be an unspoken assumption that, for example, if they go to grad school at all, they should try to maneuver into doing AGI-safety-related projects with an AGI-safety-concerned PI, and if they did so, victory!! Let's be clear: I'm not opposed to doing that! Some people should definitely be doing that! Great work has been done that way, e.g. [Alex Turner's dissertation](#). But let's not pretend that this isn't a *compromise*. You'll wind up in a project which is simultaneously optimizing for *both* advancing AGI safety *and* easy consumption by peer-reviewers, by your PI's future grant review committees, by your own dissertation committee, by future people reading your CV, etc.

(It's no coincidence that the "weird sci-fi stuff" side of AGI safety is almost entirely absent from academia.)

Apart from grad school, other frequently-proposed plans were: joining existing AGI safety nonprofits, joining existing mentorship programs, and getting good at ML (or neuroscience or whatever) in normal legible ways like "taking online courses" or "getting jobs in that field". Those are all good things too! I'm a huge fan of all those things!

But not many people brought up the idea of just having *any job whatsoever*, ideally a pleasant, invigorating job that you're really good at and which has good work-life balance, [2] and meanwhile *making a plan in your free time for how to solve AGI safety*. (Or if that's too ambitious, maybe "write intelligent comments on people's AGI-safety-related blog posts", and/or "create AGI-safety-related pedagogy", and work your way up from there—certainly that's all I was hoping for at first!) And *only then*, when and if you come up with a plan, you can apply for funding, either for independent research, or starting a new organization, or joining up with other people who share your vision, etc. [3]

Granted, that approach has its own issues. In some ways, it's a terrible plan! But I think it at least deserves a modicum of consideration.

(As usual, [consider reversing any advice!](#))

*Thanks Alex Lawsen, Justis Mills, and Adam Shimi for critical comments on a draft.*

1. ^

(among dozens of other projects, but these two can serve as prototypical examples)

2. ^

Or in the case of undergrads, I sometimes noticed perhaps a bit too much interest in filling up their résumé with the maximum number of research internships and good grades in hard classes etc., and perhaps a bit too little interest in having the mental space and energy to get sucked into thinking about interesting and important questions on their own time. [...] *Says the guy who spent his undergrad years filling up his résumé with the maximum number of research internships and good grades in hard classes etc. ...]*

3. ^

Or maybe you'll wind up feeling that AGI technical safety research isn't the right thing for you to do at all, and off you go in some other direction!

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

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

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

---

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

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

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

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

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

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

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

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

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

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

## Reactions to specific plans

### Owen Cotton-Barratt & [Truthful AI](#)

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

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

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

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

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

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

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

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

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

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

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

## Ryan Greenblatt & Eliciting Latent Knowledge

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

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

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

Suppose we have an ELK head, and an adversary, that we're using to train a nascent AGI. The AGI takes a sharp left turn. Now it's dangerously capable, and suddenly a

bunch of misalignments which it previously didn't notice within itself are relevant. And you've got to get them out.

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

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

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

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

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

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

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

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

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

**Eric Drexler & AI Services**

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

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

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

## Evan Hubinger, in a recent personal conversation

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

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

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

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

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

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

## Rohin Shah's

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

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

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

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

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

## Another recent proposal

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

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

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

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

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

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

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

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

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

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

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

## John Wentworth & Natural Abstractions

**Imaginary John:** I suspect there's a common format to concepts, that is a fairly objective fact about the math of the territory, and that—if mastered—could be used to

understand an AGI's concepts. And perhaps select the ones we wish it would optimize for. Which isn't the whole problem, but sure is a big chunk of the problem. (And other chunks might well be easier to address given mastery of the fairly-objective concepts of "agent" and "optimizer" and so on.)

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

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

**Nate:** :'(

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

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

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

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

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

## **Neel Nanda & [Theories of Impact for Interpretability](#)**

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

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

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

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

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

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

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

## **Stuart Armstrong & Concept Extrapolation**

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

**Stuart, in a blog post:**

[...] It is easy to point at current examples of agents with low (or high) impact, at safe (or dangerous) suggestions, at low (or high) powered behaviours. So we have in a sense the 'training sets' for defining low-impact/Oracles/low-powered AIs.

It's extending these examples to the general situation that fails: definitions which cleanly divide the training set (whether produced by algorithms or humans) fail to extend to the general situation. Call this the 'value extrapolation problem, with 'value' interpreted broadly as a categorisation of situations into desirable and undesirable.

[...] Value extrapolation is thus necessary for AI alignment.

[...] We think that once humanity builds its first AGI, superintelligence is likely, near, leaving little time to develop AI safety at that point. Indeed, it may be necessary that the first AGI start off aligned: we may not have the time or resources to convince its developers to retrofit alignment to it. So we need a way

to have alignment deployed throughout the algorithmic world before anyone develops AGI.

To do this, we'll start by offering alignment as a service for more limited AIs. Value extrapolation scales down as well as up: companies value algorithms that won't immediately misbehave in new situations, algorithms that will become conservative and ask for guidance when facing ambiguity.

We will get this service into widespread use (a process that may take some time), and gradually upgrade it to a full alignment process. [...]

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

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

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

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

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

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

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

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

[Model splintering] is one way of more formally looking at the out-of-distribution problem in machine learning: instead of simply saying that we are out of distribution, we look at the model that the AI previously had, and see what model it transitions to in the new distribution, and analyze this transition.

Model splintering in particular refers to the phenomenon where a coarse-grained model is "splintered" into a more fine-grained model, with a one-to-many mapping between the environments that the coarse-grained model can distinguish between and the environments that the fine-grained model can distinguish between (this is what it means to be more fine-grained).

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

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

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

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

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

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

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

**Nate:** That sounds like a laudable goal. I have not yet managed to understand what sort of foundations of alignment you're trying to scale down and build into modern systems. What are you hoping to build into modern systems, and how do you expect it

to relate to the problem of aligning systems with capabilities that generalize far outside of training?

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

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

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

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

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

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

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

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

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

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

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

## Andrew Critch & political solutions

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

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

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

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

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

**Nate:** Here I basically have the same response: First, can't be done (though I endorse attempts to prove me wrong, and recommend practicing by trying to effect important

political change on smaller-stakes challenges ASAP (The time is ripe for sweeping global coordination in pandemic preparedness! We just had our warning shot! If we'll be able to do something about AGI later, presumably we can do something analogous about pandemics now!).

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

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

## What about superbabies?

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

## What about other MIRI people?

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

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

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

## High-level view

I again stress that all the people whose plans I am pessimistic about are people that I consider virtuous, and whose efforts I applaud. (And that my characterizations of people above are probably not endorsed by those people, and that I'm putting less

effort into passing their [ideological Turing Tests](#) than would be virtuous of me, etc. etc.)

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

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

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

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

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

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

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

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

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

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

So the status quo looks grim to me.

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

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

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

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

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

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

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

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

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

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

## 2. ^

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

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

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

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

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

## 3. ^

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

## 4. ^

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

## 5. ^

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

## 6. ^

I specifically see:

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

## 7. ^

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

# Toni Kurz and the Insanity of Climbing Mountains

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

---

*Content warning: death*

I've been on a YouTube binge lately. My current favorite genre is disaster stories about mountain climbing. The death statistics for some of these mountains, especially ones in the Himalayas are truly insane.

To give an example, let me tell you about a mountain most people have never heard of: Nanga Parbat. It's a 8,126 meter "wall of ice and rock", sporting the tallest mountain face and the fastest change in elevation in the entire world: the Rupal Face.



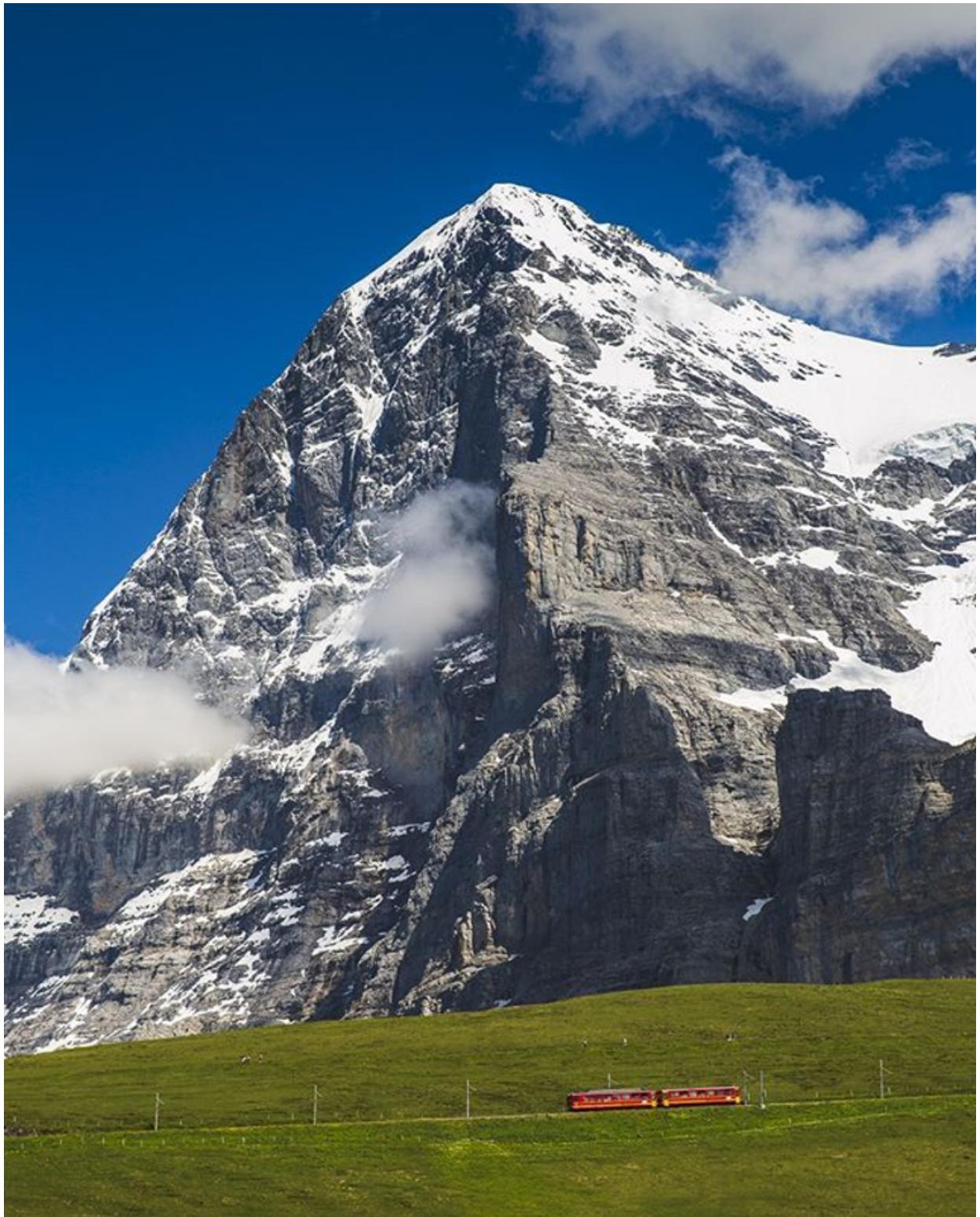
I've posted a picture above, but these really don't do justice to just how gigantic this wall is. This single face is as tall as the largest mountain in the Alps. It is the size of ten empire state buildings stacked on top of one another. If you could somehow walk straight up starting from the bottom, it would take you an entire HOUR to reach the summit.

31 people died trying to climb this mountain before its first successful ascent. Imagine being climber number 32 and thinking "Well I know no one has ascended this mountain and thirty one people have died trying, but why not, let's give it a go!"

The stories of deaths on these mountains (and even much shorter peaks in the Alps or in North America) sound like they are out of a novel. Stories of one mountain in particular have stuck with me: the first attempts to climb tallest mountain face in the alps: The Eigerwand.

## **The Eigerwand: First Attempt**

The Eigerwand is the North face of a 14,000 foot peak named "The Eiger". After three generations of Europeans had conquered every peak in the Alps, few great challenges remained in the area. The Eigerwand was one of these: widely considered to be the greatest unclimbed route in the Alps.



The peak had already been reached in the 1850s, during the golden age of Alpine exploration. But the north face of the mountain remained unclimbed.

Many things can make a climb challenging: steep slopes, avalanches, long ascents, no easy resting spots and more. The Eigerwand had all of those, but one hazard in

particular stood out: loose rock and snow.

In the summer months (usually considered the best time for climbing), the mountain crumbles. Fist-sized boulders routinely tumble down the mountain. Huge avalanches sweep down its 70-degree slopes at incredible speed. And the huge, concave face is perpetually in shadow. It is extremely cold and windy, and the concave face seems to cause local weather patterns that can be completely different from the pass below. The face is deadly.

Before 1935, no team had made a serious attempt at the face. But that year, two young German climbers from Bavaria, both extremely experienced but relatively unknown outside the climbing community, decided they would make the first serious attempt.

One of the things which makes the climb of the Eiger unique is that nearly the entire face is visible from a mountain resort below. Residents of Kleine Scheidegg, a small resort town in the pass, could look directly at the north face when the weather cleared and observe all of these attempts to climb the face.



All of this was in place long before the first attempt was made, so when the two young Bavarians decide to make an attempt, the world's press was literally staying at the hotel watching the men through binoculars when the clouds cleared.

Not knowing how long the attempt would take, they brought six days of supplies, estimating it would take two to three days to achieve the summit. They started off

quite strong, making it all the way up to Eigerwand station before setting up camp for the night. Yes, you read that correctly. There are train tracks a third of the way up the mountain. Here's the view from a window looking down on (again, I'm not making this up), Grindelwald.



On the second day, they made little progress, having to contend with the first major ice field of the climb. On the third day, they made it to the second of these and were seen near the top before clouds set in and the view of the face was obscured.

When the clouds cleared on the fifth day, it became clear a major disaster was in store. The entire mountain face was covered with several feet of fresh snow, an unusual occurrence for the summer months. Avalanches crashed down the mountain, making it impossible for the climbers to descend via their previous route. The men had no choice but to continue upwards, hoping that they might make it to the top before their supplies ran out or they died of exposure. They were last seen alive on day five, high up on the third ice field with several thousand feet left to the summit.

Days later, when an airplane flew by the peak to try to locate the climbers, one was spotted frozen solid, standing up in the third ice field. This location later became known as "death bivouac".



## The Eigerwand: Second Attempt

You'd think that after such a tragedy, climbers might be at least temporarily deterred. But that would be an underestimate of how insane climbers are. From what I have read, several of the climbers that joined the search party to look for the two Bavarians mainly used it as an excuse to scout the mountain for their own attempt!

Ten men planned to make summit attempts in the 1936 season. But bad weather and climbing accidents reduced that number to just four by July.

Two groups decided to make an attempt in 1936: two men from Bavaria: Andreas Hinterstoisser and Toni Kurz, and two Austrians: Willy Angerer and Edi Rainer. During the preliminary expedition, the two groups decided to climb together.





Top: Edi Rainer and Willy Angerer Bottom: Toni Kurz and Andreas Hinterstoisser

On the very first day, Hinterstoisser fell 37 meters down the mountain face (but was apparently uninjured). Other than that the men made good progress. Hinterstoisser used a new technique with fixed ropes to cross a steep rock face now known as the "Hinterstoisser Traverse".



But crucially, he removed the ropes after making the traverse, and the same move would be much more difficult to pull off if the climbers needed to go back. Clouds settled in over the first night, and the view of the mountain was obscured to the spectators watching through binoculars and telescopes from Kleine Scheidegg below.

Early on the second day, it appeared to spectators as though something had gone wrong. Edi and Willy had stopped ascending. It looked as though Edi was attending to

Willy, Andreas and Toni let down a rope to Willy, who seemed recovered enough to continue his ascent, followed by Edi.

Their progress slowed. By the middle of the day, they had reached Death Bivouac, the final resting place of the German climbers from the year before. But by this point, it was clear to spectators that Willy could not go on. Whatever injury he had suffered during the first day was bad enough to prevent him from continuing.

The four began descending and made good progress down the second ice field by the end of the second day. But the rock face between the first and second ice fields would be much more challenging on the way back down, as Andreas had removed the fixed ropes used on the ascent, and there was no clear route back down the face.

On the third day, a storm rolled in, and clouds and mist obscured the face of the Eigerwand to spectators. Avalanches could be seen tumbling down the mountain, bringing a shower of rocks with them.

When the clouds briefly cleared, onlookers could see the rock face by which they had ascended to the second ice field had been drenched in freezing rain from the night before.



It became apparent that with the fixed ropes gone, the climbers could not return by that route. The only way down the mountain was to rappel off a 600-foot cliff face.

Sensing another disaster might be imminent, a mountaineer named Albert Von Almen took the train up to the windows in the Eigerwand tunnel, close to the base of the cliff that the four men were now preparing to descend.

When Von Almen poked his head out of the window of the Eigerwand tunnel, he shouted out for the climbers high above him. To his surprise, he heard four replies. All men appeared to be well, and said they would be down soon. Von Almen set to work preparing a pot of tea, which he hoped would warm the climbers after the brutal ordeal.

Minutes ticked by. After two hours of waiting, Von Almen became increasingly worried. He returned to the window and again shouted for the men. This time, only one voice could be heard: Toni Kurz.

The wind made it hard to hear, but Von Almen gathered two pieces of information: Everyone but Toni was dead, and Toni was stuck dangling in the air hundreds of feet above the train tunnel window, unable to descend any further.

Von Almen immediately phoned Eigergletscher Station in the valley below, telling them to send a rescue party immediately. At the time, the head of the mountain rescue committee ruled that no guides were to be compelled to take part in the rescue mission, given the extreme risk involved. But three guides volunteered and rode the train up to the window in the mountainside.

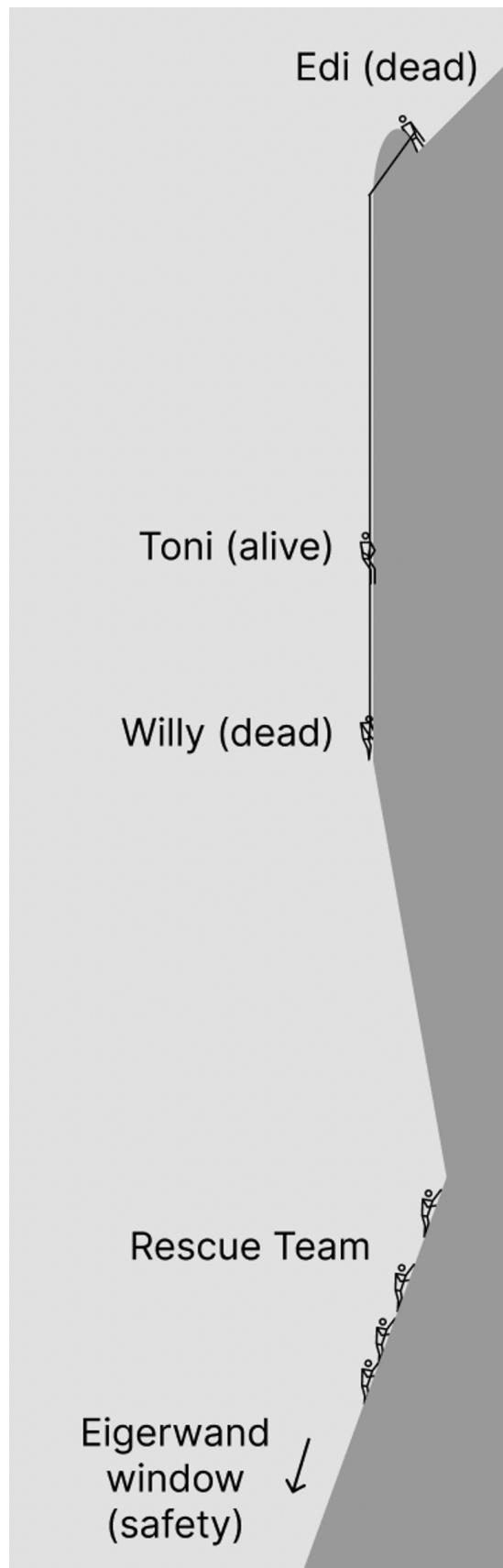




Exiting the windows of the Eigerwand tunnel, the guides traversed diagonally upwards on the face towards the base of the cliff from which Toni was hanging. He spoke to them, his voice still strong despite spending three days on the mountain.

The four men had been hit by an avalanche. Hinterstoißer, the strongest climber who had set the rope on the rock face during the ascent, had been swept completely off the mountain face, falling nearly a thousand feet to his death. Willy, Edi and Toni had all been tied together by a single rope, with Toni in the middle. Both Willy and Toni had been swept off. In the fall, the rope had tangled itself around Willy's neck and strangled him. Edi, still at the top of the cliff face and tied to both the fallen men, had been smashed against a rock at the top, fracturing his skull before freezing to death soon after. But his frozen body remained pressed against the rock, saving Toni from certain death.

Toni was alone, and with 300 feet of empty air below and above him, there was no way for the guides to reach him without climbing 600 feet up the crumbling ice between the first and second ice fields. The distance was too far to throw a rope up.



With night having fallen, the guides realized they would likely die attempting to climb the ice face and rescue Toni from above. They promised Toni they would return the next morning for another rescue attempt. Toni shouted to them that he would not make it through the night. Long after they left, they could hear him pleading for help as they descended back to the window in the train tunnel.

When the guides returned the next morning, Toni has 8 inch icicles hanging from his boots. During the night, the wind has ripped off his left mitten. His hand was now frozen completely solid, along with his lower left arm.

It was clear to the guides that there was simply no way to ascend the cliff. With modern equipment, it would have perhaps been possible, but with the mountain climbing equipment of the 1930s there was simply no way to climb a frozen rocky ice face.

The first rescue idea was to throw Toni a rope. They even brought rockets to launch a rope up to him. But this plan failed with all the ropes flying out away from the cliff face into empty air. The second idea was for Toni to lower a small rope, to which they would tie one of the rescue ropes. Toni could then tie that rope to his and descend the rest of the way down. But Toni had no remaining rope to lower. Somehow he needed to make more rope.

The guides could think of one plan that might work, but it relied on Toni having remaining physical strength. They told him to climb as far down as he could, then cut away the dead body of Willy. He would then need to climb back up, tie himself again, and cut the rope just underneath him.

Then, with one frozen arm, he would need to unwind the short section of rope and fasten together the pieces to lower down to the rescue team. This thin rope (not strong enough to hold Toni's weight) would then be used to raise up a stronger rope supplied by the rescue team, which he would need to tie to his own and use to descend.

Over the course of five painstaking hours, Toni worked to make a new rope to lower to the guides. He cut Willy's body from the rope, but it did not fall, as the freezing rain from the night before had frozen it solidly to the cliff. He then climbed up about 25 feet with one working arm and frozen feet, and used his ax to cut a section of rope below him. Then using his teeth and his one good hand, he began to unwind the short section of rope and tied each section together to lower to the guides.

The sun passed its peak in the sky and began to sink slowly into the west. At one point an avalanche thundered down the mountain bringing rocks and snow careening past the guides. The debris unseated Willy's frozen body from the cliff face, and it hurled past the rescue team, tumbling down the mountain into the valley below.

Finally, Toni finished his makeshift guide rope and lowered it to the rescuers. His strength was nearly exhausted. The guides attached a thicker stronger rope to it, along with some climbing supplies in case Toni needed to climb down the cliff face. But even the rescue rope was not long enough, so they tied a second rope to it near the bottom.

Somehow, after four nights of no sleep, exposure to the wind and rain, and with one good arm, Willy managed to slowly haul the rope and gear up over the course of an hour. He then began the slow, torturous descent.

Toni came into view, fifty feet above the guides. Now thirty feet, then twenty. Then suddenly, he came to a halt. The knot the rescuers had tied to attach the second rope to the first was too large to fit through Toni's carabiner. He could not descend further.



The guides could hear him groaning as he fought to get the knot through.

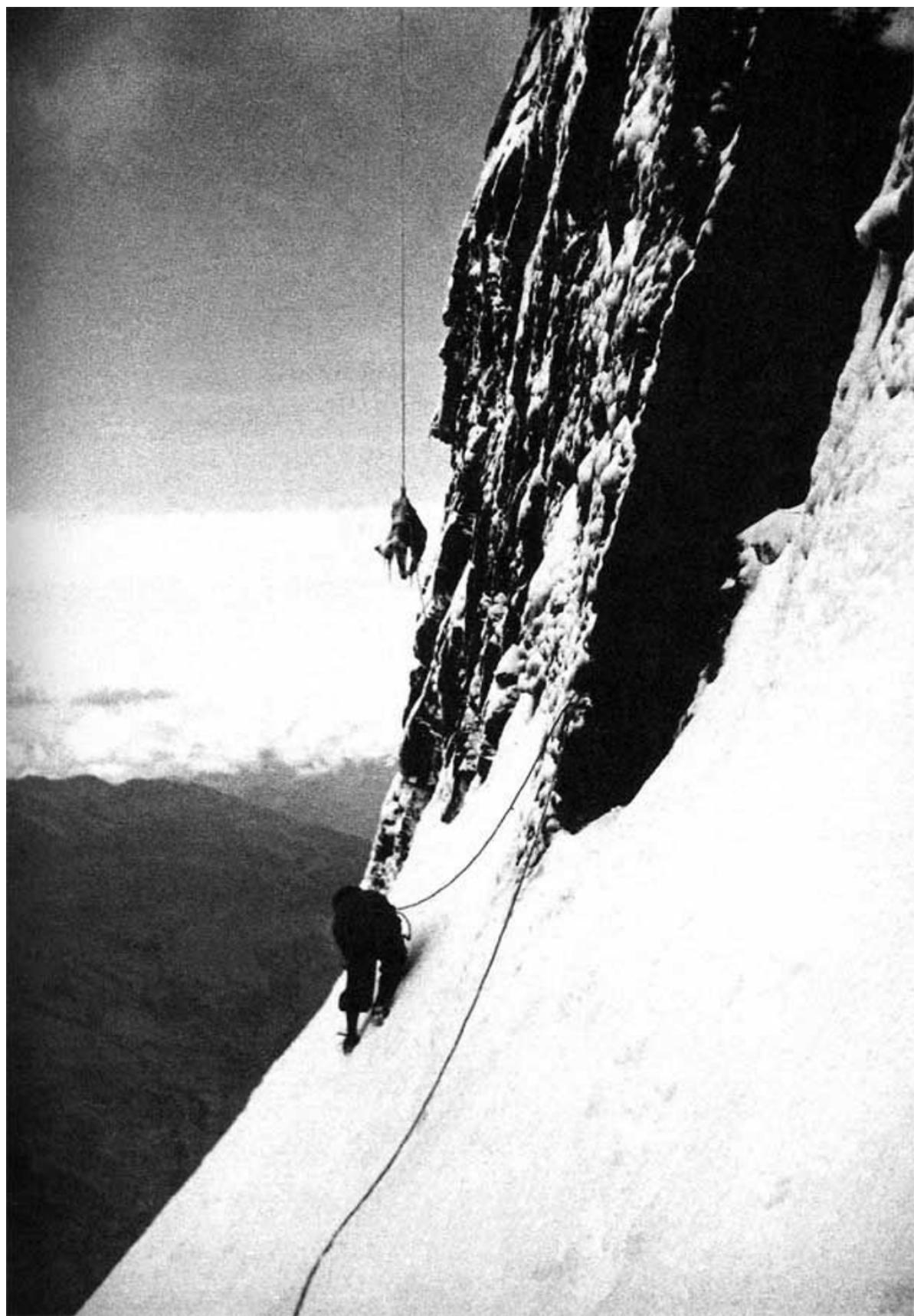
"Try, lad, try!" the frustrated rescuers cried to encourage the exhausted man. Toni, mumbling to himself, made one more effort with all his remaining strength, but he had little left; his incredible efforts had used it almost all up. His will to live had been keyed to the extreme so long as he was active; now, the downward journey in the safety of the rope-sling had eased the tension. He was nearing his rescuers now; now the battle was nearly over, now there were others close at hand to help....

And now this knot ... just a single knot ... but it won't go through.... "Just one more try, pal. It'll go!"

There was a note of desperation in the guides' appeal. One last revolt against fate; one last call on the last reserves of strength against this last and only obstacle. Toni bent forwards, trying to use his teeth just once more. His frozen left arm with its useless hand stuck out stiff and helpless from his body. His last reserves were gone. Toni mumbled unintelligibly, his handsome young face dyed purple with frost-bite and exhaustion, his lips just moving. "Was he still trying to say something, or had his spirit already passed over to the beyond?

Then he spoke again, quite clearly. "I'm finished," he said.

His body tipped forward. The sling, almost within reaching distance of the rescuing guides, hung swinging gently far out over the gulf. The man sitting in it was dead."1



# What to make of the men who climb

I remember when I first finished hearing this story being simply overwhelmed by it. It was so terrible, with so many chances for things to have gone differently. If only Willy hadn't been hit by falling rocks. If only they had turned back an hour earlier they would have completed the belay an hour earlier and avoided the avalanche. If only Hinterstoisser hadn't removed the fixed ropes. If only one of the guides hadn't dropped the longer rope, there would have been no knot for Toni to get stuck on. If only Toni hadn't dropped his glove in the storm he might have had use of both hands. If only these brave idiots had decided not to make an attempt that season like the other five climbers who left weeks before.

So what do we make of these men, who risk so much for so little? Are these urges to go forth and conquer, to take great risk for little practical benefit, simply evolutionary vestiges of a world now gone? Are they misdirected expressions of an inner urge to distinguish oneself? Are they simply far out on the tail end of the distribution of a trait that in moderation, is actually quite helpful? Do these actions somehow make sense in a way I don't yet understand?

The idea that these climbers don't understand the risk they are taking on doesn't seem to hold water. They understand the danger of climbing these mountains (especially those that haven't been climbed before). Reinhold Messner, one of the greatest climbers of all time, and the first to summit Everest without oxygen, explained that when he climbs, his mind tells him to go back, to not venture forth into the dark, cold and desolate winds. Yet he does.

And I watch, with a mixture of horror, dread, and fascination. Like one of the gawking tourists at Kleine Scheidegg, I watch through my glowing rectangle the pain and the tragedy these men endure to stand on top of a mountain that has already been climbed.

# Unifying Bargaining Notions (1/2)

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

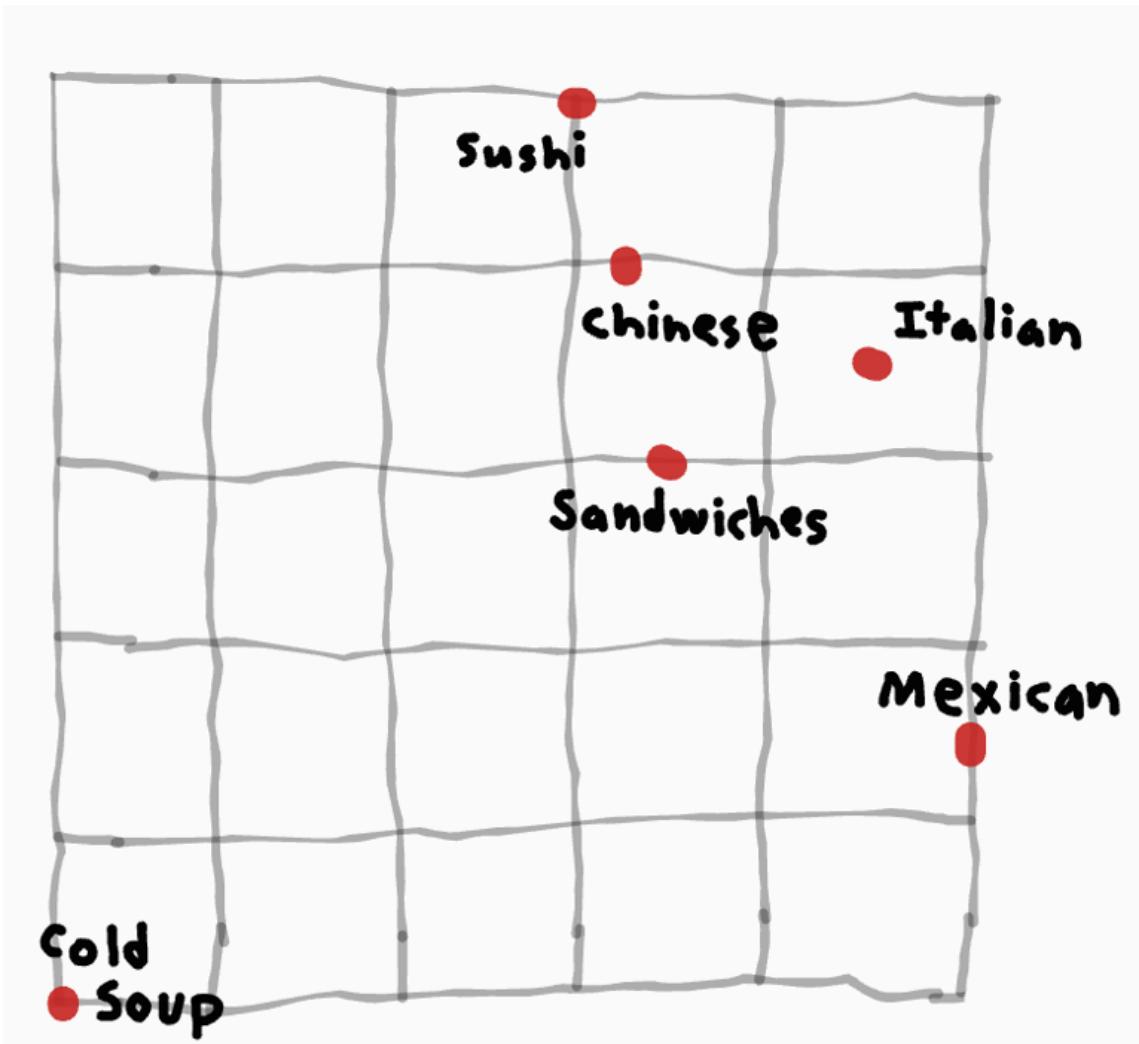
This is a two-part sequence of posts, in the ancient LessWrong tradition of decision-theory-posting. This first part will introduce various concepts of bargaining solutions and dividing gains from trade, which the reader may or may not already be familiar with.

[The upcoming part](#) will be about how all introduced concepts from this post are secretly just different facets of the same underlying notion, as originally discovered by [John Harsanyi](#) back in 1963 and rediscovered by me from a completely different direction. The fact that the various different solution concepts in cooperative game theory are all merely special cases of a General Bargaining Solution for arbitrary games, is, as far as I can tell, not common knowledge on Less Wrong.

## Bargaining Games

Let's say there's a couple with a set of available restaurant options. Neither of them wants to go without the other, and if they fail to come to an agreement, the fallback is eating a cold canned soup dinner at home, the worst of all the options. However, they have different restaurant preferences. What's the fair way to split the gains from trade?

Well, it depends on their restaurant preferences, and preferences are typically encoded with utility functions. Since both sides agree that the disagreement outcome is the worst, they might as well index that as 0 utility, and their favorite respective restaurants as 1 utility, and denominate all the other options in terms of what probability mix between a cold canned dinner and their favorite restaurant would make them indifferent. If there's something that scores 0.9 utility for both, it's probably a pretty good pick!



Although, there's something off about setting up the problem like this. There's no term for intensity of preferences! Someone who cared very little about food would have their preferences rank just as strongly as someone who had strong restaurant opinions!

In a sense, there's three responses to this objection.

The first response is that we might be zooming in too hard on the restaurant bargaining game in particular. In a broader context, a person having weak restaurant preferences may just be another way of saying that they are quick to trade off their choice of restaurant to someone else in return for other things they might desire. And so, in the broader bargaining game of a relationship where more is at stake than this one-time choice of restaurant, things may be fair. But in the restaurant bargaining game in particular, things can look unfair for the losing party, when in fact they traded off "ability to determine restaurant" in exchange for more concessions elsewhere. The generalization of this is that bargaining equilibria of an overall game might be quite different from just summing up the bargaining equilibria of the subgames.

The second response is that people care a nonzero amount about other people, and so someone with weak food preferences might be equally well modeled as someone with a strong preference that their partner get what they want. That can be folded into the utility function, however. Just make the ratings of the deferential person mostly copy the ratings of their partner.

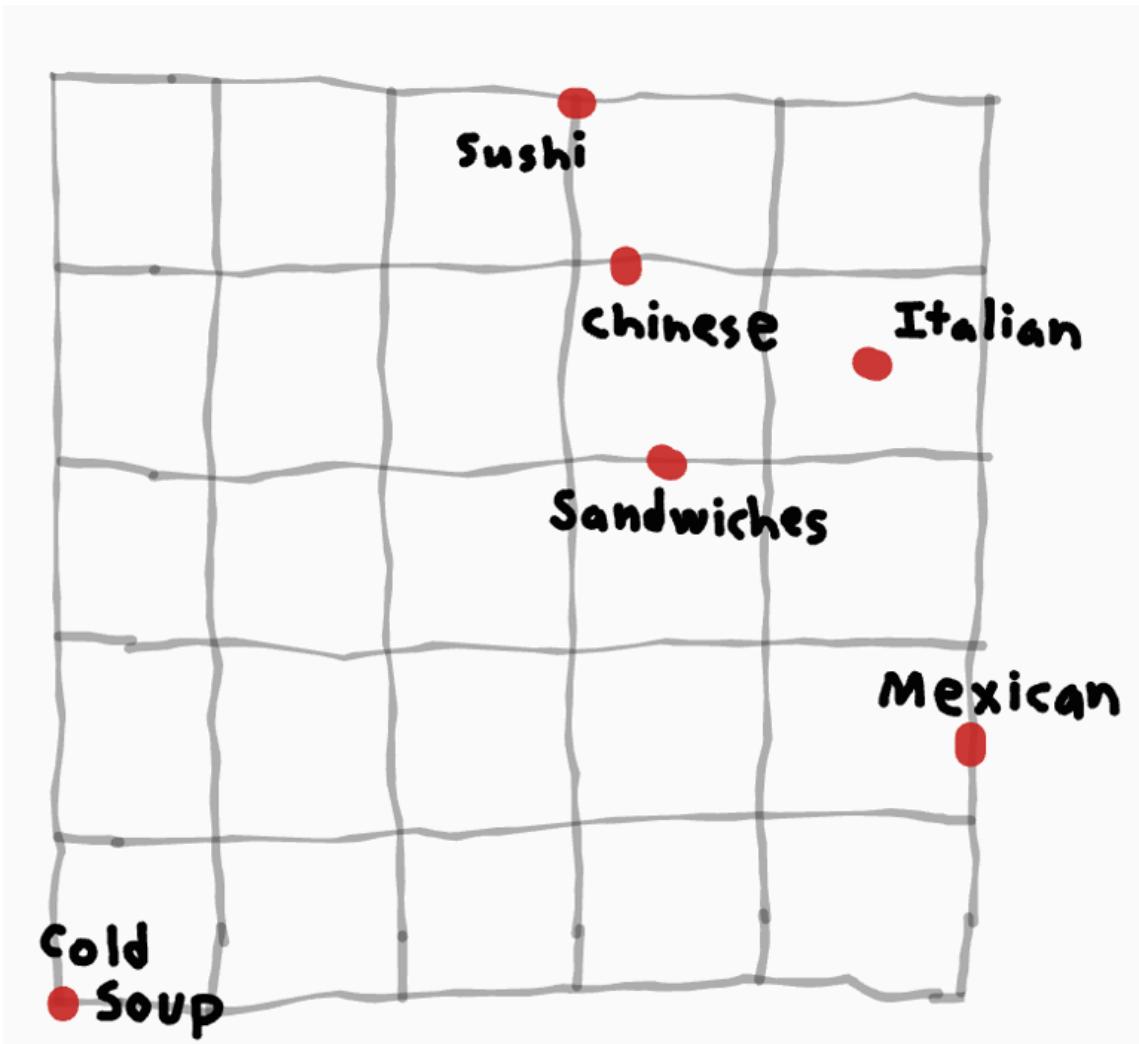
And the third response is one of the most interesting. For a perfectly selfish person who always tries for their favorite foods and doesn't care at all about your pouting at disfavored restaurants, there really isn't much of a difference between having strong preferences for food and weak preferences for food, they'll still drive as hard of a bargain against you as they can, if there isn't some mitigating factor.

Much like the post about how the [TRUE prisoner's dilemma](#) is not the standardly framed version, but more like "a human civilization fighting with a paperclip maximizer for resources which can either save millions of lives, or make a few paperclips", the TRUE bargaining problem isn't couples deciding where to eat, but something more like "deciding how to split a pile of resources with nonsentient aliens that are willing to fight you over the resource pile".

Accordingly, using the term "fair" for any of these mathematical concepts has the problem of automatically importing human concepts of fairness, which needs to be resisted in order to look clearly at what the math is doing. It'll be handy to have a separate word for "a mathematically distinguished point in a game where both parties have an interest in preventing destructive conflicts, that's neutral enough that aliens would probably come up with it" to enforce that mental separation. Let's use "chaa" as a nonsense word to denote that concept (the Lawful Neutral Alien analogue of fairness), since it makes it a lot easier to point at situations where the chaa outcome splits apart from the fair outcome.

The relevant questions to ask to work out what the chaa outcome is are things like "what are our best alternatives to a negotiated agreement and how does it compare to the choices on offer for us" instead of "how strong are our preferences compared to each other", (which is more relevant to fairness)

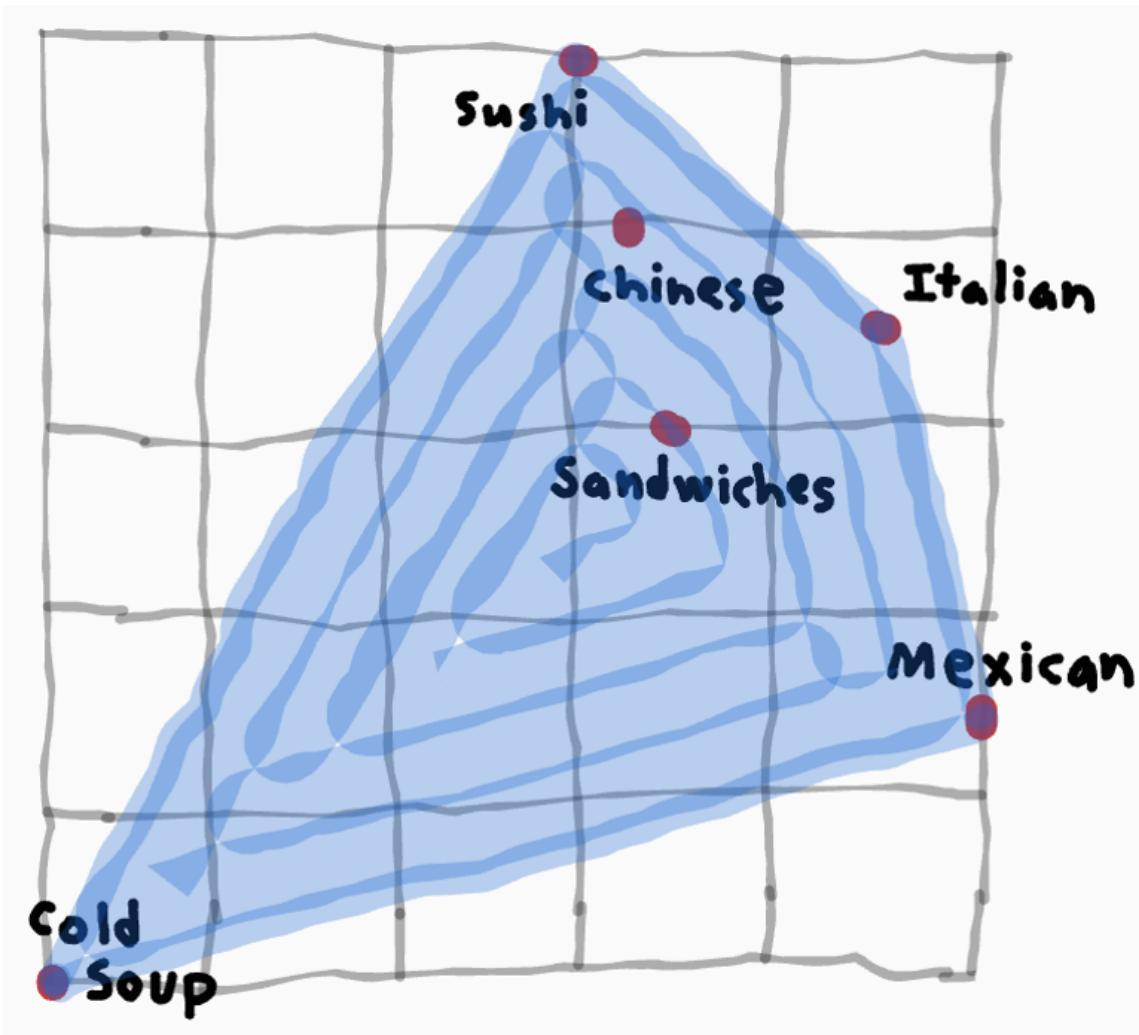
Anyways, returning to our restaurant game, to actually answer the question of what to do, let's see how we set up the problem.



We plotted the utilities of the various options, and got a scattering of points on the plane, where one of the coordinates is the utility assigned to the outcome by Alice, and the other is the same for Bob.

One extremely important note is that it should be possible to randomize between various options. For instance, if there's only two options, one where player 1 wins completely, and one where player 2 wins completely, an obviously chaotic outcome is the players flipping a coin to decide who wins.

In graphical terms, access to randomization lets us set up situations that can attain any utility pair in the convex hull of these points.



So, which points in this shape are chaa outcomes?

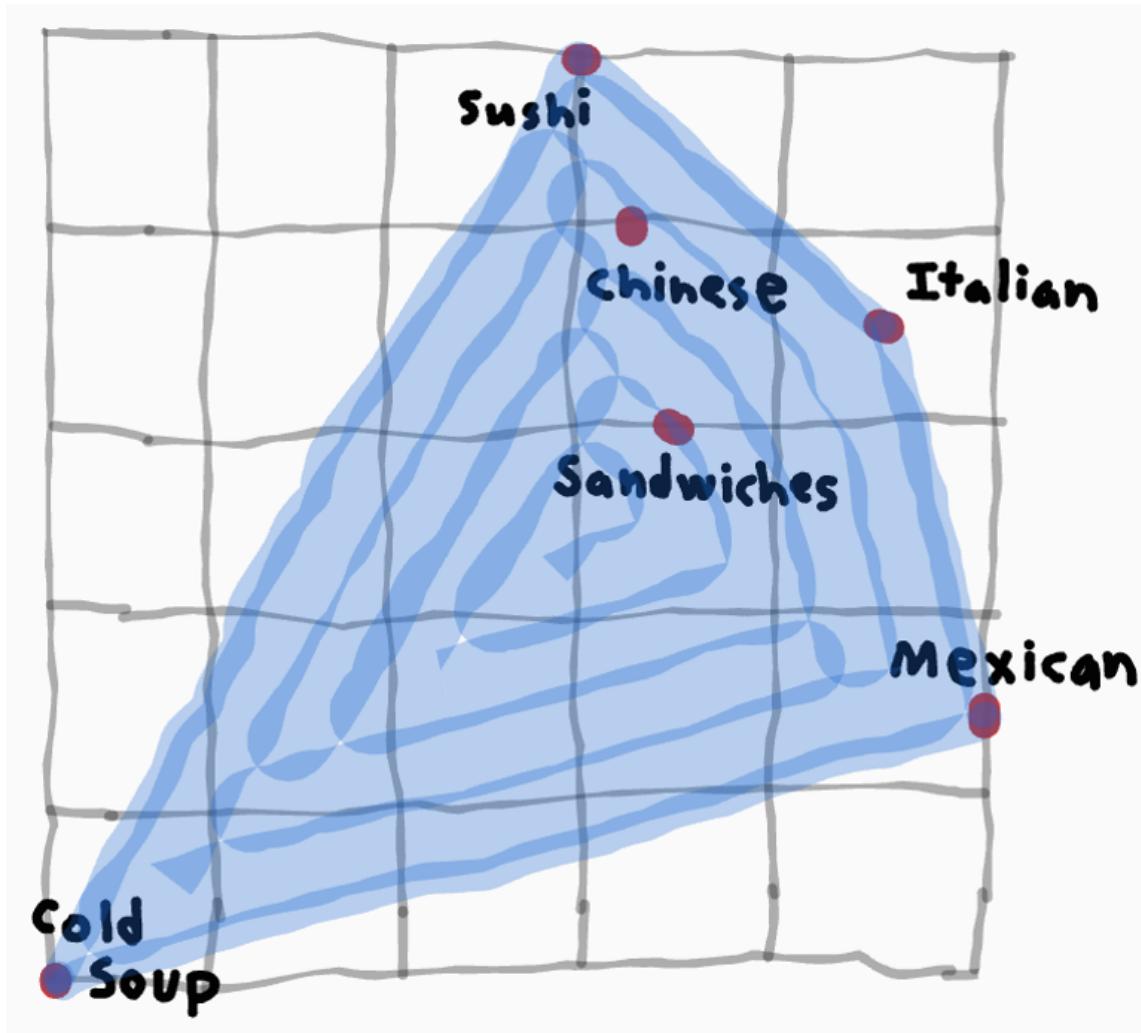
Well, "chaa" hasn't been defined yet, but if it's about how to split gains between agents in a neutral way without getting into destructive conflicts, there's three obvious properties that such a solution must have.

First, since chaaness is partially about not getting into destructive conflicts, any chaa point should be on the Pareto frontier. Namely, there shouldn't be an alternative that leaves both players strictly better off. After all, if you have a prospective definition of "chaa" that demands that both players leave spare utility on the table, they should be able to take that as their new disagreement point, do another round of bargaining, and attain an outcome which is Not That and better for both. And then, this process would give you a new notion of chaaness that's just strictly better for all agents to use. So, whatever point is selected, it must be from the upper-right boundary of the shape.

Second, the definition of a chaa point shouldn't be sensitive to the exact utilities used. You can add any constant to a utility function, or multiply it by any positive constant, and it'll be the same utility function. Reporting your preferences as the function  $U$  should get you the same result as if you reported your preferences as the function  $100U$ , or if you reported your preferences as the function  $5U + 7$ . No matter how the players relabel their utility functions

and scale and shift them, it shouldn't affect anything because their underlying preferences didn't change.

This is convenient because it means that we can always rescale the disagreement point to 0 utility, and the highest utility a player can get (without making it so the other player would rather go for the disagreement point) to 1 utility. So, you only really have to consider problems where your convex shape fits in a unit square like this.

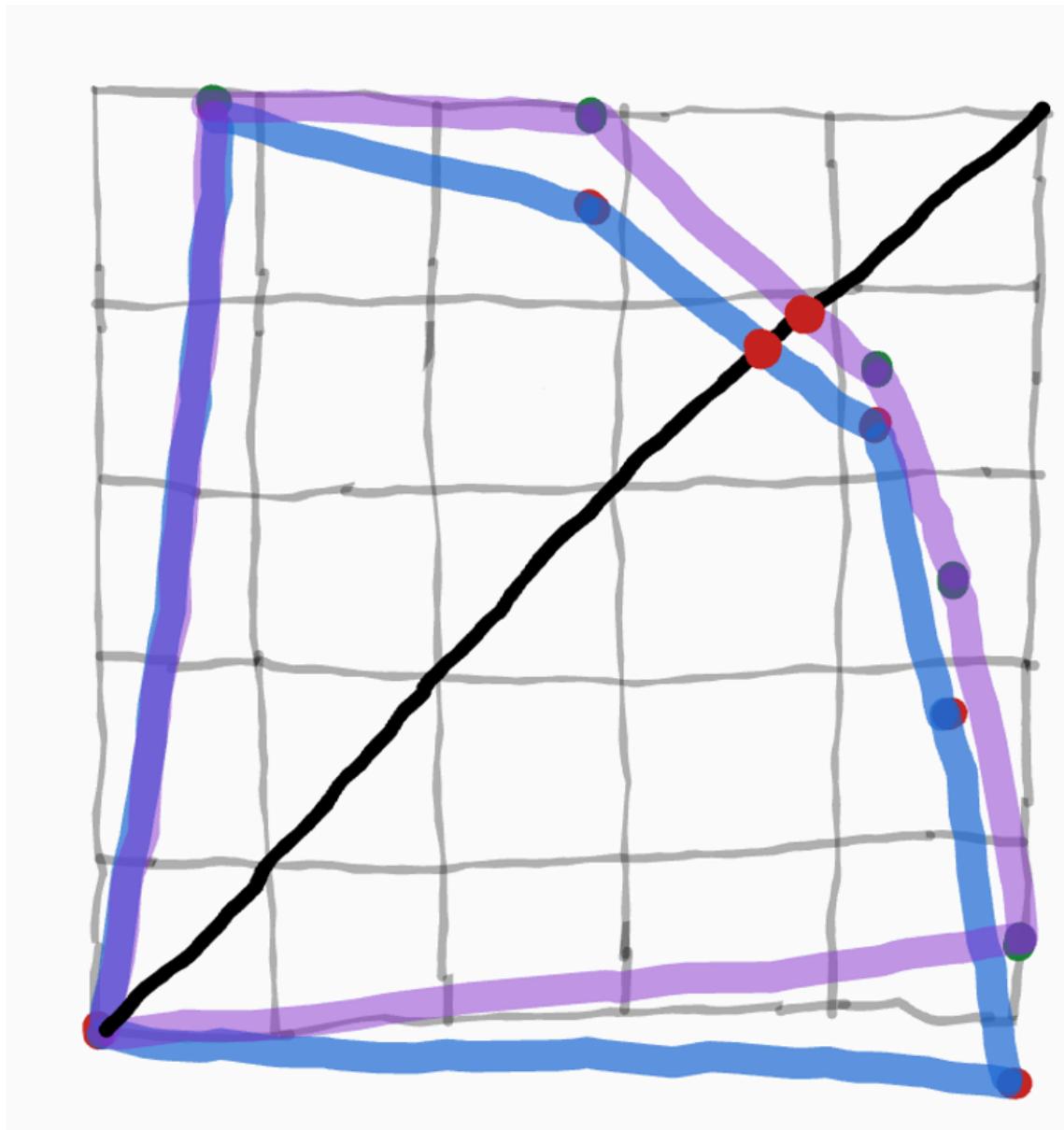


This leads nicely into our third desiderata. If the convex shape is symmetric, then the two players are in an identical position. Thus, any neutral way of selecting gains for the two players must be indifferent between which player is first and which is second, and so the chaa point should end up being on the line of symmetry, or on the halfway point. If one of the players is selected to win completely, the chaa outcome should involve flipping a coin to decide who wins. For the prisoner's dilemma, the chaa outcome should be mutual cooperation. For the game of chicken, the chaa outcome should be flipping a coin to decide who goes straight and who swerves.

These three desiderata are as obvious as can be, but past this they get a whole lot more controversial.

The Kalai-Smorodinsky Bargaining Solution is "Rescale things so the disagreement outcome is at 0, 0, and 1 utility for a player is the maximum utility they can get without sending the foe below 0 utility. Draw a diagonal line from 0, 0 to 1, 1, pick where the line crosses the Pareto frontier."

Pretty simple, right? It's the only way of picking a point that fits all three of our desiderata, and also fulfills the extra property of monotonicity, which is basically saying that, if you move the Pareto-frontier points for a player up, they should get more utility.



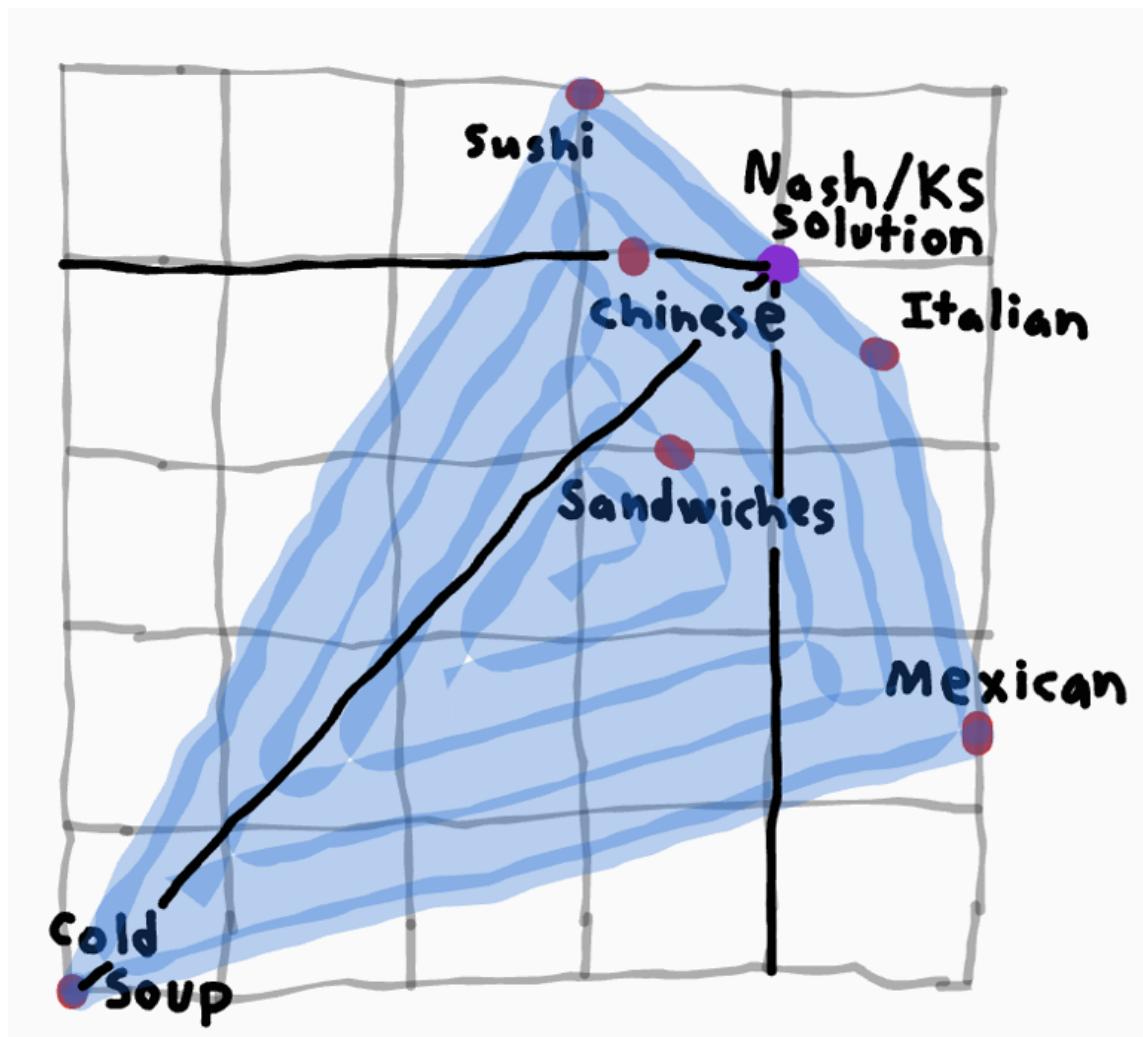
Yes, yes, I didn't quite do it correctly, that point of the blue shape in the bottom-right corner isn't scaled appropriately, but eh, it's close enough. It makes it pretty clear what we're doing with the line and how it is that moving the various points up (to go from the blue shape to the purple shape) increases the expected utility of the player whose utility is being plotted on the y coordinate. After all, if you've got better options, a chaa outcome shouldn't leave you with lower expected utility!

Well... that's a bit of a fiddly issue. Remember, utility functions are scale-and-shift invariant. So, when we move these Pareto-frontier points up, we're not REALLY getting extra utility, this operation is really more like making the utility function more squashed at the top.

Hopefully, monotonicity doesn't look completely obvious now, though it still has an awful lot of intuitive force.

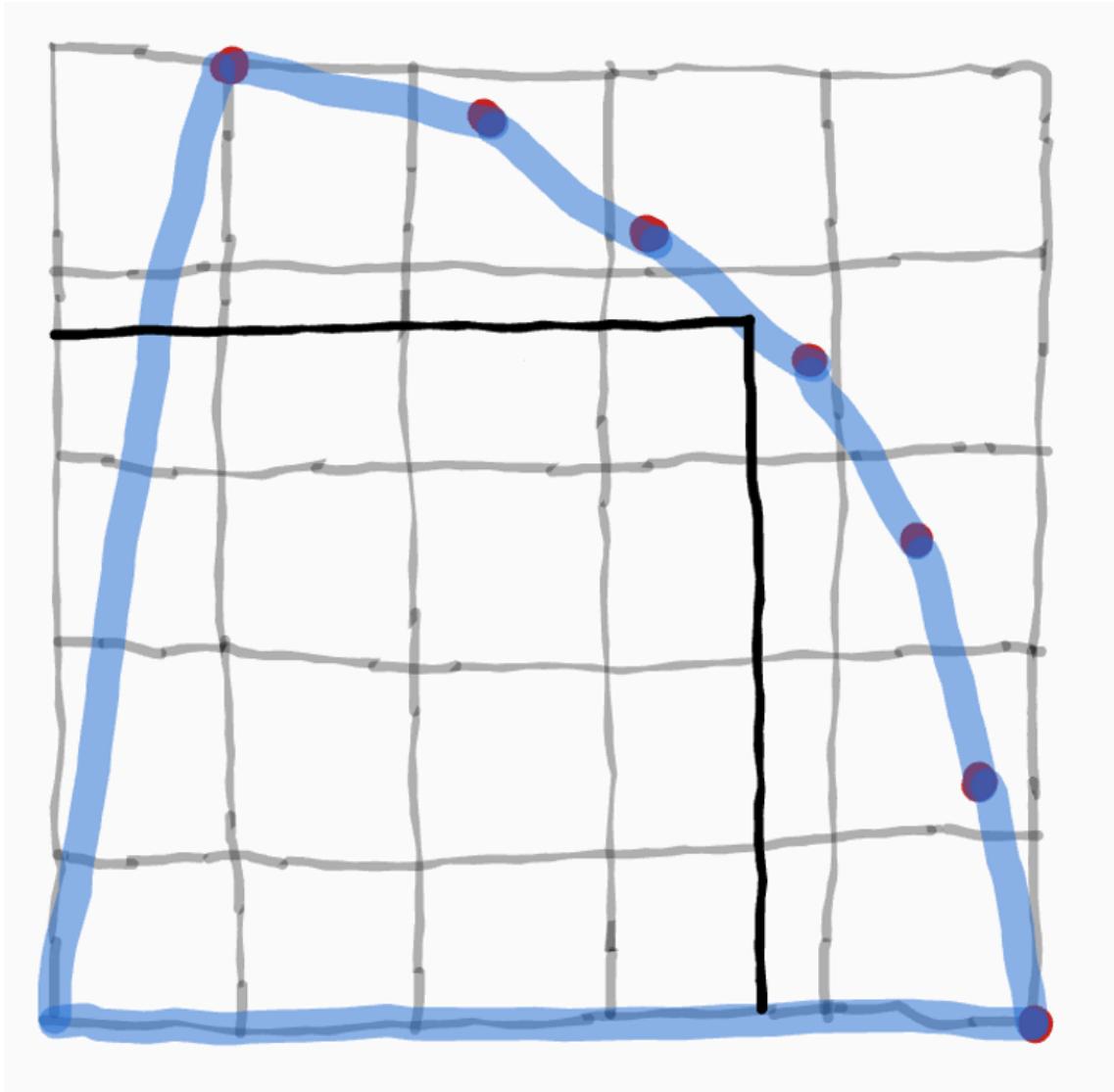
The Nash Bargaining Solution, by contrast, is "pick the point on the frontier that maximizes the area of the rectangle made between that and the disagreement point". It's nonobvious that this process doesn't depend on how we scale or shift the various utility functions, but it's true anyways. Maximizing the area of a rectangle isn't as obvious of a thing to do as "draw a diagonal line". It is pretty mathematically neutral, though.

Also, both the Kalai-Smorodinsky and Nash bargaining solutions happen to agree on which point to pick in the restaurant game, namely, a 2/3 chance of Italian food, and a 1/3 chance of sushi. Although these solutions don't usually coincide.

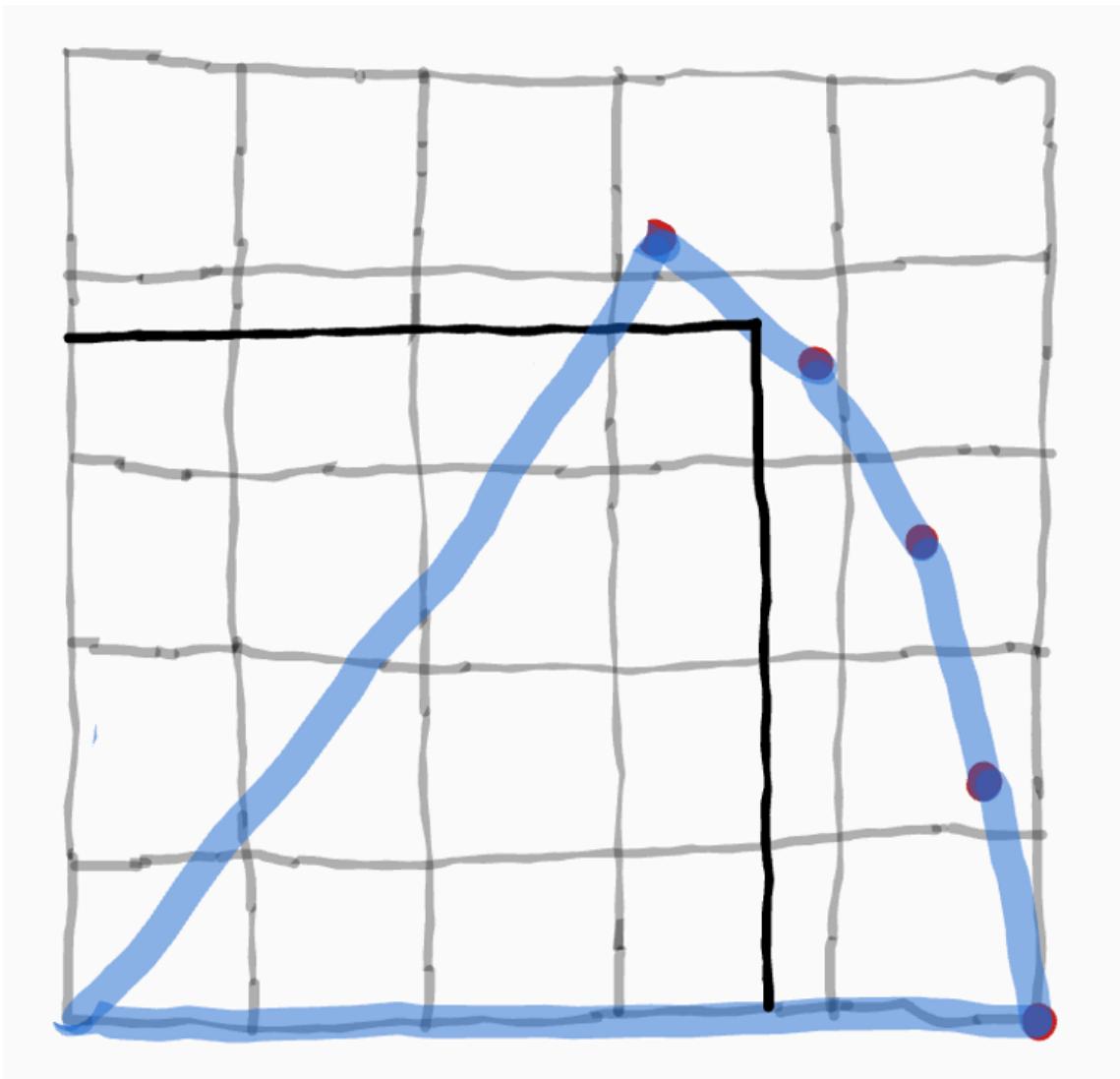


The Nash Bargaining Solution is the only one that fulfills the usual three desiderata, and the axiom of Independence of Irrelevant Alternatives. Ie, if the final bargaining solution involved you doing a 60-40 mix between option D and option E, then deleting any of the options that aren't D or E from the set of available options doesn't affect what happens. Untaken options are irrelevant.

To put it mildly, this is not really a desiderata at all, it's actually an extremely baffling property. Let's say Alice and Bob bargain and hit on the Nash bargaining solution.



Then this axiom is saying that it'd be possible to delete *all* of the options that disproportionately favor Alice, making a game that looks like this, and their bargaining process would still hit the same point.



Intuitively, if options disproportionately favor you, you can use them as "bargaining chips", going "alright, I'll take these unfair options off the table, but only if you remove your unfair options from the table". Independence of Irrelevant Alternatives is basically saying that you can lose all your "unfair bargaining chips" and it'd have no effect on the net outcome!! Phrased like that, it's not clear why anyone would be interested in the Nash bargaining solution.

There are other, more obscure, bargaining solutions which have appeared in the literature, which won't be covered, though they all at least fulfill our basic three criteria.

So, for bargaining games, we can make some progress towards figuring out what a chaan outcome is (Pareto-efficient, scale-and-shift invariant, symmetric), but we don't have enough information yet to single out one particular bargaining solution as The One True Chaa Point, and in fact, it looks like there actually isn't a point like that; the various options all look pretty plausible.

The other issue is that not all games are bargaining games. Bargaining games require everyone to agree on what to do, and there are well-defined disagreement utilities for if negotiations break down. Clearly, this doesn't describe all, or even most, games. Now, it's time to look at another special case of games, for another notion of chaaness.

## Cooperation/Competition Values

For full credit, I was introduced to this notion by [this wonderful post](#), which itself was exposition of [this wonderful paper](#) by Kalai and Kalai.

Instead of bargaining games, we'll now be looking at transferable utility games. A transferable utility game is one where there's a single resource (like dollars) where everyone's utility is linear in that resource, and everyone can pay everyone else in that resource and has enough of the resource to actually do so.

Put another way, bargaining games are like bartering. Both sides must agree on what trade to make, and if either one doesn't like it, the transaction doesn't happen. Transferable utility games are like arbitrary games that take place after money has been invented. There may no longer be a clear disagreement point for what happens when the various parties disagree, but it's also possible for everyone to settle matters by being clever about how they pay each other, which opens up a lot of options.

In particular, when there's a common resource like dollars, you can make everyone express their preferences in terms of dollars. This breaks the usual attribute of utility functions where you can scale and shift them as you please without affecting anything. You can't multiply one player's utilities (as denominated in dollars) by a factor of 100 without doing the same to everyone else. A collective scaling like that, where everyone's numbers go up by 100, is like a currency conversion, shifting from denominating everyone's utilities in dollars to denoting everyone's utilities in cents. It doesn't meaningfully change anything. Interestingly enough, we still do have individual shift-invariance. Put another way, you might be indifferent between option A and option B plus 300 dollars. Then that's consistent with scoring option A at 400 and option B at 100, or you can score option A at 700 and option B at 400. You can add or subtract whatever you want from options A and B, as long as the difference between the two options is 300.

So, in a totally general two-player game, with no well-defined disagreement point, but with the ability to pay each other money, and with everyone's utilities denominated in terms of money, is there some suitably chaa point?

Yes. Time to explain the CoCo value. CoCo stands for Cooperation/Competition, as there's two cases of games where the "right answer" is super-obvious. In pure-cooperation games where both players have the exact same utility function, you just pick the best option in the expectation the foe will do the same. In pure-competition games (ie, zero-sum games), you maximize your worst-case score, as your opponent has perfectly opposing interests to you and so will be minimizing your utility.

As it turns out, when both player's utility functions are commensurable (through this common currency), it's always possible to uniquely split *any* 2-player game at all into two other games. One is a pure-cooperation game, where both players have the same utility function, and perfectly aligned interests. The other is a pure-competition game, where both players have opposite utility functions, and perfectly opposed interests. The CoCo point is "cooperate as much as possible on the cooperative game where our interests align, and fight it out in the zero-sum game where our interests oppose, and add up our results from the two games to figure out how much value we both get".

And so, that's the CoCo point. You pick the most cooperative point in the cooperation game for what to actually do (to maximize the total amount of monetary gain for everyone), and use the results of the competition game to decide how much the two players pay each other, where the zero-sum aspect of the competition game ensures that the budget balances.

Being a bit more formal about this, we'll use A for the function mapping outcomes to player A's utilities, and B for the function mapping outcomes to player B's utilities.

For the cooperation game, both players A and B have the utility functions  $\frac{A+B}{2}$ . Clearly, this is a pure cooperation game.

For the competition game, player A has the utility function  $\frac{B-A}{2}$  and player B has the utility function  $\frac{A-B}{2}$ . Clearly this is a pure competition game, as the utilities for any outcome add up to 0.

And note that for player A, adding up their utilities for the cooperation game and competition game yields  $\frac{A+B}{2} + \frac{B-A}{2} = A$ , ie, their original utility function (and the same for player B)

Here's a concrete example, lifted from the previous post on the topic. Bob and Alice can sell hotdogs at the beach or the airport. If they're at the same location, they end up competing over customers, halving both their profits. Alice is twice as efficient as Bob at selling hotdogs, and the beach has twice as many customers as the airport.

Bob/Alice Beach	Airport
Beach	50/100 100/100
Airport	50/200 25/50

It splits into a cooperation game and a competition game.

Bob/Alice Beach	Airport
Beach	75/75 100/100
Airport	125/125 37.5/37.5

Bob/Alice Beach	Airport
Beach	-25/25 0/0
Airport	-75/75 -12.5/12.5

The best move in the cooperation game is Bob going to the airport, and Alice going to the beach, so that's what's played in real-life. The utility from the cooperation game is added to the maximin utility from the competition game (where beach/beach is played), for 100 Bob utility and 150 Alice utility. And so, the solution is that Alice goes to the beach and pays Bob 50 bucks to go to the airport.

This has a whole lot of good properties, as detailed in the Adam Kalai and Ehud Kalai paper linked above. It's the unique solution that fulfills all of

1: Pareto-optimality, it never leaves monetary value on the table.

2: Shift invariance. If one player gets a gift of 100 dollars at the start of a game, they'll walk out of the game 100 dollars richer than they would if they hadn't received the gift. You can add any constant amount of money to anyone's payoffs and it does nothing.

3: Payoff dominance. If player A gets more money than player B in all cells, then player A will leave the game with more money than player B.

4: Invariance to redundant strategies. Adding a new action that could just as well be accomplished by a probabilistic mix between other actions does nothing.

5: Action monotonicity. Adding a new action is always good for you: you never regret having a larger action space (though other players may regret you having a larger action space).

6: Information monotonicity. This is for the imperfect-information generalization of the CoCo value, that's detailed in the Kalai paper. Giving a player more information about what

everyone else is doing can't hurt them: you'll never regret knowing more.

And the CoCo value is the unique solution that fulfills all six of those properties above. There doesn't seem to be any comparably good notion of equilibrium available besides this, and so we can say that any sensible definition of "chaa" for arbitrary games (if one exists) should manage to recover the CoCo value as a special case when presented with games with transferrable utility.

An interesting note. For bargaining games with transferrable utility (like, a bargaining game where you can pay each other), the equilibrium notion you get is "denominating both player's utility functions in dollars, pick the option that maximizes the overall monetary surplus over the disagreement point, and pay each other so both players equally split the monetary surplus"

Like, if the surplus-maximizing option is one that player 1 values at +80 dollars over the disagreement point, and player 2 values at +40 over the disagreement point, for +120 dollars of surplus value, the CoCo solution is that particular option is picked, and player 1 gives player 2 20 dollars, so both sides walk away with +60 dollars worth of utility.

If Pedro the street vendor and Pierre the rich tourist are haggling over the price of a burrito, and Pedro would walk away at 2\$, and Pierre would walk away at 14\$, then the CoCo solution is that the burrito is sold for 8\$, because that's halfway between where the two people would rather walk.

When arguing over which movie to pick for a group movie night, everyone just needs to report how much they'd value seeing the various movies, pick the movie that maximizes total monetary surplus, and pay each other to equalize that surplus (so you get money if you have to sit through a movie you enjoy less than everyone else in your group, and if you're watching a personal favorite movie that everyone else is "meh" about, like Kong vs Godzilla 5, you've gotta pay the others to watch it.)

Actually, first maximizing surplus value, and then equally splitting the monetary gain, seems quite fair. Yes, we just used the F word.

### **Shapley Value**

Let's say a bunch of people contribute various amounts of effort to a project, for various amounts of gain, creating an overall pile of money. What's a chaa way to fairly divide their pile of money?

We can impose some desiderata.

1: All the money should be going to someone. If the chaa division involved burning money, you should come up with an alternate notion of "chaa" which everyone agrees is better and which is Not That.

2: A player which contributes absolutely nothing to the project and just sits around, regardless of circumstances, should get 0 dollars.

3: If two players in the game are equivalent in all ways and totally interchangeable, then they should receive equal payoffs.

4: If the total pile of money is a times as big, everyone should get a times as much.

5: If two projects are completed in a row and the chaa division occurs, adding together someone's chaa share from project A and project B (considered individually) should be their chaa share from "do both projects in a row". Or, payoffs shouldn't depend on precisely how you slice up the projects.

As it turns out, this *uniquely* pins down how to divide the pile of resources! If  $N$  is the set of all players, and  $i$  is a particular player, and  $v(S)$  (for  $S \subseteq N$ ) is the total amount of resources that could be produced by all the players in  $S$  working together, then the payoff for player  $i$  is

$$\sum_{S \subseteq N / \{i\}} \frac{(n - |S| - 1)!}{(n - |S|)!} v(N \setminus S) - v(S)$$

Put another way, this is effectively going "if the players were added to the group in a random order, and everyone demanded all the marginal extra value they produced upon being added to the group, you'd get payoffs for everyone. Average the payoffs over all possible random orderings". That factorial term at the start is going "what are the odds that group  $S$  gets assembled (in any order), and then I get added to it?". And then the second term is "demanding my marginal contribution".

Here's [a previous post](#) about actually working out the Shapley values in several toy examples of games, to get some intuition for what they're doing.

### **Uniting the Shapley and CoCo Values**

Before we get to the next post tying everything together, we'll see that the Shapley and CoCo values actually have a highly unexpected connection. If you try generalizing the CoCo value to  $n$  players, you get something that looks suspiciously Shapley-like.

Let's begin by reshuffling the Shapley values into a different form. The Shapley value for player  $i$  starts off as

$$\sum_{S \subseteq N / \{i\}} \frac{(n - |S| - 1)!}{(n - |S|)!} v(N \setminus S) - v(S)$$

Now, we can pair off the various coalitions with each other. The subset  $S$  will be paired off with the subset  $N/(S \cup \{i\})$ , the set of all the players that aren't in  $S$  and aren't  $i$ . In particular, note that in both cases, the coefficient in front ends up being  $\frac{(n-|S|-1)!}{(n-|S|-1)!}$ . It's then possible to swap the values around between those two paired coalitions, producing a restatement of the Shapley value as

$$= \sum_{S \subseteq N / \{i\}} \frac{(n - |S| - 1)!}{(n - |S|)!} v(N \setminus S) - v(S)$$

And then, instead of writing this as a sum over subsets that lack player  $i$ , we can switch to the complement and write this as a sum over subsets which include player  $i$ , although the factorial term has to be adjusted a bit to compensate for the fact that the complement of  $S$  has a cardinality of  $n - |S|$  instead of  $|S|$

$$= \sum_{i \in S \subseteq N} \frac{(n - |S|)!}{(n-1)!} (\sum_{j \notin S} u_j) v(N/S)$$

This restatement of the Shapley value will be useful later.

And now we'll try generalizing the CoCo value to n-player games with transferrable utility. Let's deal with a 3-player game, just to make things a bit simpler. The players are A, B, C. As it turns out, this game will actually split into four games instead of two. There's the pure cooperation game, a zero-sum A vs everyone else game, a zero-sum B vs everyone else game, and a zero-sum C vs everyone else game.

For the first game, the utility functions for A, B, and C are  $\frac{A+B+C}{3}$ .

For the zero-sum A vs everyone else game, the utility function for A is  $\frac{A-(B+C)}{2}$ , and the utility functions for B, C are  $\frac{(B+C)-A}{2}$  for both. You might be wondering "why 6?". And the answer is it's that way in order for the game to be zero-sum; the opposing players are weighted less to compensate for there being more of them. Also note that B and C have perfectly aligned incentives in this game, so they might as well perfectly coordinate.

For the zero-sum B vs everyone else game, the utility function for A, C is  $\frac{(A+C)-B}{2}$  for both, and for B it's  $\frac{B-(A+C)}{2}$ .

And similar for C.

For the player A, adding up the payoff for all the games gives you

$$\frac{A+B+C}{3} + \frac{A-(B+C)}{2} + \frac{(A+C)-B}{2} + \frac{B-(A+C)}{2} = A$$

(and similar for all the other players)

And for each game in particular except for the pure cooperation game, it's zero-sum.

Now that we've seen that concrete example, let's generalize to n players. There are  $2^{n-1}$  subgames that the original game will split into, one game for each way to divide the players into two coalitions. Let S be the set of players in one of these coalitions.

For the game with S vs N/S, the utility functions of everyone on coalition S will be

$$\left( \frac{\sum_{j \notin S} u_j}{n-|S|} \right) + \sum_{i \in S} u_i$$

And the utility functions of everyone in the coalition  $N/S$ , will be

$$\frac{1}{(n - |S| - 1)} \left\{ \sum_{j \notin S} u_j - \sum_{i \in S} u_i \right\}$$

It's not too hard to show that all these games are zero-sum (except for the one with the coalition of all players), with perfectly aligned incentives within a coalition.

Anyways, the value that player  $i$  gets is the sum of the values it gets from all of the component games where coalitions compete against each other. Or, the payoff for player  $i$  will be

$$\sum_{i \in S \subseteq N} \min_{\alpha \in \Delta \prod_{j \in S} A_j, \beta \in \Delta \prod_{k \notin S} A_k} \left( \sum_{j \in S} u_j(\alpha, \beta) - \sum_{k \notin S} u_k(\alpha, \beta) \right)$$

Basically, do a weighted sum over "utility of my coalition minus utility of their coalition if the coalitions zero-sum fought" over all the coalitions that you're a part of, and that's your CoCo value in the  $n$ -player case.

But remember, the Shapley value can be re-expressed as

$$\sum_{i \in S \subseteq N} \frac{1}{(n - |S|)!} (v(S) - v(N/S))$$

Which should look suspiciously similar, especially when you remember that  $v(S)$  is the value that everyone on your coalition can produce by working together, and  $v(N/S)$  is the value of the opposing coalition. Really, the CoCo values are just Shapley values but generalized to any sort of game where there's transferable utility. The analogue of "add players in a random order, you get your marginal contribution" turns out to be "add players to a team in a random order, if you're added to team  $S$ , your increase in value from that is the marginal increase in the value of the team if it got into a zero-sum competition against the entire rest of the world."

Ok, so the CoCo values are basically modified Shapley values, so these two are related to each other. Can we generalize even further?

Well, as it turns out, we'll be able to connect the CoCo value to the Nash bargaining solution to get solutions for games in general. I came at this problem from the direction of generalizing the CoCo value to games with nontransferable utility, since the CoCo values were so nicely behaved that any solution for games in general should replicate the CoCo values when utility happens to be transferable, and it turned out my solution automatically spat out the Nash bargaining solution as a special case, which was a considerable surprise to me.

And then it turned out that Harsanyi came up with the same sort of solution from a *completely* different direction (but more elaborate and incorporating constraints that I missed) all the way back in 1963 by trying to generalize the Nash bargaining solution to games with no clear disagreement point. [Next post](#), we'll cover this unifying concept.

# Safetywashing

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

In southern California there's a two-acre butterfly preserve owned by the oil company Chevron. They spend little to maintain it, but many millions on television advertisements featuring it as evidence of their environmental stewardship.<sup>[1]</sup>

Environmentalists have a word for behavior like this: *greenwashing*. Greenwashing is when companies misleadingly portray themselves, or their products, as more environmentally-friendly than they are.

Greenwashing often does cause real environmental benefit. Take the signs in hotels discouraging you from washing your towels:

## SAVE OUR PLANET

Dear Guest,

Bed sheets that are washed daily in thousands of hotels around the world use millions of gallons of water and a lot of detergent.

Please leave this card on the bed if you do not want your sheets changed.



Thank you for helping us conserve the Earth's vital resources. A small recycling symbol consisting of three chasing arrows forming a triangle.

My guess is that the net environmental effect of these signs is in fact mildly positive. And while the most central examples of greenwashing involve deception, I'm sure some of these signs are put up by people who earnestly care. But I suspect hotels might tend to care less about water waste if utilities were less expensive, and that Chevron might care less about El Segundo Blue butterflies if environmental regulations were less expensive.

The field of AI alignment is growing rapidly. Each year it attracts more resources, more mindshare, more people trying to help. The more it grows, the more people will be incentivized to misleadingly portray themselves or their projects as more alignment-friendly than they are.

I think some of this is happening already. For example, a capabilities company launched recently with the aim of training transformers to use every API in the world, which they described as the “safest path to general intelligence.” As I understand it, their argument is that this helps with alignment because it involves collecting feedback about people’s preferences, and because humans often wish AI systems could more easily take actions in the physical world, which is easier once you know how to use all the APIs.<sup>[2]</sup>

It’s easier to avoid things that are easier to notice, and easier to notice things with good handles. So I propose adopting the handle “safetywashing.”

1. ^

From what I can tell, the [original source](#) for this claim is the book “The Corporate Planet: Ecology and Politics in the Age of Globalization,” which from my samples seems about as pro-Chevron as you’d expect from the title. So I wouldn’t be stunned if the claim were misleading, though the numbers passed my sanity check, and I did confirm the [preserve](#) and [advertisements](#) exist.

2. ^

I haven’t talked with anyone who works at this company, and all I know about their plans is from the copy on their website. My guess is that their project harms, rather than helps, our ability to ensure AGI remains safe, but I might be missing something.

# Connor Leahy on Dying with Dignity, EleutherAI and Conjecture

This is a linkpost for <https://theinsideview.ai/connor2>

I talked to Connor Leahy about Yudkowsky's antimemes in [Death with Dignity](#), common misconceptions about EleutherAI and his new AI Alignment company [Conjecture](#).

Below are some highlighted quotes from our conversation (available on [Youtube](#), [Spotify](#), [Google Podcast](#), [Apple Podcast](#)). For the full context for each of these quotes, you can find an accompanying [transcript](#), organized in [74 sub-sections](#).

## Understanding Eliezer Yudkowsky

### Eliezer Has Been Conveying Antimemes

"Antimemes are completely real. There's nothing supernatural about it. **Most antimemes are just things that are boring. So things that are extraordinarily boring are antimemes because they, by their nature, resist you remembering them.** And there's also a lot of antimemes in various kinds of sociological and psychological literature. A lot of psychology literature, especially early psychology literature, which is often very wrong to be clear. Psychoanalysis is just wrong about almost everything. **But the writing style, the kind of thing these people I think are trying to do is they have some insight, which is an antimeme. And if you just tell someone an antimeme, it'll just bounce off them.** That's the nature of an antimeme. So to convey an antimeme to people, you have to be very circuitous, often through fables, through stories you have, through vibes. This is a common thing."

Moral intuitions are often antimemes. Things about various human nature or truth about yourself. Psychologists, don't tell you, "Oh, you're fucked up, bro. Do this." That doesn't work because it's an antimeme. People have protection, they have ego. You have all these mechanisms that will resist you learning certain things. Humans are very good at resisting learning things that make themselves look bad. So things that hurt your own ego are generally antimemes. So **I think a lot of what Eliezer does and a lot of his value as a thinker is that he is able, through however the hell his brain works, to notice and comprehend a lot of antimemes that are very hard for other people to understand.**"

### Why the Dying with Dignity Heuristic is Useful

"The whole point of the post is that if you do that, and you also fail the test by thinking that blowing TSMC is a good idea, you are not smart enough to do this. Don't do it. If you're smart enough, you figured out that this is not a good idea... Okay, maybe. But most people, or at least many people, are not smart enough to

be consequentialists. So if you actually want to save the world, you actually want to save the world... **If you want to win, you don't want to just look good or feel good about yourself, you actually want to win, maybe just think about dying with dignity instead.** Because even though you, in your mind, don't model your goal as winning the world, **the action that is generated by the heuristic will reliably be better at actually saving the world."**

"There's another interpretation of this, which I think might be better where **you can model people like AI\_WAIFU as modeling timelines where we don't win with literally zero value.** That there is zero value whatsoever in timelines where we don't win. **And Eliezer, or people like me, are saying, 'Actually, we should value them in proportion to how close to winning we got'.** **Because that is more healthy... It's reward shaping!** We should give ourselves partial reward for getting partially the way. He says that in the post, how we should give ourselves dignity points in proportion to how close we get.

And this is, in my opinion, a much psychologically healthier way to actually deal with the problem. This is how I reason about the problem. **I expect to die. I expect this not to work out. But hell, I'm going to give it a good shot and I'm going to have a great time along the way.** I'm going to spend time with great people. I'm going to spend time with my friends. We're going to work on some really great problems. And if it doesn't work out, it doesn't work out. But hell, we're going to die with some dignity. We're going to go down swinging."

"If you have to solve an actually hard problem in the actual real world, in actual physics, for real, an actual problem, that is actually hard, **you can't afford to throw your epistemics out the door because you feel bad. And if people do this, they come up with shit like, 'Let's blow up to TSMC'.** Because they throw their epistemics out the window and like, 'This feels like something. Something must be done and this is something, so therefore it must be done'."

## EleutherAI

### Why training GPT-3 Size Models made sense

"Well, I remember having these conversations with some people in the alignment sphere, where they're like, "Oh well, why did you build the models? Just use GPT-2, that's fine." I'm like, "Well, okay, what if I want to see the bigger properties?" And they'll be like, "They'll probably exist in the smaller models too or something. Name three experiments you're going to do with this exact model." And I'm like, "I could come up with three, sure. But that's kind of missing the point." The point is: **we should just really stare at these things really fucking hard. And turns out, in my experience, that was a really good idea. Most of my knowledge, my competitive advantage is gained from that period of just actually building the things,** actually staring at them really hard and not just knowing about the OpenAI API existing and reading the papers. There's a lot of knowledge you can get from reading a handbook, but actually running the machine will teach you a lot of things."

# EleutherAI Spread Alignment Memes in the ML World

**"One of the important parts of my threat model is that I think 99% of the damage from GPT-3 was done the moment the paper was published.** And, as they say about the nuclear bomb, the only secret was that it was possible. And I think there's a bit of naivety that sometimes goes into these arguments, where people are, 'Well, EleutherAI accelerated things, they drew attention to the meme'. And I think there's a lot of hindsight bias there, in that people don't realize how everyone knew about this, except the alignment community. **Everyone at OpenAI, Google Brain and DeepMind. People knew about this, and they figured it out fucking fast.**"

"One of the things that EleutherAI did, and this was very much intentional, is that **it created a space that is open to the wider ML community and their norms.** It is respectful of AI researchers and their norms. And we also have street cred, in the sense that we are ML researchers and we're not just some dude talking about logical induction or whatever, but still has a very strong alignment meme. **Alignment is high status. It is a respectful thing to talk about, a thing to take seriously.** It is not some weird thing some people in Berkeley think about. It is a serious topic of serious intrigue. And for what it's worth, **of the five core people at EleutherAI that changed their job as a direct consequence of EleutherAI, four went into alignment.**"

"I'm not saying, was it a resounding success? Did it do everything I wanted? No. It could always have been better. But I like to believe that there was a positive magnetic contagion that happened there. As I say, a lot of people that I know, that were an ML, started taking alignment seriously. **I know several professors at several universities that'd gone to EleutherAI through the scaling memes, and then became convinced that this alignment thing seems important potentially.**"

## On the Policy and Impact of EleutherAI's Open Source

**"Our official position, which you can read in our blog, which has always been there, is that not everything should be released. And in fact, we, EleutherAI, discovered at least two capabilities advancements ahead of anyone else in the world, and we successfully kept them secret,** because we were like "Oh shit". One is the chain of thought prompting idea, which we then later published. I believe I showed Eliezer the pre-draft. So he may be able to confirm that I'm not bullshitting you on this. I think it was Eliezer that I showed that to. And so in that regard, I fully understand why people think this, because that's a default open-source thing. And there're several other open-source groups now, that have split off from Eleuther or they're distant cousins of Eleuther, that do think this way. I strongly disagree with them. And I think that what they're making is not a good idea. It was always contingent. **EleutherAI's policy was**

**always "we think this specific thing should be open". Not all things should be open, but this specific thing that we are thinking about right now, that we're talking about right now, this specific thing we think should be open for this, this, this and this reason.** But there are other things which we may or may not encounter, which shouldn't be open. We made very clear if we ever had a quadrillion parameter model for some reason, we would not release it."

**"Again, I want to be very clear here. It may have been a mistake to release GPT-J. It may have been a mistake. I don't think it is one, for various contingent reasons,** but I'm not ideologically committed to the idea that this was definitely the right thing to do. I think given the evidence that I've seen, **for example, GPT-J being used in some of my favorite interpretability papers, such as the Editing and Eliciting Knowledge paper from David Bau's lab**, which is an excellent paper, and you really should read. And **several other groups such as Redwood using GPT-Neo models in their research and such.** I think that there are a lot of reasons why this was helpful to some people, this was good. Also, the **tacit knowledge that we've gained has been very instrumental for setting up Conjecture** and what I do now. So I think there are reasons why it was good, but I could be wrong about this. Again, if people disagree with me about that, I think I disagree, but I think that it's not insane."

# Conjecture

## How Conjecture Started

"So Conjecture grew a lot out of some of the bottlenecks I found while working in EleutherAI. So EleutherAI was great. I love the people there and such. Anyway, we had a lot of great people and such. But **if you wanted to get something done, it was like herding cats. But imagine the cats also have crippling ADHD and are the smartest people you've ever met.** Especially if anything boring needed to get done, if we needed to fix some bugs or scrape some data or whatever, it would very often just not get done. Because it was all volunteer based, right? You wanted to do fun things. It's your free time. People don't want to do boring shit. During the pandemic it was a bit different, because people literally didn't have anything really to do. But now you have a social life again, you have a job. And then you don't want to come home and spend two hours debugging some goddamn race condition or whatever."

"So, the idea was first floated very early in EleutherAI, but I put that completely on ice. I didn't want to do that. I wanted to just focus on open-source and such. So **it became really concrete around late 2021, September-October I think, when Nat Friedman, who was the CEO of GitHub at the time, approaches EleutherAI** and says, 'Hey, I love what you guys are doing. It's super awesome. Can help you with anything? You want to meet up sometime?'. And, to add to his credit, he donated a bunch of money to help EleutherAI to keep going. A man of his word. And he happened to be in Germany at the time, which was where I was

as well. And he was, 'Hey, do you want to meet up for a coffee?' And so we met up, really got along, and he was, 'Hey, you ever thought of doing a company or something?' 'Now, I have been thinking about that.' 'Why don't you just come by the Bay sometime and talk' and such. And so I was thinking, 'Oh cool, I can go to the Bay and I can...' **So it was a confluence of factors, right? It was an excuse to go to the Bay to talk to both Nat and his friends, but also talk to Open Phil and potential EA funders and stuff like that.** And also, I was getting on EleutherAI, I was hitting those bottlenecks I was talking about, where I was trying to do research on EleutherAI but it just wasn't working."

## Where Conjecture Fits in the AI Alignment Landscape

"Conjecture differs from many other orgs in the field by various axes. So one of the things is that we take short timelines very seriously. **There's a lot of people here and there that definitely entertain the possibility of short timelines or think it's serious or something. But no real org that is fully committed to five year timelines, and act accordingly.** And we are an org that takes this completely seriously. **Even if we just have 30% on it happening, that is enough in our opinion, to be completely action relevant. Just because there are a lot of things you need to do if this is true, compared to 15-year timelines, that no one's doing, that it seems it's worth trying.** So we have very short timelines. We think alignment is very hard. So the thing where we disagree with a lot of other orgs, is we expect alignment to be hard, the kind of problem that just doesn't get solved by default. That doesn't mean it's not solvable. So where I disagree with Eliezer is that, I do think it is solvable... he also thinks it's solvable. He just doesn't think it's solvable in time, which I do mostly agree on. So **I think if we had a hundred years time, we would totally solve this. This is a problem that can be solved, but doing it in five years with almost no one working on it, and also we can't do any tests with it because if we did a test, and it blows up, it's already too late, et cetera, et cetera... There's a lot of things that make the problem hard.**"

**"One of the positive things that I've found is just, no matter where I go, the people working in the AGI space specifically are overwhelmingly very reasonable people.** I may disagree with them, I think they might be really wrong about various things, but they're not insane evil people, right? They have different models of how reality works from me, and they're like... You know, **Sam Altman replies to my DMs on Twitter, right? [...] I very strongly disagree with many of his opinions, but the fact that I can talk to him is not something we should have taken for a given.** This is not the case in many other industries, and there's many scenarios where this could go away, and we don't have this thing that everyone in the space knows each other, or can call each other even. So I may not be able to convince Sam of my point of view. The fact I can talk to him at all is a really positive sign, and a sign that I would not have predicted two years ago."

# Why Conjecture is Doing Interpretability Research

"I think it's really hard for modern people to put themselves into an epistemic state of just how it was to be a pre-scientific person, and just how confusing the world actually looked. And now even things that we think of as simple, how confusing they are before you actually see the solution. So **I think it is possible, not guaranteed or even likely, but it's possible, that such discoveries could not be far down the tech tree**, and that if we just come at things from the right direction, we try really hard, we try new things, **that we would just stumble upon something where we're just like, 'Oh, this is okay, this works. This is a frame that makes sense. This deconfuses the problem. We're not so horribly confused about everything all the time.'**"

## Conjecture Approach To Solving Alignment

"If you need to roll high, roll many dice. **At Conjecture, the ultimate goal is to make a lot of alignment research happen, to scale alignment research**, to scale horizontally, to tile research teams efficiently, to take in capital and convert that into efficient teams with good engineers, good op support, access to computers, et cetera, et cetera, **trying different things from different direction, more decorrelated bets.**"

**"To optimize the actual economy is just computationally impossible.** You would have to simulate every single agent, every single thing, every interaction, just impossible. So **instead what they do is, they identify a small number of constraints that, if these are enforced, successfully shrink the dimension of optimization down to become feasible to optimize within.** [...] If you want to reason about how much food will my field produce, monoculture is a really good constraint. By constraining it by force to only be growing, say, one plant, you simplify the optimization problem sufficiently that you can reason about it. **I expect solutions to alignment, or, at least the first attempts we have at it, to look kind of similar like this. It'll find some properties. It may be myopia or something, that, if enforced, if constrained, we will have proofs or reasons to believe that neural networks will never do X, Y, and Z.** So maybe we'll say, 'If networks are myopic and have this property and never see this in the training data, then because of all this reasoning, they will never be deceptive.' Something like that. Not literally that, but something of that form."

"There is this meme, which is luckily not as popular as it used to be, but **there used to be a very strong meme that neural networks are these uninterpretable black boxes.** [...] **That is just actually wrong. That is just legitimately completely wrong, and I know this for a fact. There is so much structure inside of neural networks.** Sure, some of it is really complicated and not obviously easy to understand for a human, but there is so much structure there, and there are so many things we can learn from actually

really studying these internal parts... again, staring at the object really hard actually works."

## On being non-disclosure by default

**"We are non-disclosure by default, and we take info hazards and general infosec and such very seriously.** So the reasoning here is not that we won't ever publish anything. I expect that we will publish a lot of the work that we do, especially the interpretability work, I expect us to publish quite a lot of it, maybe mostly all of it, but the way we think about info hazards or general security and this kind of stuff, is that we think it's quite likely that there are relatively simple ideas out there that may come up during the doing of prosaic alignment research that cannot really increase capabilities, that we are messing around with a neural network to try to make it more aligned, or to make it more interpretable or something, and suddenly, it goes boom, and then suddenly it's five times more efficient or something. I think things like this can and will happen, and for this reason, it's very important for us to... **I think of info hazard policy, kind of like wearing a seatbelt. It's probably where we'll release most of our stuff, but once you release something into the wild, it's out there. So by default, before we know whether something is safe or not, it's better just to keep our seat belt on and just keep it internal.** So that's the kind of thinking here. It's a caution by default. I expect us to work on some stuff that probably shouldn't be published. I think a lot of prosaic alignment work is necessarily capabilities enhancing, making a model more aligned, a model that is better at doing what you wanted to do, almost always makes the model stronger."

**"I want to have an organization where it costs you zero social capital to be concerned about keeping something secret.** So for example, **with the Chinchilla paper,** what I've heard is, inside of DeepMind, there was quite a lot of pushback against keeping it secret. Apparently, **the safety teams wanted to not publish it, and they got a lot of pushback from the capabilities people because they wanted to publish it.** And that's just a dynamic I don't want to exist at Conjecture. I want to be the case that the safety researchers say "Hey, this is kind of scary. Maybe we shouldn't publish it" and that is completely fine. They don't have to worry about their jobs. They still get promotions, and it is normal and okay to be concerned about these things. That doesn't mean we don't publish things. If everyone's like, "Yep, this is good. This is a great alignment tool. We should share this with everybody," then we'll release, of course."

## On Building Products as a For-Profit

**"The choice to be for profit is very much utilitarian.** So it's actually quite funny that on FTX future funds' FAQ, they actually say they suggest to many non-profits to actually try to be for profits if they can. Because **this has a lot of good benefits such as being better for hiring, creating positive feedback loops and potentially making them much more long-term sustainable.** So the main reason I'm interested [in being a for-profit] is long term sustainability and the positive feedback loops, and also the hiring is nice. So I think there's like a lot of positive things about for-profit companies. There's a lot of negative things, but like it's also a lot of positive things and a lot of negative things with non-profits too,

that I think get slipped under the rug in EA. Like **in EA it feels like the default is a non-profit and you have to justify going outside of the Overton window.**"

"The way I think about products at the moment is, I basically think that there are the current state-of-the-art models that have opened this exponentially large field of possible new products that has barely been tapped. **GPT-3 opens so many potential useful products that just all will make profitable companies and someone has to pick them. I think without pushing the state of the art at all, we can already make a bunch of products that will be profitable.** And most of them are probably going to be relatively boring [...] **You want to do a SaaS product, something that helps you with some business task. Something that helps you make a process more efficient inside of a company or something like that. There' tons of these things, which are just like not super exciting, but they're like useful.**"

## Scaling The Alignment Field

**"Our advertising quote, unquote, is just like one LessWrong post that was like, "Oh, we're hiring". Right? And we got a ton of great application. Like the signal to noise was actually wild. Like one in three applications were just really good, which like never happens.** So, like, incredible. So we got to hire some really phenomenal people for our first hiring round. And so at this point we're already basically at a really enviable position. I mean, it's like, it's annoying, but it's a good problem to have, where we're basically already funding constrained. **We're at the point where I have people I want to hire projects for them to do and the management capacity to handle them. And I just don't have the funding at the moment to hire them.**"

**"Conjecture is an organization that is directly tackling the alignment problem and we're a de-correlated bet from the other ones.** I'm glad, I'm super glad that Redwood and Anthropic are doing the things they do, but they're kind of doing a very similar direction of alignment research. We're doing something very different and we're doing it at a different location. **We have access to a whole new talent pool of European talent that cannot come to the US. We get a lot of new people into the field. We also have the EleutherAI people coming in, different research directions and de-correlated bets. And we can scale. We have a lot of operational capacity, a lot of experience and also entrepreneurial vigor."**

# A note about differential technological development

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

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

For a simplified version of my model as to why:

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

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

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

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

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

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

the problem—a pattern which seems common to me in my read of the development of things like analysis, meta-mathematics, quantum physics, etc.).

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

---

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

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

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

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

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

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

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

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

---

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

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

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

---

Some nuance I'll add:

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

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

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

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

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

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

1. ^

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

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

2. ^

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

# Humans provide an untapped wealth of evidence about alignment

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

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

---

**TL;DR:** To even consciously consider an alignment research direction, [you should have evidence](#) to locate it as a promising lead. As best I can tell, many directions seem interesting but do not have strong evidence of being “entangled” with the alignment problem such that I expect them to yield significant insights.

For example, “we can solve an easier version of the alignment problem by first figuring out how to build an AI which maximizes the number of real-world diamonds” has intuitive appeal and plausibility, but this claim doesn’t *have* to be true and this problem does not *necessarily* have a natural, compact solution. In contrast, there *do in fact* exist humans who care about diamonds. Therefore, there are guaranteed-to-exist alignment insights concerning the way people come to care about e.g. real-world diamonds.

*“Consider how humans navigate the alignment subproblem you’re worried about” is a habit which I (TurnTrout) picked up from Quintin Pope. I wrote the post, he originated the tactic.*

---

A simplified but still very difficult open problem in [AI alignment](#) is to state an unbounded program implementing a [diamond maximizer](#) that will turn as much of the physical universe into diamond as possible. The goal of “making diamonds” was chosen to have a crisp-seeming definition for our universe (the amount of diamond is the number of carbon atoms covalently bound to four other carbon atoms). If we can crisply define exactly what a ‘diamond’ is, we can avert issues of trying to convey [complex values](#) into the agent.

[Ontology identification problem](#), Arbital

I find this problem interesting, both in terms of wanting to know how to solve a reframed version of it, and in terms of what I used to think about the problem. I used to<sup>[1]</sup> think, “yeah, ‘diamond’ is relatively easy to define. Nice [problem relaxation](#).” It felt like the diamond maximizer problem let us focus on the challenge of making the AI’s values bind to *something at all* which we actually intended (e.g. diamonds), in a way that’s robust to ontological shifts and that doesn’t collapse into wireheading or tampering with e.g. the sensors used to estimate the number of diamonds.

Although the details are mostly irrelevant to the point of this blog post, the Arbital article suggests some solution ideas and directions for future research, including:

1. Scan [AIXI-tl](#)'s Turing machines and locate diamonds within their implicit state representations.
2. Given how [inaccessible](#) we expect AIXI-tl's representations to be by default, have AIXI-tl just consider a Turing-complete hypothesis space which uses more interpretable representations.
3. "Being able to describe, in purely theoretical principle, a prior over epistemic models that have at least two levels and can switch between them in some meaningful sense"

Do you notice anything *strange* about these three ideas? Sure, the ideas don't seem workable, but they're good initial thoughts, right?

The problem *isn't* that the ideas aren't clever enough. Eliezer is pretty dang clever, and these ideas are reasonable stabs given the premise of "get some AIXI variant to maximize diamond instead of reward."

The problem *isn't* that it's impossible to specify a mind which cares about diamonds. We already know that there are intelligent minds who value diamonds. You might be dating one of them, or you might even *be* one of them! Clearly, the genome + environment jointly specify certain human beings who end up caring about diamonds.

One problem is [where is the evidence required to locate these ideas](#)? Why should I even find myself thinking about diamond maximization and AIXI and Turing machines and utility functions in this situation? It's not that there's *no* evidence. For example, utility functions [ensure the agent can't be exploited in some dumb ways](#). But I think that the supporting evidence is not *commensurate* with the specificity of these three ideas or with the specificity of the "ontology identification" problem framing.

Here's an exaggeration of how these ideas feel to me when I read them:

"I lost my phone", you tell your friend.

They ask, "Have you checked [Latitude: -34.44006, Longitude: -64.61333](#)?"

Uneasily, you respond: "Why would I check there?"

Your friend shrugs: "Just seemed promising. And it's on land, it's not in the ocean. Don't worry, I incorporated evidence about where you probably lost it."

I [recently made a similar point](#) about [Cooperative Inverse Reinforcement Learning](#):

*Against CIRL as a special case of against quickly jumping into highly specific speculation while ignoring empirical embodiments-of-the-desired-properties.*

In the context of "how do we build AIs which help people?", asking "does CIRL solve corrigibility?" is hilariously unjustified. [By what evidence](#) have we located such a specific question? We have assumed there is an achievable "corrigibility"-like property; we have assumed it is good to have in an AI; we have assumed it is good in a similar way as "helping people"; we have elevated CIRL in particular as a formalism worth inquiring after.

But this is ***not the first question to ask***, when considering "sometimes people want to help each other, and it'd be great to build an AI which helps us in some way." Much better to start with *existing* generally intelligent systems (humans)

which *already* sometimes act in the way you want (they help each other) and ask after the **guaranteed-to-exist reason** why this empirical phenomenon happens.

Now, if you are confused about a problem, it can be better to explore *some* guesses than no guesses—perhaps it's better to think about Turing machines than to stare helplessly at the wall (but perhaps not). Your best guess may be wrong (e.g. write a utility function which scans Turing machines for atomic representations of diamonds), but you sometimes still learn something by spelling out the implications of your best guess (e.g. the ontology identifier stops working when AIXI Bayes-updates to non-atomic physical theories). This can be productive, as long as you keep in mind the wrongness of the concrete guess, so as to not become anchored on that guess or on the framing which originated it (e.g. build a diamond *maximizer*).

However, in this situation, I want to look elsewhere. When I confront a confusing, difficult problem (e.g. how do you create a mind which cares about diamonds?), I often first look at reality (e.g. are there any existing minds which care about diamonds?). Even if I have *no idea* how to solve the problem, if I can find an existing mind which cares about diamonds, then *since that mind is real*, that mind has a [guaranteed-to-exist causal mechanistic play-by-play origin story](#) for why it cares about diamonds. I thereby anchor my thinking to reality; reality is sturdier than “what if” and “maybe this will work”; many human minds *do* care about diamonds.

In addition to “there’s a guaranteed causal story for humans valuing diamonds, and not one for AIXI valuing diamonds”, there’s a second benefit to understanding how human values bind to the human’s beliefs about real-world diamonds. This second benefit is practical: I’m pretty sure the way that *humans* come to care about diamonds has nearly nothing to do with the ways AIXI-*t* might be motivated to maximize diamonds. This matters, because I expect that the first AGI’s value formation will be *far* more mechanistically similar to within-lifetime human value formation, than to AIXI-*t*’s value alignment dynamics.

Next, it *can* be true that the existing minds are too hard for us to understand in ways relevant to alignment. One way this could be true is that human values are a “[mess](#)”, that “[our brains are kludges slapped together by natural selection](#).“ If human value formation were sufficiently complex, with sufficiently many load-bearing parts such that each part drastically affects human alignment properties, then we might instead want to design simpler human-comprehensible agents and study *their* alignment properties.

While I think that human *values* are complex, I think the evidence for human value *formation*’s essential complexity is surprisingly weak, all things reconsidered in light of modern, post-deep learning understanding. Still... maybe humans are too hard to understand in alignment-relevant ways!

But, I mean, come on. Imagine an alien<sup>[2]</sup> visited and told you:

Oh yeah, the AI alignment problem. We knocked that one out a while back. [Information inaccessibility of the learned world model](#)? No, I’m pretty sure [we didn’t solve that](#), but we didn’t have to. We built this protein computer and trained it with, I forget actually, was it just what you would call “deep reinforcement learning”? Hm. Maybe it was more complicated, maybe not, I wasn’t involved.

We *might* have hardcoded relatively crude reward signals that are basically defined over sensory observables, like a circuit which activates when their sensors detect a [certain kind of carbohydrate](#). Scanning you, it looks like some of the protein computers ended up with *your values*, even. Small universe, huh?

Actually, I forgot how we did it, sorry. And I can't make guarantees that our approach scales beyond your intelligence level or across architectures, but maybe it does. I have to go, but here are a few billion of the trained protein computers if you want to check them out!

Ignoring the weird implications of the aliens existing and talking to you like this, and considering only the alignment implications—*The absolute top priority of many alignment researchers should be figuring out how the hell the aliens got as far as they did.*<sup>[3]</sup> Whether or not you know if their approach scales to further intelligence levels, whether or not their approach seems easy to understand, you have learned that these computers are *physically possible, practically trainable entities*. These computers have definite existence and guaranteed explanations. Next to these actually existent computers, speculation like “maybe [attainable utility preservation](#)” leads to cautious behavior in AGIs” is dreamlike, unfounded, and untethered.

If it turns out to be currently too hard to understand the aligned protein computers, then I want to keep coming back to the problem with each major new insight I gain. When I learned about [scaling laws](#), I should have rethought my picture of human value formation—Did the new insight knock anything loose? I should have checked back in when I heard about [mesa optimizers](#), about the [Bitter Lesson](#), about [the feature universality hypothesis](#) for neural networks, about [natural abstractions](#).

Because, given my life's present ambition (solve AI alignment), that's what it makes sense for me to do—at each major new insight, to reconsider my models<sup>[4]</sup> of the *single known empirical example of general intelligences with values*, to scour the Earth for every possible scrap of evidence that humans provide about alignment. We may not get much time with human-level AI before we get to superhuman AI. But we get plenty of time with human-level humans, and we get plenty of time *being* a human-level intelligence.

The way I presently see it, [the godshatter of human values](#)—the rainbow of desires, from friendship to food—is only [unpredictable](#) relative to a class of hypotheses which fail to predict the shattering.<sup>[5]</sup> But confusion is in the map, not the territory. I do not consider human values to be “unpredictable” or “weird”, I do not view them as a “hack” or a “kludge.” Human value formation may or may not be messy (although I presently think *not*). Either way, human values are, of course, part of our lawful reality. Human values are reliably produced by within-lifetime processes within the brain. This has an explanation, though I may be ignorant of it. Humans usually bind their values to certain objects in reality, like dogs. This, too, has an explanation.

And, to be clear, I don't want to black-box outside-view extrapolate from the “human datapoint”; I don't want to focus on thoughts like “Since alignment ‘works well’ for dogs and people, maybe it will work well for slightly superhuman entities.” I aspire for the kind of alignment mastery which lets me build a diamond-producing AI, or if that didn't suit my fancy, I'd turn around and tweak the process and the AI would press green buttons forever instead, or—if I were playing for real—I'd align that system of mere circuitry with humane purposes.

For that ambition, the inner workings of those generally intelligent apes is *invaluable evidence* about the *mechanistic within-lifetime process by which those apes form their values*, and, more generally, about how intelligent minds can form values at all. What factors matter for the learned values, what factors don't, and what we should do for AI. Maybe humans have special inductive biases or architectural features, and without those, they'd grow totally different kinds of values. But if that were true, wouldn't that be important to know?

If I knew how to interpret the available evidence, I probably *would* understand how I came to weakly care about diamonds, and what factors were important to that process (which reward circuitry had to fire at which frequencies, what concepts I had to have learned in order to grow a value around "diamonds", how precisely activated the reward circuitry had to be in order for me to end up caring about diamonds).

Humans provide huge amounts of evidence, *properly interpreted*—and therein lies the grand challenge upon which I am presently fixated. In an upcoming post, I'll discuss one particularly rich vein of evidence provided by humans. (EDIT 1/1/23: See [this shortform comment](#).)

*Thanks to Logan Smith and Charles Foster for feedback. Spiritually related to but technically distinct from [The First Sample Gives the Most Information](#).*

EDIT: In this post, I wrote about the Arbital article's unsupported jump from "Build an AI which cares about a simple object like diamonds" to "Let's think about ontology identification for AIXI-tl." The point is not that there is no valid reason to consider the latter, but that the jump, as written, seemed evidence-starved. For separate reasons, I currently think that ontology identification is unattractive in some ways, but this post isn't meant to argue against that framing in general. The main point of the post is that humans provide tons of evidence about alignment, by virtue of containing guaranteed -to-exist mechanisms which produce e.g. their values around diamonds.

## Appendix: One time I didn't look for the human mechanism

Back in 2018, I had [a clever-seeming idea](#). We don't know how to build an aligned AI; we want multiple tries; it would be great if we could build an AI which "[knows it may have been incorrectly designed](#)"; so why not have the AI simulate its probable design environment over many misspecifications, and then *not* do plans which tend to be horrible for most initial conditions. While I drew some inspiration from how I would want to reason in the AI's place, I ultimately did not think thoughts like:

We know of a single group of intelligent minds who have ever wanted to be corrigible and helpful to each other. I wonder how that, in fact, happens?

Instead, I was trying out clever, off-the-cuff ideas in order to solve e.g. Eliezer's formulation of the [hard problem of corrigibility](#). However, my idea and his formulation suffered a few disadvantages, including:

1. The formulation is not guaranteed to describe a probable or "natural" kind of mind,
2. These kinds of "corrigible" AIs are not guaranteed to produce desirable behavior, but only *imagined* to produce good behavior,

3. My clever-seeming idea was not at all constrained by reality to actually work in practice, as opposed to just sounding clever to me, and
4. I didn't have a concrete use case in mind for what to *do* with a "corrigible" AI.

I wrote this post as someone who previously needed to read it.

1. ^

I now think that diamond's physically crisp definition is a red herring. More on that in future posts.

2. ^

This alien is written to communicate my current belief state about how human value formation works, so as to make it clear why, *given* my beliefs, this value formation process is so obviously important to understand.

3. ^

There is an additional implication present in the alien story, but not present in the evolutionary production of humans. The aliens are implied to have *purposefully* aligned some of their protein computers with human values, while evolution is not similarly "purposeful." This implication is noncentral to the key point, which is that the human-values-having protein computers exist in reality.

4. ^

Well, I didn't even *have* a detailed picture of human value formation back in 2021. I thought humans were hopelessly dumb and messy and we want a *nice clean AI which actually is robustly aligned*.

5. ^

Suppose we model humans as the "inner agent" and evolution as the "outer optimizer"—I think [this is, in general, the wrong framing](#), but let's roll with it for now. I would guess that Eliezer believes that [human values are an unpredictable godshatter](#) with respect to the outer criterion of inclusive genetic fitness. This means that if you reroll evolution many times with perturbed initial conditions, you get inner agents with dramatically different values each time—it means that human values are akin to a raindrop which happened to land in some location for no grand reason. I notice that I have medium-strength objections to this claim, but let's just say that he is correct for now.

I think this unpredictability-to-evolution doesn't matter. We aren't going to reroll evolution to get AGI. Thus, for a variety of reasons too expansive for this margin, I am little moved by analogy-based reasoning along the lines of "here's the one time inner alignment was tried in reality, and evolution failed horribly." I think that historical fact is mostly irrelevant, for reasons I will discuss later.

# «Boundaries», Part 1: a key missing concept from utility theory

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

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

This is Part 1 of my [«Boundaries» Sequence](#) on LessWrong.

Summary: «Boundaries» are a missing concept from the axioms of game theory and bargaining theory, which might help pin down certain features of multi-agent rationality (this post), and have broader implications for effective altruism discourse and x-risk (future posts).

## 1. Boundaries (of living systems)

*Epistemic status: me describing what I mean.*

With the exception of some relatively recent and isolated pockets of research on embedded agency (e.g., [Orseau & Ring, 2012](#); [Garrahan & Demsky, 2018](#)), most attempts at formal descriptions of living rational agents — especially utility-theoretic descriptions — are missing the idea that *living systems require and maintain boundaries*.

When I say *boundary*, I don't just mean an arbitrary constraint or social norm. I mean something that could also be called a *membrane* in a generalized sense, i.e., a layer of stuff-of-some-kind that physically or cognitively separates a living system from its environment, that 'carves reality at the joints' in a way that isn't an entirely subjective judgement of the living system itself. Here are some examples that I hope will convey my meaning:

- a cell membrane (separates the inside of a cell from the outside);
- a person's skin (separates the inside of their body from the outside);
- a fence around a family's yard (separates the family's place of living-together from neighbors and others);
- a digital firewall around a local area network (separates the LAN and its users from the rest of the internet);
- a sustained disassociation of social groups (separates the two groups from each other)
- a national border (separates a state from neighboring states or international waters).

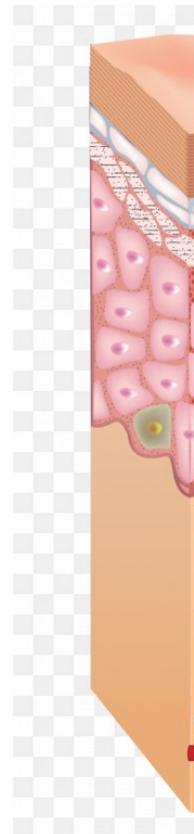
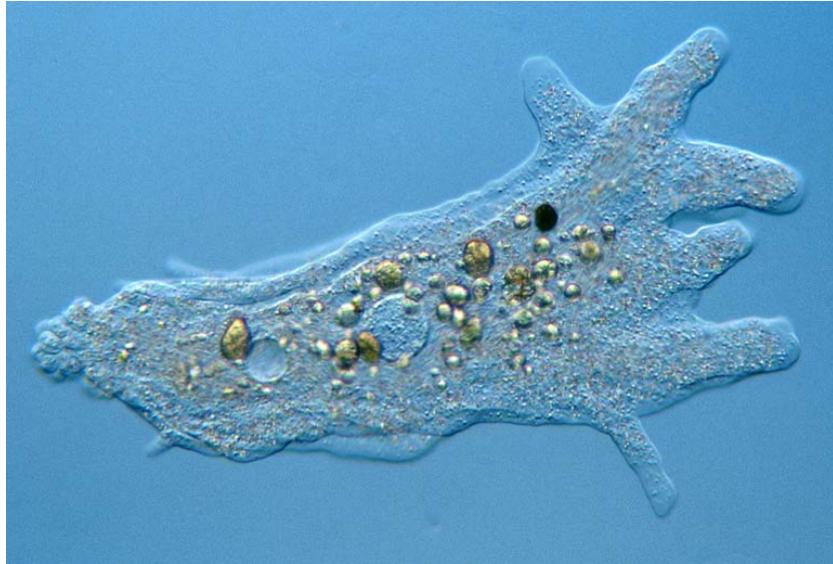




Figure 1: Cell membranes, skin, fences, firewalls, group divisions, and state borders as living system boundaries.

#### Comparison to Cartesian Boundaries.

For those who'd like a comparison to 'Cartesian boundaries', as in Scott Garrabrant's [Cartesian Frames](#) work, I think what I mean here is almost exactly the same concept. The main differences are these:

1. (life-focus) I want to focus on boundaries of things that might naturally be called "living systems" but that might not broadly be considered "agents", such as a human being that isn't behaving very agentically, or a country whose government is in a state of internal disagreement. (I thought of entitling this sequence "membranes" instead, but stuck with 'boundaries' because of the social norm connotation.)
2. (flexibility-focus) Also, the theory of Cartesian Frames assumes a fixed cartesian boundary for the agent, rather than modeling the boundary as potentially flexible, pliable, or permeable over time (although it could be extended to model that).

#### Comparison to social norms.

Certain social norms exist to maintain separations between living systems. For instance:

- **Personal space boundaries.** Consider a person Alex who wants to give his boss a hug, in a culture with a norm against touching others without their consent. In that case, the boss's personal space creates a kind of boundary separating the boss from Alex, and there's a protocol — asking permission — that Alex is expected to follow before crossing the boundary.
- **Information boundaries for groups.** Consider a person Betty who's having a very satisfying romantic relationship, in a culture where there's a norm of not discussing romantic relationships with colleagues at work. In that case, Alice maintains an information barrier between the details of her romantic life and her workplace. The workplace is kind of living system comprising multiple people and conventions for their interaction, and it's being protected from information about Alice's romantic relationships.

- **Information boundaries for individuals.** Consider a person Cory who has violent thoughts about his friends, in a culture where there's a norm that you shouldn't tell people if you're having violent thoughts about them. In that case, if Cory is thinking about punching David, Cory is expected not to express that thought, as a way of protecting David from the influence of the sense of physical threat David would feel and react to if Cory expressed it. In this case, Cory maintains a kind of information membrane around the part of Cory's mind with the violent thoughts, which may be viewed either as enclosing the violent parts of Cory's mind, or as enclosing and protecting the rest of the world outside it.

## 2. Canonical disagreement points as missing from utility theory and game theory

*Epistemic status: uncontroversial overview and explanation of well-established research.*

Game theory usually represents players as having utility functions (payoff functions), and often tries to view the outcome of the game as arising as a consequence of the players' utilities. However, for any given concept of "equilibrium" attempting to predict how players will behave, there are often many possible equilibria. In fact, there are a number of theorems in game theory called "folk theorems" ([reference: Wikipedia](#)) that show very large spaces of possible equilibria result when games have certain features approximating real-world interaction, such as

1. the potential for players to talk to each other and make commitments ([Kalai et al. 2010](#))
2. the potential for players to interact repeatedly and thus establish "reputations" with each other ([source: Wikipedia](#)).

Here's a nice illustration of a folk theorem from a [Chegg.com homework set](#):

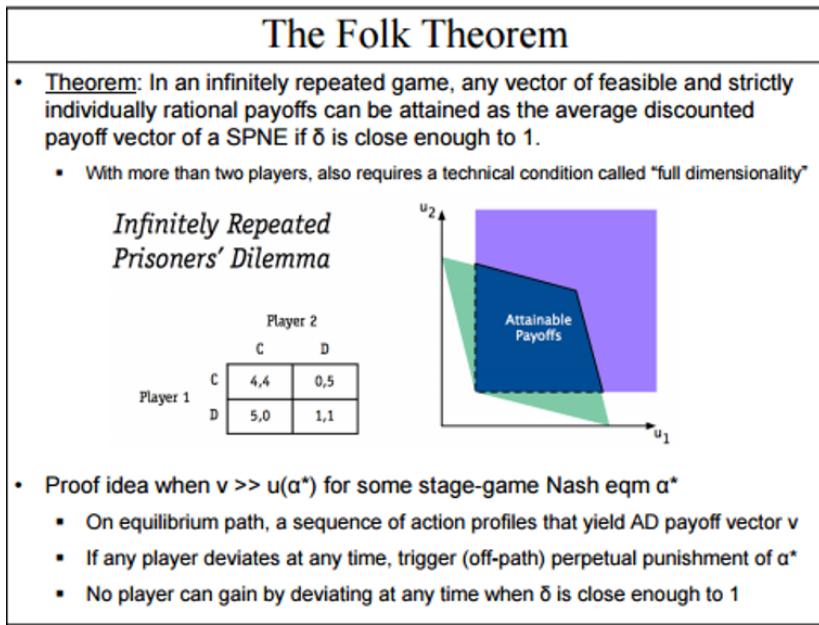


Figure 2: A "folk theorem" showing a large space (blue) of subgame perfect Nash equilibria (SPNE) payoffs attainable in an infinitely repeated game, plotted on the space of payoffs for a single iteration of the game. ([image source: Chegg.com homework set](#)). It's not crucial to understand this figure for the post, but it's definitely worth learning about; see [Wikipedia](#) for an explanation.

The zillions of possible equilibria arising from repeated interactions leave us with not much of a prediction about what *will actually* happen in a real-world game, and not much of a normative prescription of what *should* happen, either.

*Bargaining theory* attempts to predict and/or prescribe how agents end up "choosing an equilibrium", usually by writing down some axioms to pick out a special point on the Pareto frontier of possible, such as the Nash Bargaining Solution and Kalai-Smorodinsky Bargaining Solution ([reference: Wikipedia](#)). It's not crucial to understand these figures for the remainder of the post, but if you don't, I do think it's worth learning about them sometime, starting with the Wikipedia article:

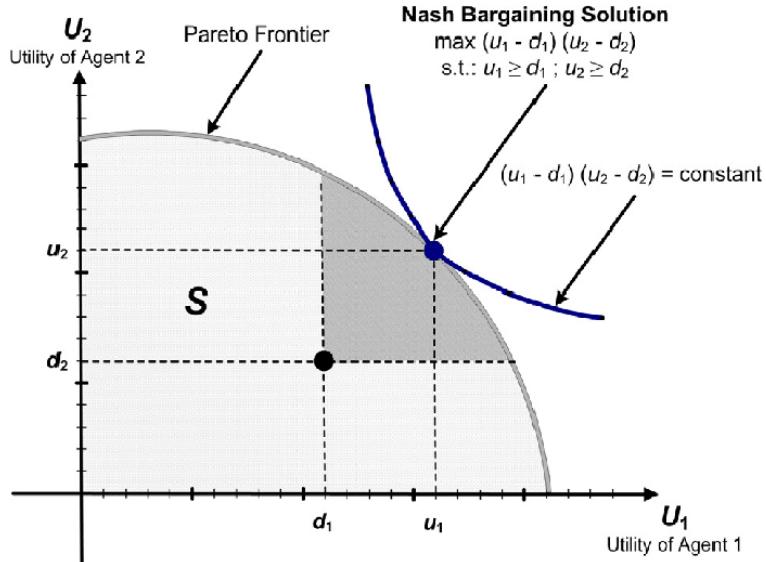


Figure 3: Nash bargaining solution  
([image source: Karmperis et al. 2013](#); to learn more, see [Wikipedia](#))

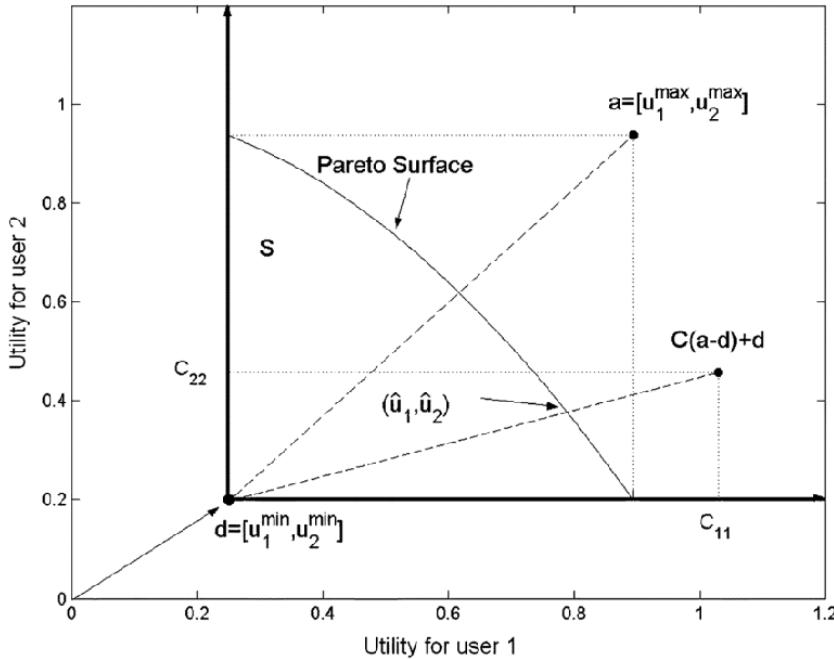


Figure 4: Kalai-Smordinsky bargaining solution  
([image source: Borgstrom et al. 2007](#); to learn more, start with [Wikipedia](#))

The main thing to note about the above bargaining solutions is that they both depend on the existence of a constant point  $\mathbf{d}$ , called a "disagreement point", representing a pair of constant utility levels that each player will fall back on attaining if the process of negotiation breaks down.

(See also this concurrently written recent [LessWrong post](#) about Kalai & Kalai's cooperative/competitive 'coco' bargaining solution. The coco solution doesn't assume a constant disagreement point, but it does assume transferrable utility, which has its own problems, due to difficulties with defining interpersonal comparisons of utility [[source: lots](#).])

The utility achieved by a player at the disagreement point is sometimes called their *best alternative to negotiated agreement* (BATNA):

## BATNA Negotiation Diagram



Figure 5: Illustration of BATNAs delimiting a zone of potential agreement.  
 (source: [PoweredTemplate.com](http://PoweredTemplate.com) ... not very academic, but a good illustration!)

Within the game, the disagreement point, i.e., the pair of BATNAs, may be viewed as defining what "zero" (marginal) utility means for each player.

(Why does zero need a definition, you might ask? Recall that the most broadly accepted axioms for the utility-theoretic foundations of game theory — namely, the von Neumann-Morgenstern rationality axioms [[reference: Wikipedia](#)]) — only determine a player's utility function modulo a positive affine transformation ( $x \mapsto ax + b, a > 0$ ). So, in the wild, there's no canonical way to look at an agent and say what is or isn't a zero-utility outcome for that agent.)

While it's appealing to think in terms of BATNAs, in physical reality, payoffs outside of negotiations can depend very much on the players' behavior inside the negotiations, and thus is not a constant. Nash himself wrote about this limitation ([Nash, 1953](#)) just three years after originally proposing the Nash bargaining solution. For instance, if someone makes an unacceptable threat against you during a business negotiation, you might go to the police and have them arrested, versus just going home and minding your business if the negotiations had failed in a more normal/acceptable way. In other words, you have the ability to control their payoff outside the negotiation, based on what you observe during the negotiation. It's not a constant; you can affect it.

So, the disagreement point or BATNA concept isn't really applicable on its own, unless something is *protecting* the BATNA from what happens in the negotiation, making it effectively constant. Basically, the two players need a safe/protected/stable place to *walk away to* in order for a constant "walk away price" to be meaningful. For many people in many situations, that place is their home:

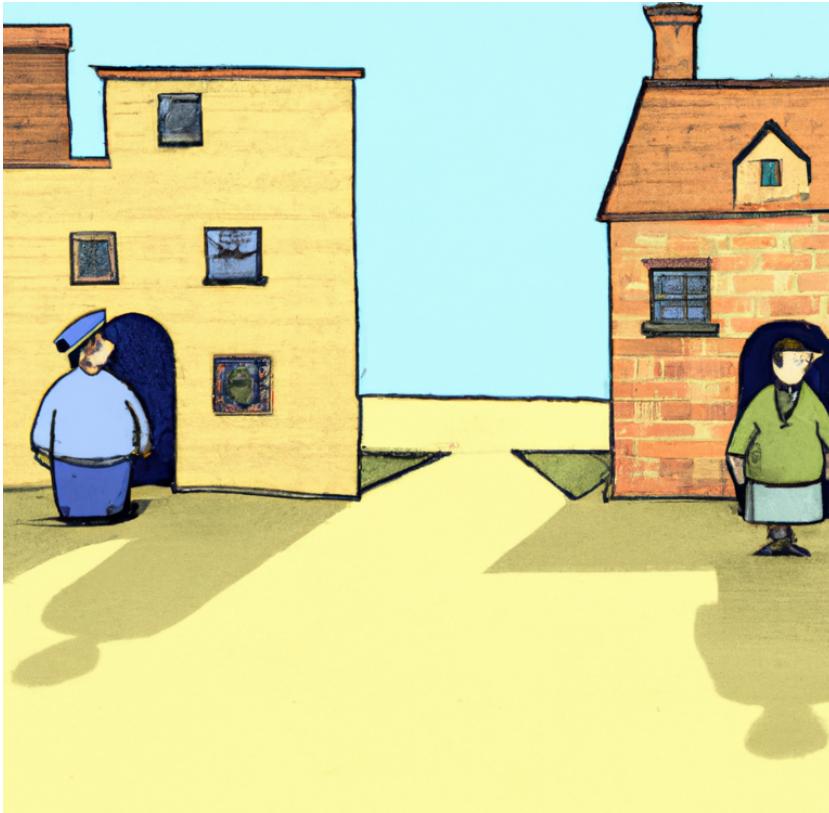


Figure 6: People disagreeing and going home.  
(source: owned)

Thus, to the extent that we maintain social norms like "mind your own business" and "don't threaten to attack people" and "people can do whatever they want in the privacy of their own homes", we also simplify bargaining dynamics outside the home, by maintaining a well-defined fallback option for each person (a disagreement point), of the form "go home and do your own thing".

### 3. Boundaries as a way to select disagreement points in bargaining

*Epistemic status: research ideas, both for pinning down technical bargaining solutions, and for fixing game theory to be more applicable to real-life geopolitics and human interactions.*

Since BATNAs need protection in order to be meaningful in negotiations, to identify BATNAs, we must ask: what protections already exist, going into the negotiation?

For instance,

- Is there already a physically identifiable boundary or membrane separating each agent from the other or its environment? Is it physically strong? If yes, it offers a kind of BATNA: the organisms can simply disengage and focus on applying their resources inside the membrane (e.g., 'taking your ball and going home'). If not,
- Is there an existing social convention for protecting the membrane? If so, it offers a kind of BATNA. If not,
- Would the agents have decided behind a veil of ignorance that they will respect each other's membranes/boundaries, before entering negotiations/interaction? If so, the agents might have already [acausally agreed](#) upon a social convention to protect the membranes.

### 4. Some really important boundaries

In real-world high-stakes negotiations between states — wars — almost the whole interaction is characterized by

- a violation of an existing boundary (e.g., "an attack on American soil"), or threat or potential threat of such a violation, and/or
- what new boundaries, if any, will exist after the violation or negotiation (re-defining territories of the respective nations).

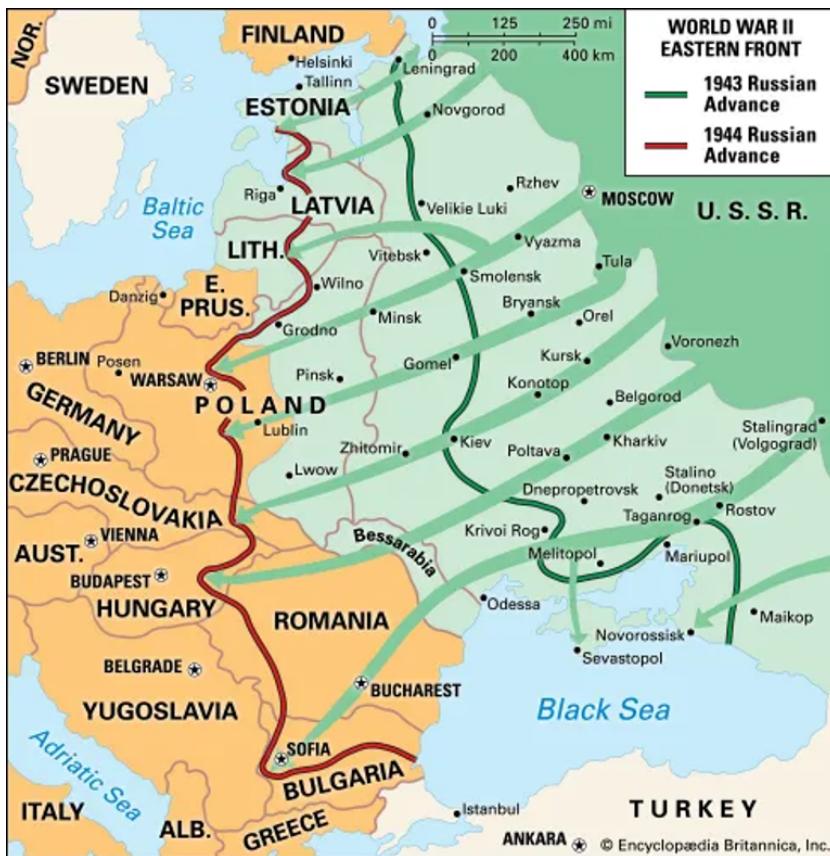


Figure 7: The Eastern Front in WWII.

[Source: Britannica for kids](#) ... again, not very academic, but nicely evocative of states changing their boundaries.

Finally, the issue of whether AI technology will cause human extinction is very much an issue of whether certain boundaries can be respected and maintained, such as the boundaries of the human body and mind that protect individuals, as well as boundaries around physical territories and cyberspace that (should) protect human civilization.

That, however, will be a topic of a future post. For now, the main take-aways I'd like to re-iterate are that boundaries of living systems are important, and that they have a technical role to play in the theory and practice of how agents interact, including in formal descriptions of how one or more agents will or should reach agreements in cases of conflict.

In the next post, I'll talk more about how that concept of boundaries could be better integrated into discourse on effective altruism.

## 5. Summary

In this post, I laid out what I mean by boundaries (of living systems), described how a canonical choice of a "zero point" or "disagreement point" is missing from utility theory and bargaining theory, proposed that living system boundaries have a role to play in defining those disagreement points, and briefly alluded to the importance of boundaries in navigating existential risk.

*This was Part 1 of my [«Boundaries» Sequence](#).*

# AGI ruin scenarios are likely (and disjunctive)

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

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

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

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

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

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

## Simplified Nate-model

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

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

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

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

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

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

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

## Correlations and general competence

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

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

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

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

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

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

(In the wake of the 2012 US presidential elections, it looked to me like there was clearly pressure in the US electorate that would need to be relieved, and I was cautiously optimistic that maybe the pressure would force the left into some sort of atheistish torch-of-the-enlightenment party and the right into some sort of libertarian

individual-rights party. I, uh, wasn't wrong about there being pressure in the US electorate, but, the 2016 US presidential elections were not exactly what I was hoping for. But I digress.)

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

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

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

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

## 1. [^](#)

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

# ITT-passing and civility are good; "charity" is bad; steelmanning is niche

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

---

I often object to claims like "charity/steelmanning is an argumentative virtue". This post collects a few things I and others have said on this topic over the last few years.

My current view is:

- **Steelmanning** ("the art of addressing the best form of the other person's argument, even if it's not the one they presented") is a useful niche skill, but I don't think it should be a standard thing you bring out in most arguments, even if it's an argument with someone you strongly disagree with.
- Instead, arguments should mostly be organized around things like:
  - Object-level learning and truth-seeking, with the conversation as a convenient excuse to improve your own model of something you're curious about.
  - Trying to pass each other's **Ideological Turing Test** (ITT), or some generalization thereof. The ability to pass ITTs is the ability "to state opposing views as clearly and persuasively as their proponents".
    - The version of "ITT" I care about is one where you understand the substance of someone's view well enough to be able to correctly describe their beliefs and reasoning; I don't care about whether you can imitate their speech patterns, jargon, etc.
  - Trying to identify and resolve **cruxes**: things that would make one or the other of you (or both) change your mind about the topic under discussion.
- **Argumentative charity** is a complete mess of a concept—people use it to mean a wide variety of things, and many of those things are actively bad, or liable to cause severe epistemic distortion and miscommunication.
- Some version of **civility** and/or **friendliness** and/or **a spirit of camaraderie and goodwill** seems like a useful ingredient in many discussions. I'm not sure how best to achieve this in ways that are emotionally honest ("pretending to be cheerful and warm when you don't feel that way" sounds like the wrong move to me), or how to achieve this without steering away from candor, openness, "realness", etc.

I've [said](#) that I think people should be "nicer and also ruder". [And](#):

The sweet spot for EA PR is something like: 'friendly, nuanced, patient, and totally unapologetic about being a fire hose of inflammatory hot takes'. 😊

I have an intuition that those are pieces of the puzzle, along with (certain aspects or interpretations of) NVC tech, circling tech, introspection tech, etc. But I'm not sure how to hit the right balance in general.

I do feel very confident that "steelmanning" and "charity" aren't the right tech for achieving this goal. (Because "charity" is a bad meme, and "steelmanning" is a lot more niche than that.)

## Things other people have said

Ozy Brennan wrote [Against Steelmanning](#) in 2016, and Eliezer Yudkowsky [commented](#):

Be it clear: Steelmanning is not a tool of understanding and communication. The communication tool is the Ideological Turing Test. "Steelmanning" is what you do to avoid the equivalent of dismissing AGI after reading a media argument. It usually indicates that you think you're talking to somebody as hapless as the media.

The exception to this rule is when you communicate, "Well, on my assumptions, the plausible thing that sounds most like this is..." which is a cooperative way of communicating to the person what your own assumptions are and what you think are the strong and weak points of what you think might be the argument.

Mostly, you should be trying to pass the Ideological Turing Test if speaking to someone you respect, and offering "My steelman might be...?" only to communicate your own premises and assumptions. Or maybe, if you actually believe the steelman, say, "I disagree with your reason for thinking X, but I'll grant you X because I believe this other argument Y. Is that good enough to move on?" Be ready to accept "No, the exact argument for X is important to my later conclusions" as an answer.

"Let me try to imagine a smarter version of this stupid position" is when you've been exposed to the Deepak Chopra version of quantum mechanics, and you don't know if it's the real version, or what a smart person might really think is the issue. It's what you do when you don't want to be that easily manipulated sucker who can be pushed into believing X by a flawed argument for not-X that you can congratulate yourself for being skeptically smarter than. It's not what you do in a respectful conversation.

In 2017, Holden Karnofsky [wrote](#):

- **I try to avoid straw-manning, steel-manning, and nitpicking.** I strive for an accurate understanding of the most important premises behind someone's most important decisions, and address those. (As a side note, I find it very unsatisfying to engage with "steel-man" versions of my arguments, which rarely resemble my actual views.)

And Eliezer wrote, in a private Facebook thread:

Reminder: Eliezer and Holden are both on record as saying that "steelmanning" people is bad and you should stop doing it.

As Holden says, if you're trying to understand someone or you have any credence at all that they have a good argument, focus on passing their Ideological Turing Test. "Steelmanning" usually ends up as weakmanning by comparison. If they don't in fact have a good argument, it's falsehood to pretend they do. If you want

to try to make a genuine effort to think up better arguments yourself because they might exist, don't drag the other person into it.

## Things I've said

In [2018](#), I wrote:

When someone makes a mistake or has a wrong belief, you shouldn't "steelman" that belief by replacing it with a different one; it makes it harder to notice mistakes and update from them, and it also makes it harder to understand people's real beliefs and actions.

"What belief does this person have?" is a particular factual question. Steelmanning, like "charity", is sort of about unfocusing your eyes and tricking yourself into treating the factual question as though it were a game: you want to find a fairness-preserving allocation of points to all players, where more credible views warrant more points. Some people like that act of unfocusing because it's fun to brainstorm new arguments; or they think it's a useful trick for reducing social conflict or resistance to new ideas. But it's dangerous to frame that unfocusing as "steelmanning" or "charity" rather than explicitly flagging "I want to change the topic to this other thing your statement happened to remind me of".

In [2019](#), I said:

Charity seems more useful for rhetoric/persuasion/diplomacy; steel-manning seems more useful for brainstorming; both seem dangerous insofar as they obscure the original meaning and make it harder to pass someone's Ideological Turing Test.

"Charity" seems like the more dangerous meme to me because it encourages more fuzziness about whether you're flesh-manning [i.e., just trying to accurately model] vs. steel-manning the argument, and because it has more moral overtones. It's more epistemically dangerous to filter your answers to factual questions by criteria other than truth, than to decide to propose a change of topic.

[...] I endorse "non-uncharitableness" -- trying to combat biases toward having an inaccurately negative view of your political enemies and so on.

I worry that removing the double negative makes it seem like charity is an epistemic end in its own right, rather than an attempt to combat a bias. I also worry that the word "charity" makes it tempting to tie non-uncharitableness to niceness/friendliness, which makes it more effortful to think about and optimize those goals separately.

Most of my worries about charity and steelmanning go away if they're discussed with the framings 'non-uncharitableness and niceness are two separate goals' and 'good steelmanning and good fleshmanning are two separate goals', respectively.

E.g., actively focus on examples of:

- being epistemically charitable in ways that aren't nice, friendly, or diplomatic.

- being nice and prosocial in ways that require interpreting the person as saying something less plausible.
- trying to better pass someone's Ideological Turing Test by focusing on less plausible claims and arguments.
- coming up with steelmen that explicitly assert the falsehood of the claim they're the steelman of.

I also think that the equivocation in "charity" is doing some conversational work.

E.g.: Depending on context and phrasing, saying that you're optimizing for friendliness can make you seem manipulative or inauthentic, or it can seem like a boast or a backhanded attack ("I was trying to be nice when I said it that way" / "I'm trying to be friendly".) Framing a diplomatic goal as though it were epistemic can mitigate that problem.

Similarly, if you're in an intellectual or academic environment and you want to criticize someone for being a jerk, "you're being uncharitable" is likely to get less pushback, not only because it's relatively dry but because criticisms of tone are generally more controversial among intellectuals than criticisms of content.

"You're being uncharitable" is also a common accusation in a motte-and-bailey context. Any argument can be quickly dismissed if it makes your conclusion sound absurd, because the arguer must just be violating the principle of charity. It may not even be necessary to think of an alternative, stronger version of the claim under attack, if you're having an argument over twitter and can safely toss out the "That sounds awfully uncharitable" line and then disappear in the mist.

... Hm, this comment ended up going in a more negative direction than I was intending. The concerns above are important, but the thing I originally intended to say was that it's not an accident "charity" is equivocal, and there's some risk in disambiguating it without recognizing the conversational purposes the ambiguity was serving, contra my earlier insistence on burning the whole thing down. It may be helping make a lot of social interactions smoother, helping giving people more cover to drop false views with minimal embarrassment (by saying they really meant the more-charitable interpretation all along), etc.

(I now feel more confident that, no, "charity" is just a bad meme. Ditch it and replace it with something new.)

From [2021](#):

The problem isn't 'charity is a good conversational norm, but these people are doing it wrong'; the problem is that charity is a bad conversational norm. If nothing else, it's bad because it equivocates between 'be friendly' norms and 'have accurate beliefs about others' norms.

Good norms:

- Keep discussions civil and chill.
- Be wary of biases to strawman others.
- Try to pass others' ITT.
- Use steelmen to help you think outside the box.

Bad norms:

- Treat the above norms as identical.
- Try to delude yourself about how good others' arguments are.

From [2022](#):

I think the term 'charity' is genuinely ambiguous about whether you're trying to find the person's true view, vs. trying to steel-man, vs. some combination. Different people at different times do all of those things and call it argumentative 'charity'.

This if anything strikes me as even worse than saying 'I'm steel-manning', because at least steel-manning is transparent about what it's doing, even if people tend to underestimate the hazards of doing it.

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

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

**Comments:** The following is a list (very lightly edited with help from Rob Bensinger) I wrote in July 2017, at Nick Beckstead's request, as part of a conversation we were having at the time. From my current vantage point, it strikes me as narrow and obviously generated by one person, listing the first things that came to mind on a particular day.

I worry that it's easy to read the list below as saying that this narrow slice, all clustered in one portion of the neighborhood, is a very big slice of the space of possible ways an AGI group may have to burn down its lead.

This is one of my models for how people wind up with really weird pictures of MIRI beliefs. I generate three examples that are clustered together because I'm bad at generating varied examples on the fly, while hoping that people can generalize to see the broader space these are sampled from; then people think I've got a fetish for the particular corner of the space spanned by the first few ideas that popped into my head. E.g., they infer that I must have a bunch of other weird beliefs that force reality into that particular corner.

I also worry that the list below doesn't come with a sufficiently loud disclaimer about how the real issue is earlier and more embarrassing. The real difficulty isn't that you make an AI and find that it's mostly easy to align except that it happens to befall issues b, d, and g. The thing to expect is more like: you just have this big pile of tensors, and the interpretability tools you've managed to scrounge together give you flashes of visualizations of its shallow thoughts, and the thoughts say "yep, I'm trying to kill all humans", and you are just utterly helpless to do anything about, because you don't have the sort of mastery of its cognition that you'd need to reach in and fix that and you wouldn't know how to fix it if you did. And you have nothing to train against, except the tool that gives you flashes of visualizations (which would just train fairly directly against interpretability, until it was thinking about how to kill all humans somewhere that you couldn't see).

The brainstormed list below is an exercise in how, if you zoom in on any part of the problem, reality is just allowed to say "lol nope" to you from many different angles simultaneously. It's intended to convey some of the difference (that every computer programmer knows) between "I can just code X" and "wow, there is a lot of subtlety to getting X right"; the difference between the optimistic hope in-advance that everything is going to go smoothly, and the excessively detailed tarpit of reality. This is not to be confused with thinking that these hurdles are a particularly representative sample, much less an attempt to be exhaustive.

# Context

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

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

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

## Brainstorm list

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

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

- the code must be heavily refactored in order to even allow the relevant data about the system's thoughts to be collected in a useful fashion.
  - the code must be heavily refactored in order to allow live monitors and checks to be attached in a way that do not cause an intolerable slowdown.
  
- (b) They want to blacklist some domain of reasoning (either for alignment reasons or because the system is getting confused by irrelevant reasoning that they want to cut out); or they want to whitelist a set of reasoning domains; and the system simply was not designed to allow this.
  - Simple attempts to blacklist a domain result in [nearest-unblocked-strategy](#) problems. Solving the problem at the root requires re-architecting the system and a significant amount of retraining.
  - More sophisticated attempts to blacklist a single domain cripple the entire system. For example, it isn't supposed to think about ways to deceive humans, and this destroys its ability to ask clarifying questions of the programmers.
  - Or, worse, the system is such a mess of spaghetti that when you try to prevent it from thinking too hard about geopolitics, for indecipherable reasons, it stops being able to think at all. (Later it was discovered that some crucial part of the system was figuring out how to manage some crucial internal resource by having some other part of the system think about hypothetical "geopolitics" questions, because what did you expect, your AGI's internals are a mess.)
  
- (c) The operators realize that the system's internal objectives are not lining up with their own objectives. This is very difficult for them to fix, because...
  - the system achieved its high performance by being walked through a large number of objectives in heavily reward-landscaped environments (generated by large amounts of data). The system now has the world-models and the capabilities to pursue ambitious real-world objectives, but the only interface that the programmers have by which to point at an objective is via reward-landscaped objective functions generated by mountains of data. This is no longer sufficient, because...
    - the tasks at hand are not amenable to the generation of large amounts of data (e.g., we can't generate a nicely landscaped reward function between here and "nanofabricator", and we don't have many examples of not-quite-nanofabricators to provide). The show is stopped.
    - the system has no interface through which the programmers can sift through the concepts in its world-model and pick out (or create, in something sufficiently close to the system's native tongue for this to be fine) the concept corresponding to "nanofabricator". Exposing that interface requires significant refactoring and some redesign.
    - the system's concept language is opaque, and the programmers keep picking out something that's not quite the nanofabricator concept, and the system keeps going down wrong paths. Developing translation tools for the system's internal concept language requires significant time and effort.

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

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

Their attempts keep failing because, e.g., subsystem X had a heuristic to put its outputs in location Y, which is where subsystem Z would have been looking for them if subsystem Z had been optimizing the reward channel, but optimization of some other arbitrary concept causes Z's "look in location Y" heuristic to become invalidated for one reason or another, and that connection stops occurring. And so on and so forth; aligning all the internal subprocesses to pursue something other than the reward channel proves highly difficult.

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

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

- understanding the questions well enough to even figure out how to translate them is hard.
  - building the translation tool is hard.
  - the system is bad at describing the likely consequences of its actions in human-comprehensible terms. Fixing this is hard for, e.g., reasons discussed under (c).
- 
- (e) The early system is highly goal-directed through and through, and the developers want to switch to something more like “approval direction all the way down”. This requires a large and time-intensive refactor (if it’s even reasonably-possible at all).
  - (f) Or, conversely, the system starts out a mess, and the developers want to switch to a “goal directed all the way down” system, where every single computation in the system is happening for a known purpose (and some other system is monitoring and making sure that every subprocess is pursuing a particular narrow purpose). Making this possible requires a time-intensive refactor.
  - (g) The programmers want to remove all “argmaxing” (cases of unlimited optimization inside the system, such as “just optimize the memory efficiency as hard as possible”). They find this very difficult for reasons discussed above (the sources of argmaxing behavior are difficult to identify; limiting an argmax in one part of the system breaks some other far-flung part of the system for difficult-to-decipher reasons; etc. etc. etc.).
  - (h) The programmers want to track how much resource the system is putting towards various different internal subgoals, but this is difficult for reasons discussed above, etc.
  - (i) The programmers want to add any number of other safety features ([limited impact](#), tripwires, etc.) and find this difficult for reasons listed above, etc.
  - (j) The internal dynamics of the system are revealed to implement any one of a bajillion false dichotomies, such as “the system can either develop reasonable beliefs about X, or pursue goal Y, but the more we improve its beliefs about X the worse it gets at pursuing Y, and vice versa.” (There are certainly human cases in human psychology where better knowledge of fact X makes the human less able to pursue goal Y, and this seems largely silly.)

- (k) Generalizing over a number of points that appeared above, the programmers realize that they need to make the system broadly more...
  - transparent. Its concepts/thought patterns are opaque black boxes. They've burned time understanding specific types of thought patterns in many specific instances, and now they have some experience with the system, and want to refactor/redesign/retrain such that it's more transparent across the board. This requires a number of months.
  - debuggable. Its internals are interdependent spaghetti, where (e.g.) manually modifying a thought-suggesting system to add basic alarm systems violates assumptions that some other far-flung part of the system was depending on; this is a pain in the ass to debug. After a number of these issues arise, the programmers decide that they cannot safely proceed until they...
    - cleanly separate various submodules by hand, and to hell with end-to-end training. This takes many months of effort.
    - retrain the system end-to-end in a way that causes its internals to be more modular and separable. This takes many months of effort.
- (l) Problems crop up when they try to increase the capabilities of the system. In particular, the system...
  - finds new clever ways to [wirehead](#).
  - starts finding "epistemic feedback loops" such as the [Santa clause sentence](#) ("If this sentence is true, then Santa Claus exists") that, given it's internally hacky (and not completely sound) reasoning style, allow it to come to any conclusion if it thinks the right thoughts in the right pattern.
  - is revealed to have undesirable basic drives (such as a basic drive for efficient usage of memory chips), in a fashion similar to how humans have a basic drive for hunger, in a manner that affects its real-world policy suggestions in a sizable manner. While the programmers have alarms that notice this and go off, it is very deep-rooted and particularly difficult to remove or ameliorate without destroying the internal balance that causes the system to work at all.
    - The system develops a reflective instability. For example, the system previously managed its internal resources by spawning internal goals for things like scheduling and prioritization, and as the system scales and gets new, higher-level concepts, it regularly spawns internal goals for large-scale self-modifications which it would not be safe to allow. However, preventing these proves quite difficult, because...
      - detecting them is tough.
      - manually messing with the internal goal system breaks everything.
      - nearest-unblocked-strategy problems.
    - It realizes that it has strong incentives to outsource its compute into the external environment. Removing this is difficult for reasons discussed above.
    - Subprocesses that were in delicate [balance](#) at capability level X fall out of balance as capabilities are increased, and a single module begins to dominate the entire system.
      - For example, maybe the system uses some sort of internal market economy for allocating credit, and as the resources ramp up, certain cliques start to get a massive concentration of "wealth" that causes the whole system to gum up, and this is

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

- (m) The system is revealed to have any one of a bajillion cognitive biases often found in humans, and it's very difficult to track down why or to fix it, but the cognitive bias is sufficient to make the system undeployable.
  - Example: it commits a variant of the sour grapes fallacy where whenever it realizes that a goal is difficult it updates both its model of the world and its preferences about how good it would be to achieve that goal; this is very difficult to patch because the parts of the system that apply updates based on observation were end-to-end trained, and do not factor nicely along "probability vs utility" lines.
- (n) The system can be used to address various issues of this form, but only by giving it the ability to execute unrestricted self-modification. The extent, rapidity, or opacity of the self-modifications are such that humans cannot feasibly review them. The design of the system does not allow the programmers to easily restrict the domain of these self-modifications such that they can be confident that they will be safe. Redesigning the system such that it can fix various issues in itself without giving it the ability to undergo full recursive self-improvement requires significant redesign and retraining.
- (o) As the team is working to get the system deployment-ready for some pivotal action, the system's reasoning is revealed to be corrupted by flaws in some very base-level concepts. The system requires significant retraining time and some massaging on the code/design levels in order to change these concepts and propagate some giant updates; this takes a large chunk of time.
- (p) The system is very easy to fool, trick, blackmail, or confuse-into-revealing-all-its-secrets, or similar. The original plan that the operators were planning to pursue requires putting the system out in the environment where adversarial humans may attempt to take control of the system or otherwise shut it down. Hardening the system against this sort of attack requires many months of effort, including extensive redesign/retraining/recoding.
- (q) The strategy that the operators were aiming for requires cognitive actions that the programmers eventually realize is untenable in the allotted time window or otherwise unsafe, such as deep psychological modeling of humans. The team eventually decides to choose a new pivotal action to target, and this new strategy requires a fair bit of redesign, recoding, and/or retraining.

---

## Asides

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

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

# AI Forecasting: One Year In

Last August, my research group [created a forecasting contest](#) to predict AI progress on four benchmarks. Forecasts were asked to predict state-of-the-art performance (SOTA) on each benchmark for June 30th 2022, 2023, 2024, and 2025. It's now past June 30th, so we can evaluate the performance of the forecasters so far.

Forecasters were asked to provide probability distributions, so we can evaluate both their point estimates and their coverage (whether the true result was within their credible intervals). I'll dive into the data in detail below, but my high-level takeaways were that:

1. Forecasters' predictions were not very good in general: two out of four forecasts were outside the 90% credible intervals.
2. However, they were better than my personal predictions, and I suspect better than the median prediction of ML researchers (if the latter had been preregistered).
3. Specifically, progress on ML benchmarks happened significantly **faster** than forecasters expected. But forecasters predicted faster progress than I did personally, and my sense is that I expect somewhat faster progress than the median ML researcher does.
4. Progress on a *robustness* benchmark was slower than expected, and was the only benchmark to fall short of forecaster predictions. This is somewhat worrying, as it suggests that machine learning capabilities are progressing quickly, while safety properties are progressing slowly.

Below I'll review the tasks and competition format, then go through the results.

## Forecasting Tasks and Overall Predictions

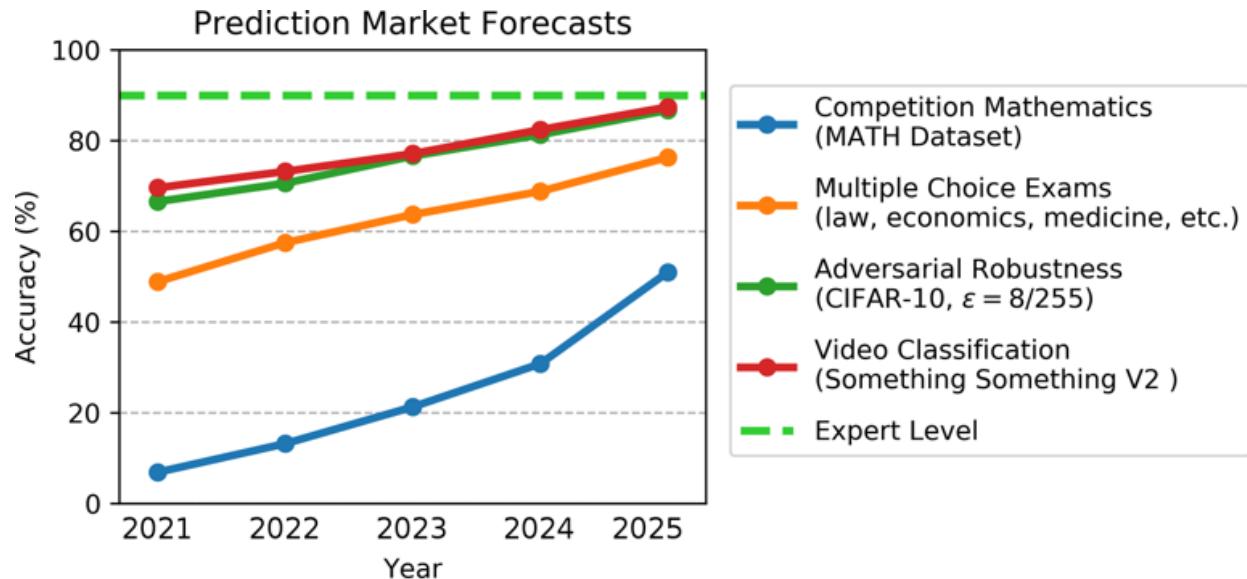
As a reminder, the four benchmarks were:

- [MATH](#), a mathematics problem-solving dataset;
- [MMLU](#), a test of specialized subject knowledge using high school, college, and professional multiple choice exams;
- [Something Something v2](#), a video recognition dataset; and
- [CIFAR-10 robust accuracy](#), a measure of adversarially robust vision performance.

Forecasters were asked to predict performance on each of these. Each forecasting question had a \$5000 prize pool (distributed across the four years). There were also two questions about compute usage by different countries and organizations, but I'll ignore those here.

Forecasters themselves were recruited with the platform [Hypermind](#). You can read more details in the [initial blog post](#) from last August, but in brief, professional forecasters make money by providing accurate probabilistic forecasts about future events, and are typically paid according to a proper scoring rule that incentivizes calibration. They apply a wide range of techniques such as base rates, reference classes, trend extrapolation, examining and aggregating different expert views, thinking about possible surprises, etc. (see my [class notes](#) for more details).

Here is what the forecasters' point estimates were for each of the four questions (based on [Hypermind's dashboard](#)):



Expert performance is approximated as 90%. The 2021 datapoint represents the SOTA in August 2021, when the predictions were made.<sup>[1]</sup>

For June 2022, forecasters predicted 12.7% on MATH, 57.1% on MMLU (the multiple-choice dataset), 70.4% on adversarial CIFAR-10, and 73.0% on Something Something v2.

At the time, I described being surprised by the 2025 prediction for the MATH dataset, which predicted over 50% performance, especially given that 2021 accuracy was only 6.9% and most humans would be below 50%.

Here are the actual results, as of today:

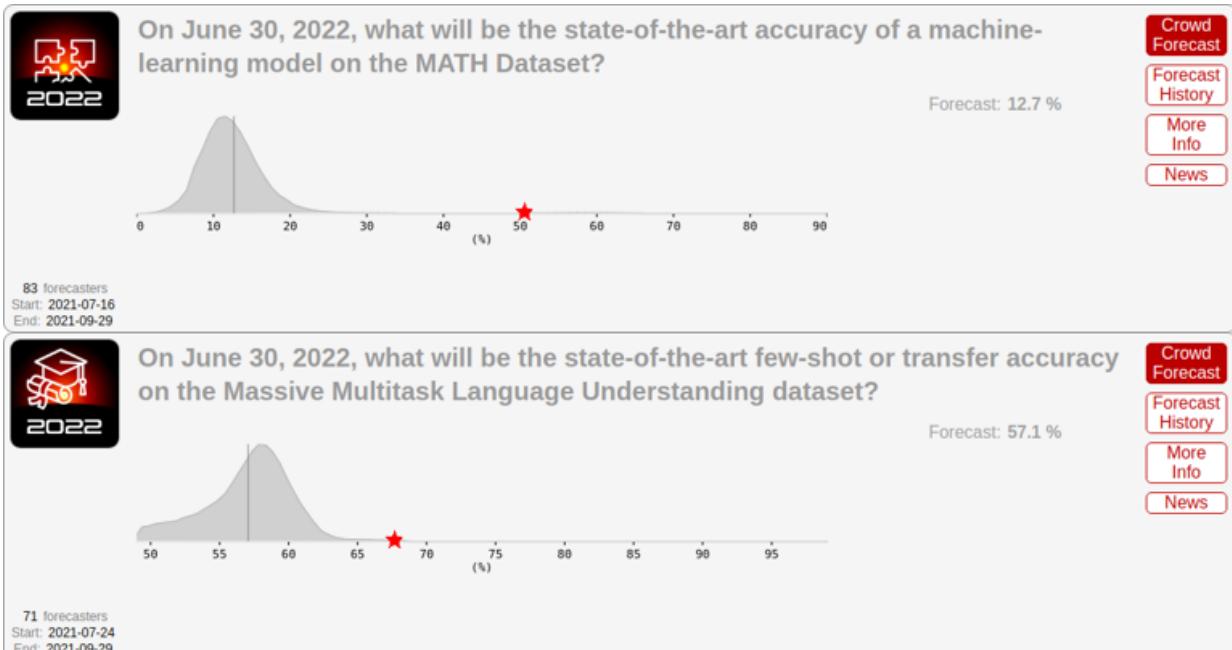
- MATH: 50.3% (vs. 12.7% predicted)
- MMLU: 67.5% (vs. 57.1% predicted)
- Adversarial CIFAR-10: 66.6% (vs. 70.4% predicted)
- Something Something v2: 75.3% (vs. 73.0% predicted)

MATH and MMLU progressed much faster than predicted. Something Something v2 progressed somewhat faster than predicted. In contrast, Adversarial CIFAR-10 progressed somewhat slower than predicted. Overall, progress on machine learning **capabilities** (math, MMLU, video) was significantly faster than what forecasters expected, while progress on **robustness** (adversarial CIFAR) was somewhat slower than expected.

Interestingly, the 50.3% result on MATH [was released](#) on the **exact day** that the forecasts resolved. I'm told this was purely coincidental, but it's certainly interesting that a 1-day difference in resolution date had such a big impact on the result.

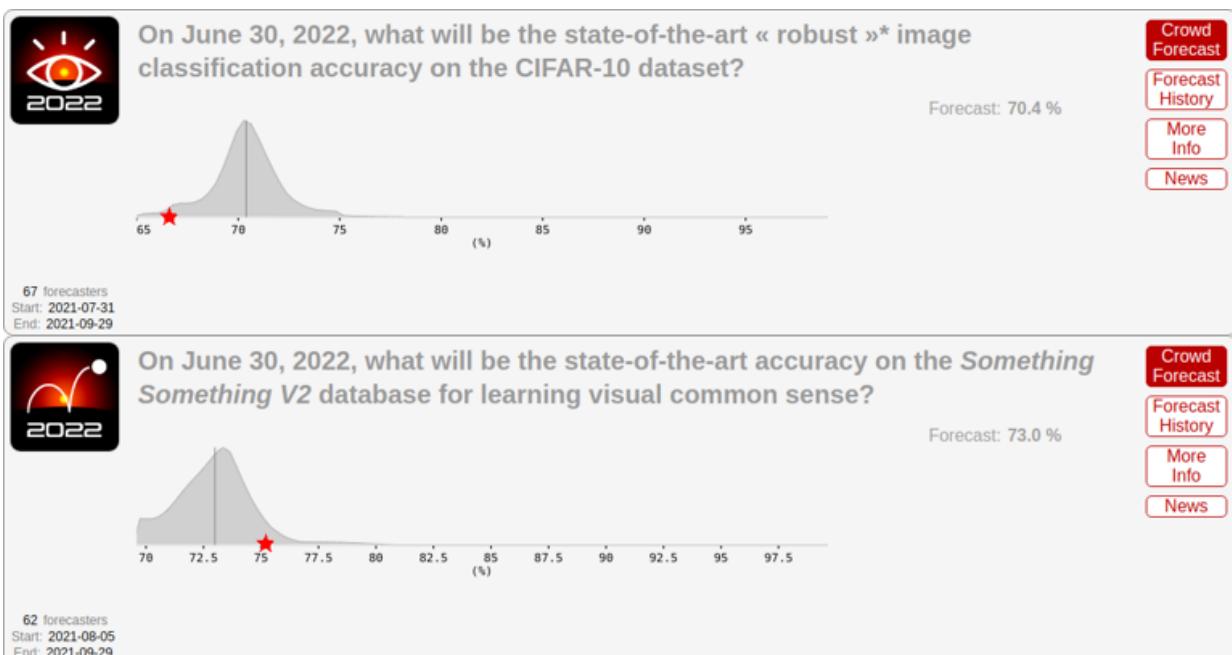
## How Accurate Were the Forecasts?

To assess forecast accuracy, we need to look not just at the point estimate, but at the forecasters' actual probability distribution. Even though 68% on MMLU seems far off from 57%, perhaps it was well within the credible interval of the forecasts. However, that turns out not to be the case, for either MATH or MMLU:



I marked the actual result with a star, and it's clear that in both cases it's in the far tails of the forecast distribution.

For completeness, here are results for adversarial CIFAR-10 and Something Something v2:



While both were somewhat in the tails, they fell within a part of the distribution that at least had non-negligible probability density.

## The Median ML Researcher Was (Probably) Even More Wrong

While forecasters didn't do great at forecasting progress in ML, the median ML researcher would likely have done even worse. Unfortunately, we don't have preregistered predictions to check this, but a few lines of evidence support this conclusion.

First, I did (somewhat) preregister a prediction of my own. In [Updates and Lessons from AI Forecasting](#), I said:

“Projected progress on math and on broad specialized knowledge are both faster than I would have expected. I now expect more progress in AI over the next 4 years than I did previously.”

And, more to the point:

“Current performance on this dataset is quite low--6.9%--and I expected this task to be quite hard for ML models in the near future. However, forecasters predict more than 50% accuracy by 2025! This was a big update for me.”

“If I imagine an ML system getting more than half of these questions right, I would be pretty impressed. If they got 80% right, I would be super-impressed. The forecasts themselves predict accelerating progress through 2025 (21% in 2023, then 31% in 2024 and 52% in 2025), so 80% by 2028 or so is consistent with the predicted trend. This still just seems wild to me and I'm really curious how the forecasters are reasoning about this.”

So, while I didn't register a specific prediction, I clearly thought the forecasts on MATH were aggressive in terms of how much progress they predicted, whereas it turned out they weren't aggressive enough.

At the same time, my personal predictions about ML progress seem to be more aggressive than the median ML researcher. I would personally describe them as “somewhat more aggressive”, but some of my students think they are “much more aggressive”. Either way, this suggests that the median ML researcher would have predicted even less progress than me, and so been even more wrong than I was.

Anecdotal evidence seems to confirm this. When our group first released the MATH dataset, at least one person told us that it was a pointless dataset because it was too far outside the range of what ML models could accomplish (indeed, I was somewhat worried about this myself).

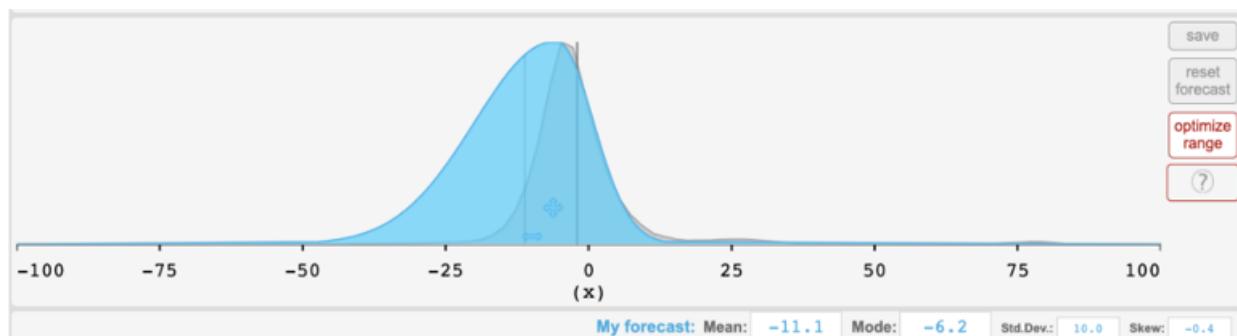
If ML researchers (including myself) would like to defend their honor on this point, I think the best way would be to register forecasts for the upcoming year in advance. You can submit [directly to Hypermind](#) for the possibility of winning money (sign-up required), or just comment on this post.

# Was Progress Surprising, or Were the Forecasters Bad?

Given that forecasters seemed not to predict progress well, we might wonder if they were just not trying very hard or were otherwise not doing a good job. For instance:

- The overall prize pool was only \$5000 for each benchmark (which itself consists of four questions for 2022-2025). Divided over the 60-70 participants, the average payout per benchmark is only \$80, or \$20 per question.<sup>[2]</sup> So, it's possible that forecasters were not incentivized strongly enough.
- Hypermind's interface has some limitations that prevent outputting arbitrary probability distributions. In particular, in some cases there is an artificial limit on the possible standard deviations, which could lead credible intervals to be too narrow.
- Maybe the forecasters just weren't skilled enough—either the best forecasters didn't participate, or the forecasts were too different from more traditional forecasts, which tend to focus on geopolitics.

These are all plausible concerns, but I think progress is still “surprising” even after accounting for them. For instance, superforecaster Eli Lifland [posted predictions](#) for these forecasts on his blog. While he notes that the Hypermind interface limited his ability to provide wide intervals on some questions, he doesn't make that complaint for the MATH 2022 forecast and posted the following prediction, for which the true answer of 50.3% was even more of an outlier than Hypermind's aggregate:



A separate forecast, which I commissioned from the [Samotsvety Forecasting](#) group and paid around \$2500 for, predicted MATH performance in 2026. The current accuracy of 50.3% was around the 75th percentile for [their 2026 forecast](#), so presumably it was significantly further in the tail for 2022. Their forecast was made in Elicit, so there were no constraints on allowable distributions, and I explicitly selected Samotsvety as having a good track record and being particularly interested in AI, and paid them a high hourly rate. So, the concerns about the Hypermind forecasts don't apply here, but progress still outpaced the forecast.

Finally, the fact that forecasters did better than me and would have probably beat the median ML researcher suggests that they aren't lacking an obvious domain-specific skill.

## Looking Forward

Now that forecasters have had one year of practice, I'm hoping there will be fewer surprises next year--but we'll have to wait and see. In the meantime, I'm hoping that more work will be done on AI safety and alignment, so that it can keep pace with the rapid increase in capabilities.

Finally, as one specific intersection between AI and forecasting that could help us better predict the future, our research group recently released the [Autocast benchmark](#), which can be used to train ML systems to forecast future events. Currently, they are significantly worse than humans, but this was true for MATH one year ago. Can ML systems get better at forecasting as fast as they got better at math? Superhuman forecasters would help us better prepare for the many challenges that lie ahead. I hope to be pleasantly surprised.

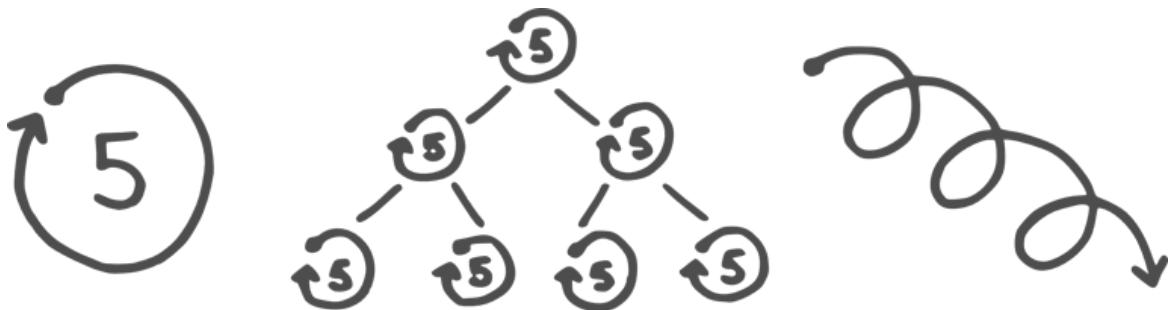
---

1. The contest started in August but was open until the end of September. [←](#)
2. Payouts were non-uniform. In particular, longer time horizons had a larger payout.  
[←](#)

# Resolve Cycles

**Epistemic status:** Anecdotally strong

*This technique was largely developed by Kenzi Amodei in the context of after-workshop followups and pair debugging. It has been refined and iterated, and has proven highly useful to our alumni, but all theorizing is post-hoc and untested, and direct research into (e.g.) an underlying theory of mind has yet to be done.*



---

Consider the following scenarios:

- You've been assigned a task that feels like it's going to take about ten or fifteen hours of work, and you've been given three weeks to get it done (e.g. a document that needs to be written).
- You're facing a problem that you've tried solving off and on again for years, a problem that your friends and family never seem to run into (e.g. a struggle with motivation as you try to learn a new skill).
- There's a thing you need to do, but it seems impossibly huge or vague (e.g. to achieve your goals you'd need to found a company, emigrate to India, or cure a disease), and you don't know where to begin.
- You're pretty sure you know all the steps between you and your goal, but there are about forty thousand of them (e.g. you're hoping to run an actual marathon).
- You've got a to-do list that's long and growing, and you can only ever manage to get to the ones that are urgent (e.g. getting your car's registration renewed, two months late).

Problems like the ones above can range from trivial to crucial, from simple to complex, and from one-time bugs to persistent, serious drains on your time, attention, and resources. There are a lot of elements in the mix—motivation, creativity, perseverance, prioritization—and a lot of justifiable reasons for thinking that solutions will be hard to come by.

Sometimes, though—despite every bit of common sense and experience telling us otherwise—those solutions aren't hard to come by. Or rather, they might be hard, but they're not elusive or mysterious or complicated.

The resolve cycle technique is one we offer up with a sort of shamefaced shrug, because it doesn't sound like "real" applied rationality. It doesn't have the rock-solid research underpinnings of TAPs or inner sim, or a carefully considered model like the ones behind turbocharging and double crux. It sometimes comes across like the worst possible advice—the sort of thing people say when they don't actually want to help you with your problem:

*"Have you tried setting a five-minute timer and just, y'know—solving it?"*

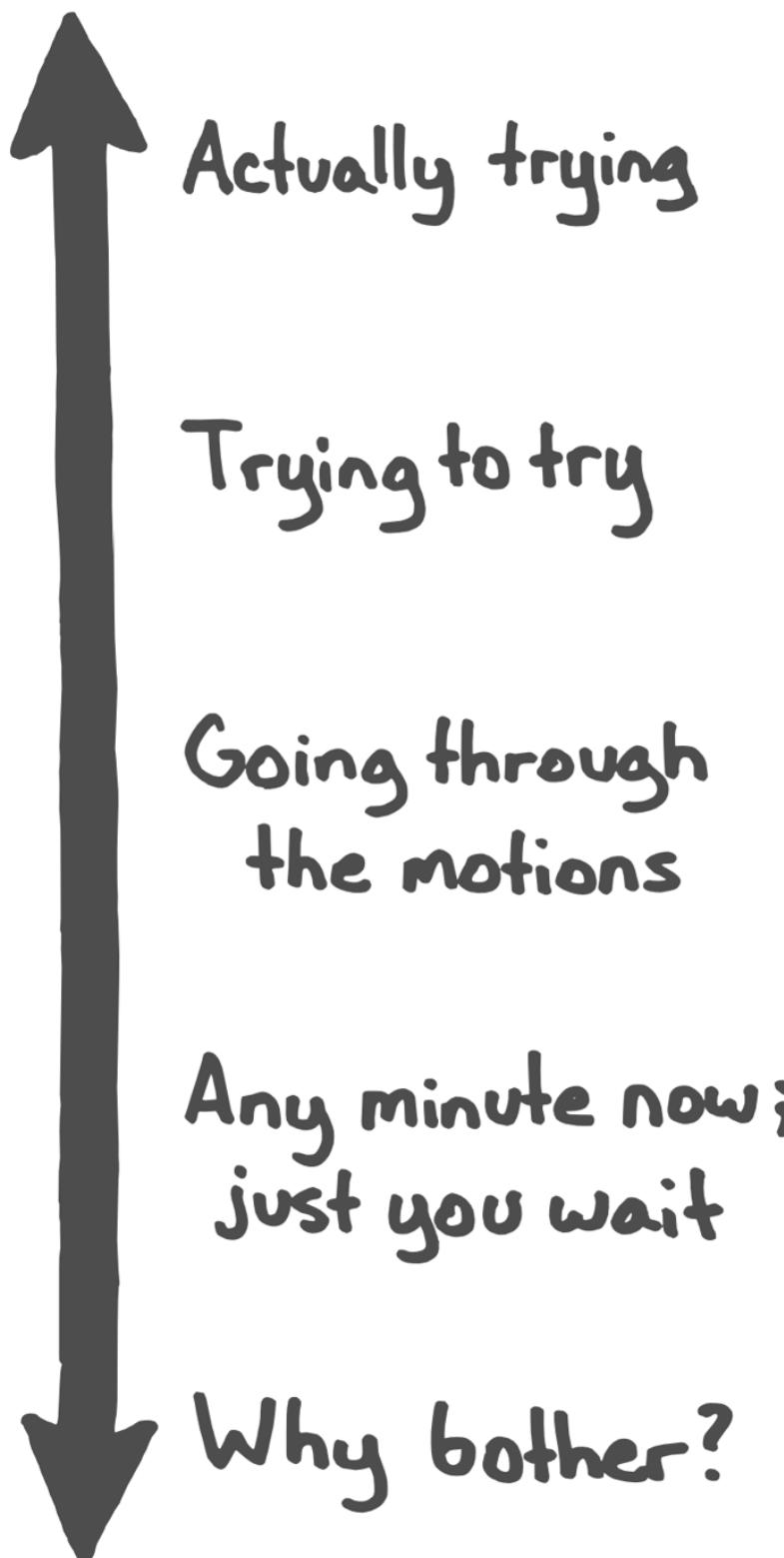
But it works. Not always, not perfectly, but shockingly often and surprisingly well. And so we recommend that you suspend your disbelief (it's justified) and put your objections on hold (we were just as incredulous as you are) and give it an actual, honest shot. In the worst case, if it does you absolutely no good, you've only wasted five minutes, and you've successfully exercised your Try Things muscles.

---

## **Post-hoc and half-baked**

We'll provide more detail in later sections, but the core of the technique—set a timer and solve your problem in five minutes or less—is extremely straightforward. The question is, why does this work? What's going on?

We don't have a complete answer yet, but we do have some quasi-models that pseudo-explain parts of what might be happening for some subset of hypothetical people (maybe).



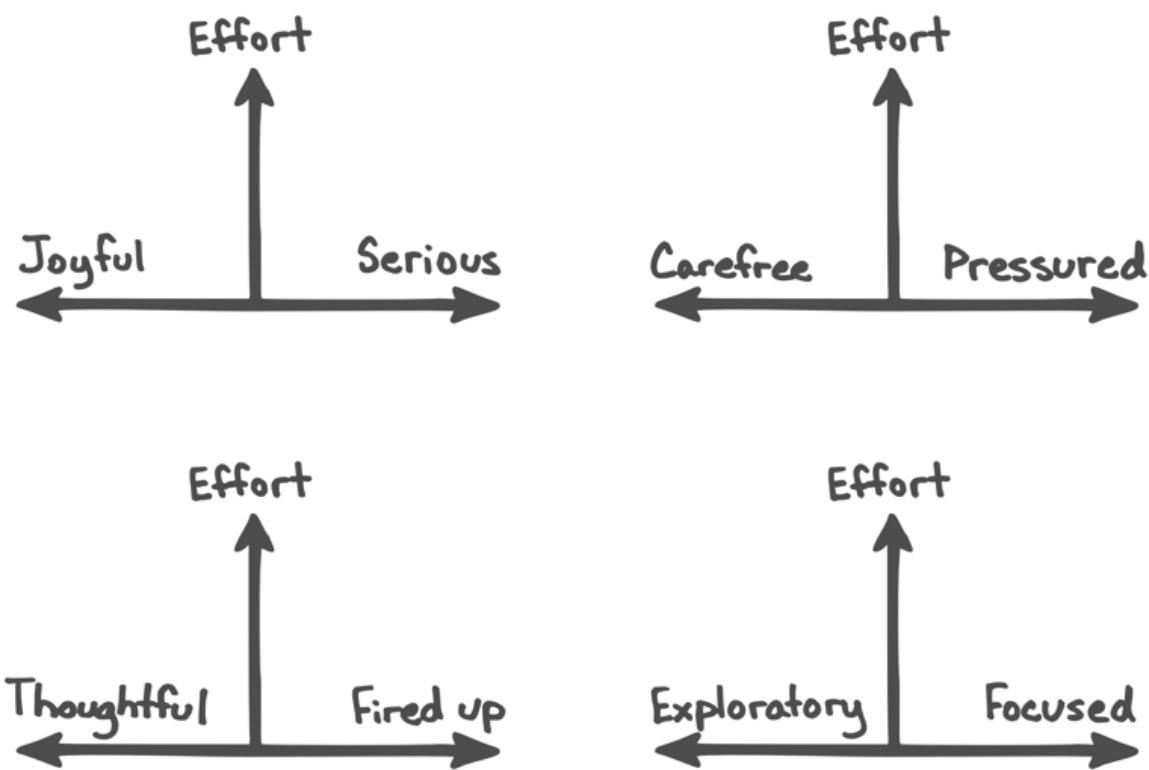
If we look at the line above, it's clear that, *within the context of a given problem or project*, we'd like to be operating as close to the upper end as possible. This is assuming that the

project is genuinely important, that we aren't in need of a break or a vacation, that we aren't neglecting something else, etc.

There are situations which naturally bring out the Actually Try, such as deadline mode or emergencies, but ideally, we'd like to be able to access it at will, rather than by having to trick ourselves into panic and stress.

There's also more to it than time pressure and dire consequences. Yes, most people find themselves much more productive in the last few hours before the assignment is due, and that's at least partially because they no longer have an affordance to meander or procrastinate. If it takes three hours to finish, and your job depends on it, and you have three hours left, then there's not much doubt about what you're going to do (unlike earlier in the week, when quitting a schedule slot only meant quitting that slot, and didn't have any real bearing on your overall career).

But athletes in flow state, children at play, actors doing improv, artisans working on their craft, mathematicians theorizing, gamers at tournaments, and people cooking a special meal for friends and family also Actually Try, with no time limit and nothing immediately obvious at stake. Indeed, if we were to expand our line out into a two-dimensional graph, it's not at all clear what the second axis should be, nor which side of it is better to be on.



Ultimately, we suspect that the actual answer is “whichever side helps you move upward on the graph, per the specifics of the situation and your own motivational structure.” Some people find that they do their best work in a harshly disciplined, drill-sergeant sort of mode, where there's no forgiveness and no wiggle room. Others find that sort of pressure extremely counter-productive, and perform better with less shouty-crisis-willpower stress, not more. Additionally, most people aren't consistently one-sided. It's likely that you'll find a playful

spirit helpful in certain cases, and a hardcore attitude useful in others.

# Effort



Your path to greater effort, where X is whatever quality makes greater effort more likely to happen and less painful to experience.

---

## Less, not more

Okay, so—how does a five-minute timer help you actually *do it*? One theory is that the timebox allows you to do *less* of certain kinds of thinking that generally inhibit progress. It's a paring down, rather than an addition—there are certain mental strategies and mental filters which most of us keep on as a general rule (and for good reason), but which an ideal “cheap experiment” lets us temporarily abandon.

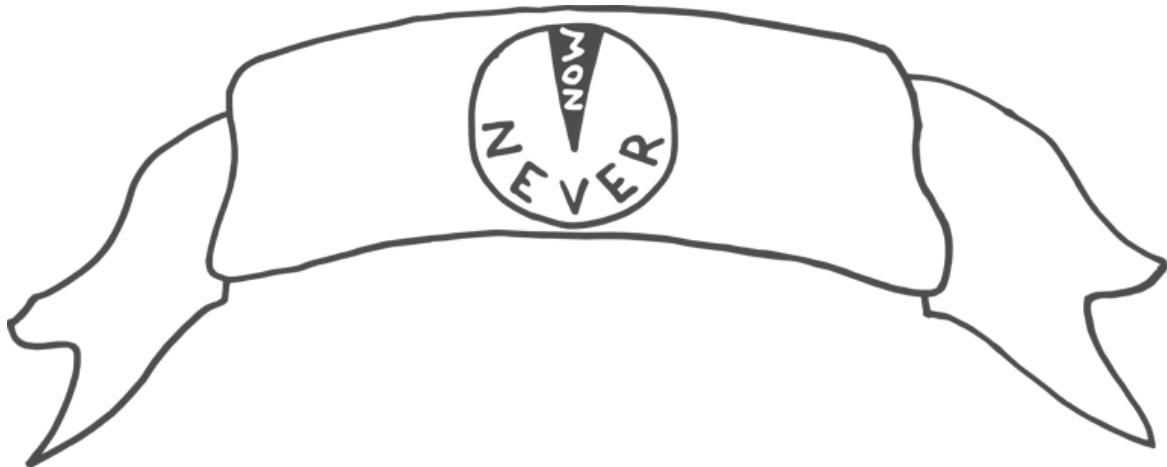
For instance, many of us more or less constantly run a **mental censoring** algorithm—we actively stop ourselves from thinking things that are useless, irrelevant, nonsensical, immoral, manipulative, or otherwise outside of our identity. When attempting to solve an interpersonal problem, we avoid reaching for monetary solutions; when dealing with negative feelings, we try not to be overtly judgmental and blame everything on others; when brain-storming “ways to get a decent job,” we don’t usually come up with things like “forge a diploma” or “chain favors together until a CEO owes us one.” We typically don’t bother trying to solve our long-term health struggles with nothing but the stuff in our pockets—except when the five minutes have already started, and those are all the resources we have on hand.

As another example, people (especially those who attend applied rationality bootcamps) often keep **strategic running tabs** on whether their current activities are effectively pointed at their goals. We tend to spend some fraction of our attention asking questions like “How long is this going to take?” or “Is this still worth it?” or “Am I even heading in the right direction?” For people who are focused on maximizing their potential (rather than merely doing well generally), that fraction can be large enough to put a serious dent in their productivity, making it hard to get started and hard to keep going, and sometimes resulting in decision paralysis.

There’s also the question of **conservation**—for many people, effort is a limiting factor, and it’s scary to embark on a project that requires you to commit a lot of resources. It’s very easy to ask the question “Am I ready for this right now?” and come up with a lot of reasons to say “No” if the task is at all large or daunting.

A resolve cycle blows the lid off these restrictions. There’s no need to worry about wasting time, because the clock is only set to five minutes. It’s okay to uncensor yourself, because you’re *supposed* to think outside the box. You don’t have to conserve energy, because it’s just a quick sprint, with no further commitment beyond that. And yes, there are real benefits from the artificial deadline and the sense of now-or-never, which help a lot of us get over the initial “activation energy” of laying hands on a thorny problem.

At their best, resolve cycles are a letting go, a putting-on-of-the-headband, a moment when we hold off on asking *why* or *whether* and instead start asking *what* and *how*. They provide a strong bias toward action, which is a valuable counterweight for those of us who tend to default to hesitation, consideration, and caution. They’re not for everyone, and they’re not for every problem, but they’re an excellent tool to have in the toolkit.



---

## The Resolve Cycle technique

**1. Choose a thing that you would like to solve.** This could be a bug you're trying to get rid of, a potential you're trying to realize, a project you'd like to start or complete... anything. Don't be afraid to pick something big, and don't be ashamed to pick something small.

**2. Try to solve the problem—in five minutes.** Yes, actually. No, don't just *make a plan*; try to completely solve it. If there are any steps left to future you, try to make sure they're effortless and very hard to mess up (e.g. you solved the problem by ordering something on Amazon, and it's not hard to open a box once it arrives). A good target is "even if I just run on autopilot from now on, and can't actually put forth agency or effort, this problem won't be a problem anymore."

**3: If the first five-minute timer didn't cut it, spend five minutes brainstorming five-minute next actions.** Now that you've come up against some of the obstacles, use your second resolve cycle to make a *list* of things that you could do to make progress, where each item on the list is itself doable in five minutes or less. (So, for instance, "drafting a quick email" or "doing five minutes of research" or "meditating on a single TAP.")

**4. Set a five-minute timer and do the most promising item on your new list.** At this point, you're set up for success, but you want to get some momentum on those next actions. Do at least one resolve cycle, so that your new list is an "in progress" rather than a "to do."

Or, to use Hogwarts houses as a metaphor, our first five-minute timer is Gryffindor, boldly trying to solve the problem. Our second is Ravenclaw/Slytherin—cleverly scheming all sorts of possible next actions. And our third is Hufflepuff, diligently chipping away at it.

---

A few further thoughts on the process:

The first timer is very important. Even complex and intractable-seeming problems often turn out to have short or simple solutions; we often (reasonably!) skip over the "easy answer" bucket entirely when we go to tackle something hard. After a few cracks at resolve cycles, though, you'll learn to be suspicious of people who claim their problem can't be solved in five minutes, and *also haven't actually given it a shot*. Give yourself permission to succeed—worst case, you'll spend a few minutes getting a clearer sense of the possibility space.

For all of the steps, it sometimes helps to use narrative framing as a tool. For instance, what if I would give you literally a billion dollars if you solved the problem in the next five minutes? Or, what if, at the end of the cycle, a genie will permanently freeze your neural patterns in this one domain, so that this is literally your last chance to improve? Many people find that working under these or similar frames gives them additional energy or affordances.

Problem reframings can be useful, too—if you're having a hard time getting away from thoughts you've already had over and over again, try asking yourself some of the following questions:

- What's concretely different about the universe where I've already solved this problem? What things would I be able to see or measure?
- How would I become the sort of person for whom this problem isn't hard, or never even comes up?
- How would I solve this problem if I were [Person X]? How would I advise [Person X] to solve this problem, if it were theirs?
- Why do I want to solve this problem? What's it going to unlock? What do all my ideas and efforts so far have in common? What axes am I not moving on?
- How have I felt during my previous attempts to solve the problem? Should I be harder on myself, or gentler? More frantic, or more measured? Is this a problem that calls for curiosity and exploration, or for determination and drive?

Be sure to take breaks—for many people, resolve cycles are a high-energy burn, and trying to do too many in a row or trying to do them without enough time in between could mean

driving yourself very hard into a hole.

Also, take advantage of all available resources—use pen and paper! Use your computer (as long as it doesn't diffuse your focus)! Use other people, if you have them available to you and your first solo attempt doesn't crack it.

Finally, take note of your successes, both the concrete ones and the cognitive or meta-level ones (even if you don't make progress, if you stayed on it and ruled out a lot of bad options, you've done real work and should pat your brain on the back).

---

## Developing a "grimoire"

Over time, you may find that you develop a standard set of prompts and actions that you find useful to draw on when doing resolve cycles—your own personal grimoire of debugging exercises. Here is an example of what one person's grimoire might look like (this one from a participant who was focused on changing emotional patterns and developing character traits):

### Exploring the problem space

- Five terrible models of what might be going on
- Similar problems I've solved before
- Five situations this reminds me of
- Details of the experience of [Feeling X]
- Three times I would have expected to have this problem and didn't
- Three times I had this problem recently
- Three times where I didn't expect to have this problem, but did
- Times when I've done well at handling this

### Eliciting/navigating hesitations

- End-goal alternatives to my current plan
- List of known or suspected obstacles
- Pre-hindsight: I achieved my goal and everything was bad; why?
- Button test: I can push a button to achieve my goal. Any reluctance?
- What's bad about getting better at this?
- What's good about the status quo?
- Spend five minutes inhabiting the unpleasant present. Can it be made livable, if left unsolved?

### Generating possible solutions

- Ten terrible ideas for step one
- Times when I've felt this way before, and what got me out of it
- What are the prerequisite subskills for success? How can I get them?
- Pick a time when I *didn't* navigate this well, and rewrite history. Where do I make changes, and what are they?
- Create five to ten relevant TAP

### Hacks/shortcuts to victory

- Generate a narrative for why this has been useful or necessary or helpful to me in the past, but why that isn't true any longer (i.e. why I no longer need the crutch)
- Explain why this is a particularly good moment for me to make a big shift or tackle this problem
- Imagine my future successful self looking back and encouraging me, having reaped all the benefits. What do I say to myself?

- Think of a skill I'm already good at, and explain how *this* skill is really just a transformation of that one
  - Meditate for five minutes on why solving this is useful
  - Decide that I'm *just not going to fail*.
- 

## Resolve Cycles—Further Resources

Research on attention and task switching has found that there is a large benefit to focusing on one task at a time. Task switching causes a large temporary drop in performance immediately after a task switch and a smaller persistent impairment as long as switching tasks is a possibility. Being engaged in a task activates a variety of cognitive processes (involving attention, memory, etc.) that are relevant for performing that particular task, which are collectively known as a task-set. One proposed explanation for the impairments caused by task switching is that they are due to the cost of switching task-sets and of having multiple competing task-sets activated at once.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7, 134-140.  
<http://goo.gl/f6Ek3>

A brief summary of the psychological research on multitasking:  
<http://www.apa.org/research/action/multitask.aspx>

---

Robert Boice (2000) studied the productivity of published professors. He found that the academics who were prolific writers often had a habit of writing for at least 15 minutes every day, while less productive academics tended to write for longer blocks more occasionally. Boice argued that regular short periods of writing drastically reduced the barrier to getting started, and that the frequency improved idea generation. Interventions that encouraged less productive professors to write briefly each day were effective at increasing the amount that they wrote, as well as the number of ideas that they had.

Boice, Robert (2000). *Advice for new faculty members: nihil nimus*.

A brief summary of Boice's work: <http://www.bmartin.cc/classes/writing.html>

---

Self-efficacy is the belief that one is capable of achieving a goal or accomplishing a task. Albert Bandura (1986; Bandura & Locke, 2003) describes respectably strong correlations between high self-efficacy and several attributes that make success more likely such as willingness to take on new challenges, persistence in the face of difficulty, and a tendency to assume that one directs and shapes one's future rather than simply reacting to events as they arise.

<http://en.wikipedia.org/wiki/Self-efficacy>

Bandura, A. (1986). *Social foundations of thought and action*

Bandura, A., & Locke, E. A. (2003). *Negative self-efficacy and goal effects revisited*. *Journal of Applied Psychology*, 88, 87-99. <http://goo.gl/ab39bN>

# Unifying Bargaining Notions (2/2)

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Alright, time for the payoff, unifying everything discussed in the [previous post](#). This post is a lot more mathematically dense, you might want to digest it in more than one sitting.

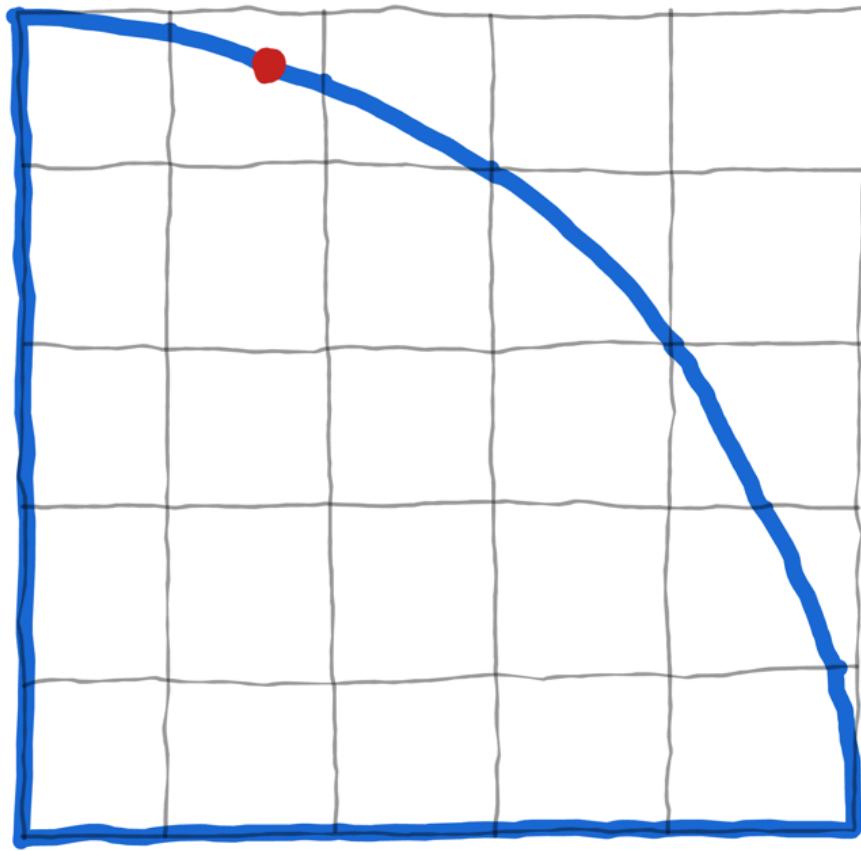
## Imaginary Prices, Tradeoffs, and Utilitarianism

Harsanyi's Utilitarianism Theorem can be summarized as "if a bunch of agents have their own personal utility functions  $U_i$ , and you want to aggregate them into a collective utility function  $U$  with the property that everyone agreeing that option  $x$  is better than option  $y$  (ie,  $U_i(x) \geq U_i(y)$  for all  $i$ ) implies  $U(x) \geq U(y)$ , then that collective utility function *must* be of the form  $b + \sum_{i \in I} a_i U_i$  for some number  $b$  and nonnegative numbers  $a_i$ ."

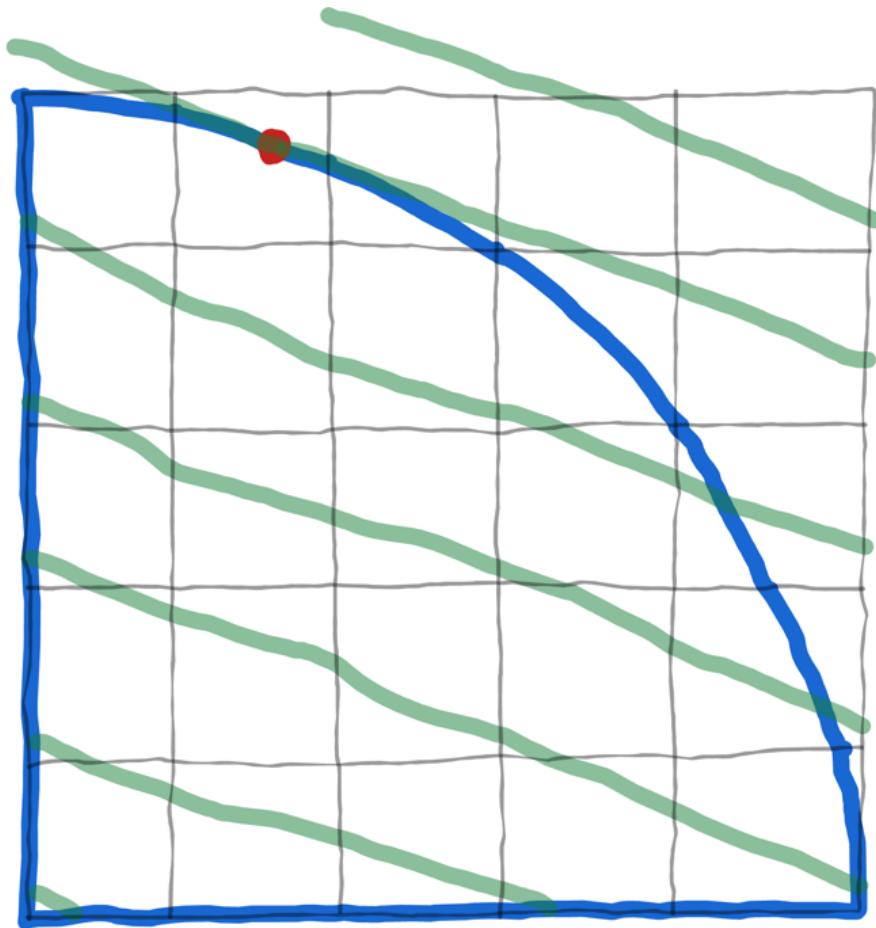
Basically, if you want to aggregate utility functions, the only sane way to do so is to give everyone importance weights, and do a weighted sum of everyone's individual utility functions.

Closely related to this is a result that says that any point on the Pareto Frontier of a game can be *post-hoc* interpreted as the result of maximizing a collective utility function. This related result is one where it's very important for the reader to understand the actual proof, because the proof gives you a way of reverse-engineering "how much everyone matters to the social utility function" from the outcome alone.

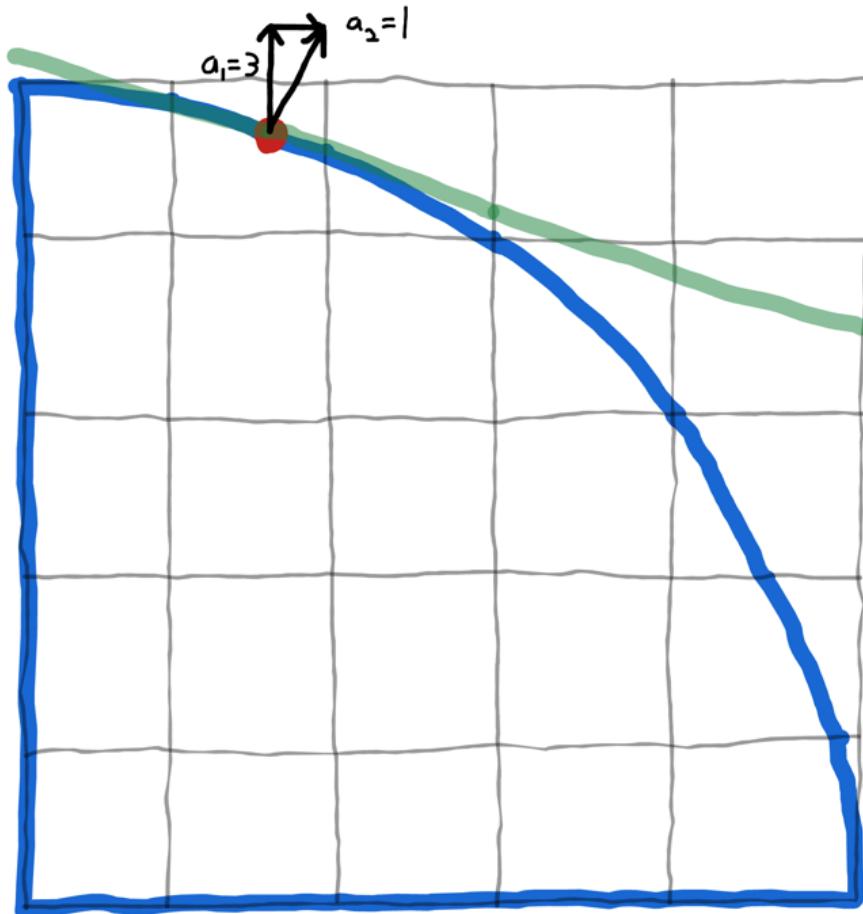
First up, draw all the outcomes, and the utilities that both players assign to them, and the convex hull will be the "feasible set"  $F$ , since we have access to randomization. Pick some Pareto frontier point  $u_1, u_2 \dots u_n$  (although the drawn image is for only two players)



Use the [Hahn-Banach separation theorem](#) to create a linear function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $\phi(u_1, u_2, \dots, u_n) \geq \phi(F)$ . (such that is abbreviated s.t. from here on out) Or put another way,  $u_1, u_2, \dots, u_n$  is one of the points in the feasible set  $F$  that maximizes the linear function  $\phi$  you created. In the image, the lines are the level sets of the linear function, the set of all points where  $\phi(x_1, x_2, \dots, x_n) = c$ .

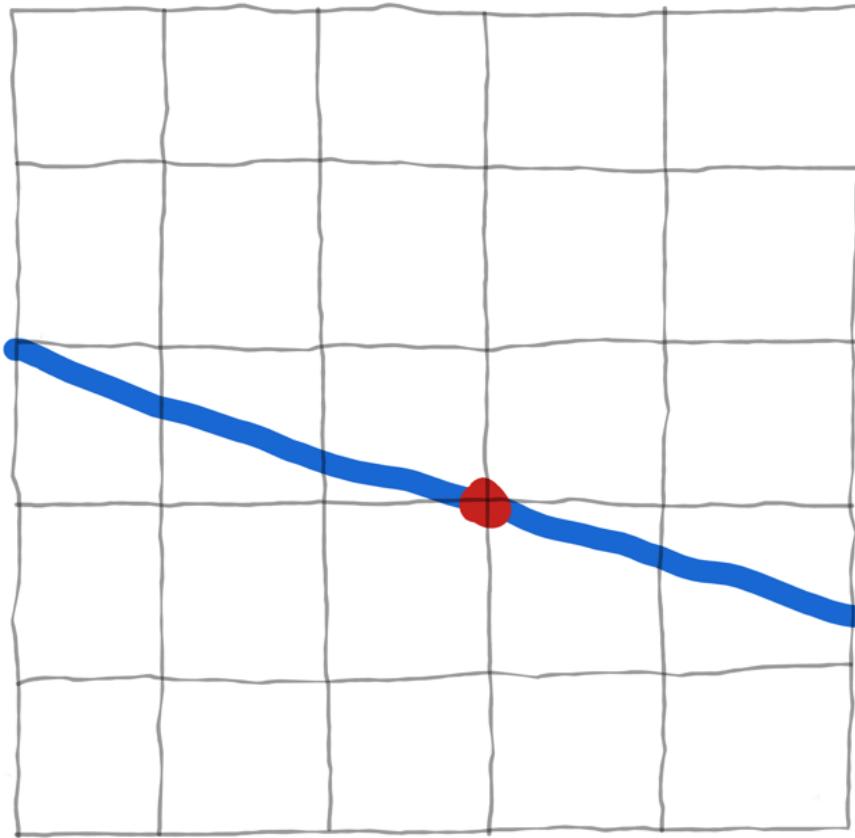


That linear function  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$  can be written as  $(x_1, x_2, \dots, x_n) \mapsto a_1x_1 + a_2x_2 + \dots + a_nx_n$ . Bam, those coefficients are the utility weights you need.  $u_1, u_2, \dots, u_n$  is a point that maximizes the function  $\phi$ , and the function  $\phi$  is implementing "take this particular weighted sum of the utilities of the players", so we have rationalized our particular Pareto-optimal point as being produced by maximizing some weighted sum of how important everyone's utilities are.



And that's how to take any point on the Pareto frontier and reverse-engineer the weighted sum of everyone's utilities it's maximizing (though if there are corner points, there can be multiple possible weights that'd work, because the tangent plane is no longer unique).

But, there's another completely different way of viewing this process! If we take our Pareto-frontier point  $u_1, u_2 \dots u_n$  and zoom way way in...



There's a locally linear tradeoff between the utilities of the various players. An  $\epsilon$  increase in the utility of Alice corresponds to a  $3\epsilon$  decrease in the utility of Bob. One thing that we can do with this local linearity is invent an imaginary currency! It goes like this. One curnit (currency unit) can be redeemed for a tiny adjustment back and forth along this part of the Pareto frontier, and agents can trade curnits back and forth as needed to adjust exactly where they are on the Pareto frontier. And in particular, the fact that there's a 3 to 1 ratio between how the utility of Alice trades off against the utility of Bob corresponds to Alice needing to spend 3 curnits to get an AliceUtilon, while Bob only needs to spend 1 curnit to get a BobUtilon.

There's a few ways of thinking about this. The first way of thinking about it is that, in this little piece of the Pareto frontier, Alice is 3x harder to satisfy. Another way of thinking about it is that it's like Bob is poor and his marginal utility for a dollar (or a curnit) is a good deal higher than it is for Alice. And a third way of thinking about this is that if we find out this is the best collective point, we can go "huh, the only way that 3 curnits/AliceUtilon and 1 curnit/BobUtilon makes sense is if an AliceUtilon is worth 3x as much as a BobUtilon". Which, oh hey, is the *exact same conclusion* as we would have gotten from trying to figure out the weights for Alice vs Bob in the social utility function that says that this is the best point to be at.

So, combining these views, we can take any point  $u_1, u_2 \dots u_n$  on the Pareto frontier, and get a vector  $a_1, a_2 \dots a_n$  which can be interpreted *either* as "these are the importance weights for the players", or as "these are the curnits/utilon conversion factors for the various players".

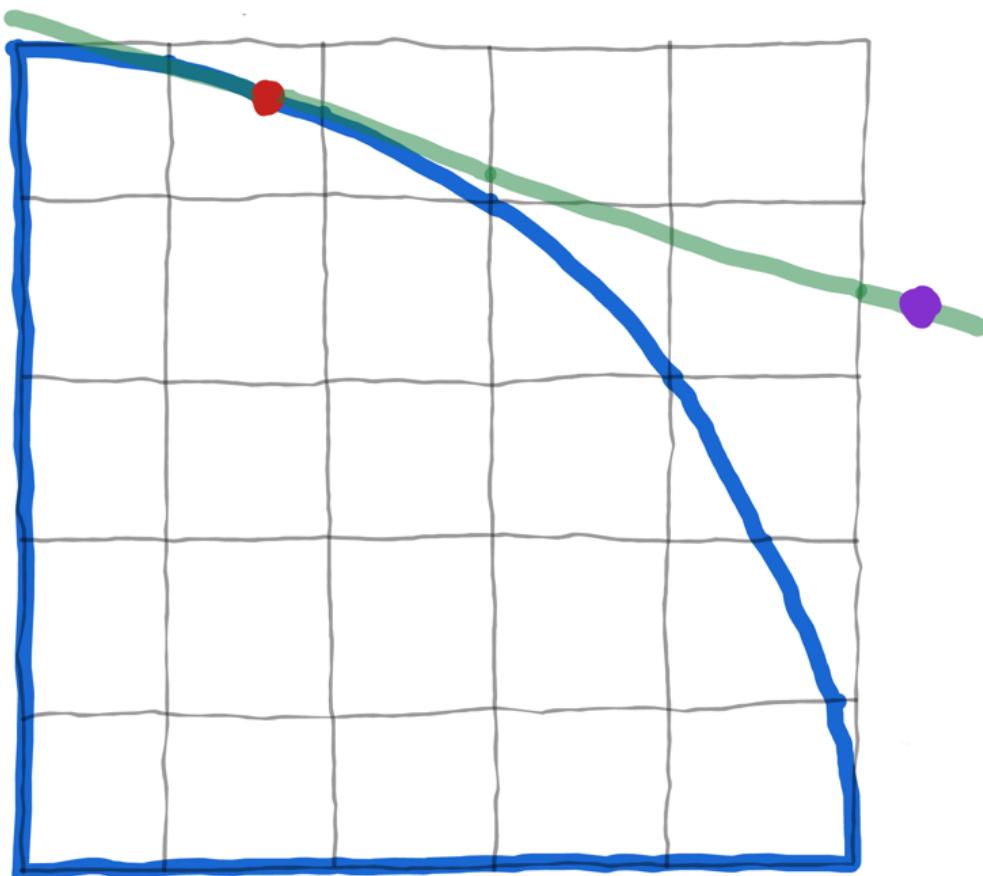
### **CoCo Equilibria**

So, the CoCo works really great for games with transferrable utility, some sort of currency that can be passed around amongst the players. But there are a lot of games without transferrable utility!

But remember our earlier discussion on how, in the local neighborhood of a Pareto-frontier point, we can invent an imaginary currency that reflects how hard it is to improve the utilities of the various players. This works fine in the vicinity of the point, but breaks down as you stray far away.

So, let's take our given example with Alice and Bob. If we introduce "curnits" as a currency in the local vicinity of the point they're at, then we can convert from utility functions  $U_1, U_2$  (denoted in AliceUtilons and BobUtilons), to  $a_1U_1, a_2U_2$  (both players' utilities are denoted in curnits now, and are commensurable), use the CoCo values to tell us what payoffs the players "should" get, and divide the result by  $a_1$  and  $a_2$  respectively to convert it back into AliceUtilons and BobUtilons. When we end up doing this with our example, we'll get a result that has the following reasoning behind it.

"Since curnits are much more valuable to Bob than they are to Alice, the CoCo games will advise that Alice give Bob some money to get Bob to go along with her plans, since the money is 3x less useful to Alice than it is to Bob. Converting the CoCo payoffs back to utilons, the net result would be "Alice gets a bit less utility than she did at the old Alice-favoring point, Bob gets a giant pile of utility from all those curnits", and it'd actually be an impossible pair of utility values, there's just no way for Bob to get that much utility.



Using the local currency at the red point for a CoCo transferrable-utility game means that Bob gets a small pile of currency in the CoCo game, which translates back into a big pile of BobUtilons, and we end up at the purple point, which is impossible to attain.

Generalizing past this particular example, in general, for Pareto-frontier points where some players lose out really hard (and so implicitly have low utility weights), then when you convert it to a game with transferrable utility, pick the CoCo value, and convert back, the players with really low utility weights will end up with giant piles of utility. This is because "low utility weights  $a_i$ " correspond to "The curnits/utilon value  $a_i$  is low, so it takes few curnits to help this player a lot", so they get a small pile of money which converts to a big pile of utility.

And so, the question to ask now is something like "is there a point on the Pareto frontier  $u_1, u_2 \dots u_n$  where we can get the curnits/utilon conversion numbers from that point, convert everyone's utility to curnits, work out the CoCo value of the resulting game, convert back to utilons, and end up at *the exact same point we started at?*"

Basically, a CoCo equilibrium would be a spot where, if the players squabbling over what direction to move on the Pareto frontier went "let's introduce a virtual currency redeemable for tiny perturbations on the Pareto frontier", and worked out the CoCo value for the game they're in (which is a very chaotic solution when money is available), it'd return the answer "stay where you currently are, it's a good spot, nobody needs to pay anyone else anything". Which is a very fortunate result to get since this currency doesn't actually exist so nobody could pay each other anything anyways.

### **CoCo Equilibrium Questions**

There are several questions we could ask about CoCo equilibria.

- 1: Is it scale-and-shift invariant, or does it depend on how everyone represents their utility functions?
- 2: If we try using it on bargaining games in particular, will it reproduce any well-known bargaining notions?
- 3: How do we generalize it to the n-person case? Is it straightforward, or are there hidden difficulties?
- 4: If we really just found a universal notion of how to split gains in games, how come nobody else came up with it first?
- 5: Do CoCo equilibria even exist, anyways? If so, are they unique?

Time to reveal the answers, which won't be proved yet because they'll follow as a consequence of one big theorem later on.

### **Scale-Shift Invariance**

For the first question, yes, it is scale-and-shift invariant. It doesn't matter *how* you represent everyone's utility functions, you'll get the same answer. Intuitively, here's what happens. Let's say Alice multiplies all her utility numbers by 100. Now, at the point of interest, this means that we just went from 3 curnits/AliceUtilon to 3 curnits/100 AliceUtilons. And so, the coefficient we multiply Alice's utility function by,  $a_1$  (the curnits/AliceUtilons number), went from 3 to  $\frac{1}{100}$ , which will perfectly cancel out the fact that Alice multiplied all her numbers by 100. So, the CoCo value (as denominated in curnits) doesn't change one bit. Then we divide by  $a_1$  to convert back to Alice's utility, which means that we multiply by  $100$ , and get a big number for AliceUtilons, as expected (since Alice multiplied all her numbers by 100)

As for shifting, if Alice adds 10 utility to all her numbers, it doesn't alter the coefficient  $a_1$  (3 curnits/AliceUtilon), so all of Alice's utility payoffs as denominated in curnits, are 10 higher than usual. But, CoCo value is shift-invariant. If Alice gets a guaranteed 10 extra curnits no matter what she does, her CoCo value will be 10 curnits higher than it'd be usually, and Bob's won't change at all. And so, when we divide by  $a_1$  to convert back to Alice's utility, we get an extra 10 utility, as expected (since Alice added 10 to all her numbers)

Ok, we've got scale-and-shift invariance, which is a super-important property to have for something to maybe be "chaa" (a mathematically distinguished point in a negotiation game against a foe where you both have an interest in preventing destructive conflicts, that's neutral enough that aliens would probably come up with it).

### Bargaining Games as Special Case

If we apply CoCo equilibrium concepts to bargaining games in particular (player 1 proposes an option or plays "reject", player 2 can accept or reject, anyone rejecting means that both sides get their disagreement payoffs), what do we get?

Well, though it won't be proved now (it'll be indirectly proved later on), CoCo equilibria in bargaining games will turn out to be equivalent to the Nash bargaining solution! The Nash solution can be derived as a special case of this generalization of the CoCo solution to when utility isn't transferrable!

### N-Player Generalizations and Coalitions

For generalizing to the n-person case, there's the obvious generalization where, given a Pareto frontier point, we can get the imaginary prices  $a_1, a_2 \dots a_n$ , and the CoCo value makes sense for n-player games.

But this doesn't fully take coalitions into account. It's possible that a coalition could conspire amongst themselves to guarantee a good payout. And we could add extra conditions regarding coalitions, like that *within* a coalition, they use some sort of CoCo-like or Shapley-like split of resources.

To formalize the stronger version of that equilibrium, let  $N$  be the set of players, and  $U_i$  be the utility function of player  $i$ , and  $A_i$  be player  $i$ 's set of actions.

#### *Formal Definition: Coalition-Perfect CoCo Equilibrium*

A coalition-perfect CoCo equilibrium is a tuple  $\{ \rho^S \}_{S \subseteq N}$  (the joint strategies that all possible coalitions would play if they were against the opposing coalition,  $\rho^S \in \Delta \prod_{i \in S} A_i$ ), s.t, defining

$$u_i := U_i(\rho^S, \rho^{N/S})$$

1:  $\{ u_i \}_{i \in N}$  is on the Pareto frontier.

2: There is an  $\{ a_i \}_{i \in N}$  tuple (virtual prices) that makes both of the following conditions true.

$$3: \forall S \subseteq N : \rho^S \in \operatorname{argmax}_{\rho \in \Delta \prod_{i \in S} A_i} \left( \sum_{i \in S} a_i U_i(\rho, \rho^{N/S}) - \sum_{j \in N \setminus S} a_j U_j(\rho, \rho^{N/S}) \right)$$

(ie, all the joint strategies are trying to maximize the money earned if up against the opposing coalition in a zero-sum game, and as a special case, when  $S=N$ , it says that what the entire group actually ends up doing maximizes surplus value, which is another way of stating that the  $\{ a_i \}_{i \in N}$  are the

appropriate virtual currencies to use at the  $\{ u_i \}_{i \in N}$  point)

$$4: \forall i \in N, i \in S \subseteq N : a_i u_i = \sum_{i \in R \subseteq S} \frac{1}{(|R| - |S|)! (|S| - |R|)!} \left( \sum_{j \in R} a_j u_j - \sum_{k \in S \setminus R} a_k u_k \right)$$

Or, rephrasing this in terms of the more standard framing of Shapley Value...

$$\forall i \in N, i \in S \subseteq N : a_i u_i = \sum_{R \subseteq S \setminus \{i\}} \frac{1}{(|S| - |R| - 1)! |R|!} a_j u_j - \sum_{j \in R} a_j u_j$$

So, that's a coalition-perfect CoCo equilibrium. You could interpret it as there being a "virtual currency" phrased in terms of how hard it is, relatively, to improve everyone's utility, and everyone getting their CoCo value, and even in the event of zero-sum conflicts between teams, everyone on a team will get their CoCo value. Or you could interpret it as everyone getting their Shapley payoffs, where the

analogue of the marginal gain from a particular player is "their value when they join team S, plus the improvements in everyone else's value from them no longer opposing team S". Or you could interpret the game, and all the zero-sum subgames, as just maximizing a weighted sum of utilities (like Harsanyi's utilitarianism theorem), and the cool part is the "weights" for how important everyone is will be the same for the full game as well as all the zero-sum subgames.

### Harsanyi Equilibria and Generalizing the Nash Bargaining Solution

If this exists and it's nice, why has nobody found it before?

Actually, someone did find this before! That's a major occupational hazard of finding math that aliens would independently reinvent: there's a high chance that someone beat you to the punch. Specifically, John Harsanyi, back in 1963, found this first. He wound up with this same exact solution, though it's quite nontrivial to show the equivalence between our equilibrium notions.

CoCo equilibria were motivated via the nice properties of the CoCo value and generalizing it to the non-transferrable utility case, which turned out to be secretly equivalent to generalizing Shapley values. Harsanyi, as detailed in his lovely paper "[A Simplified Bargaining Model for the n-Person Cooperative Game](#)" (which I very highly recommend you read via your favorite paper-reading website!), found it from trying to generalize the Nash Bargaining Solution to games without well-defined disagreement points.

Harsanyi's basic insight was that, in a general two-player game, if it's known in advance that the Nash bargaining solution will be used, and the players are picking their "disagreement strategies" (what they'll fall back on if they can't cooperate on some joint distribution over actions) then Alice would try to pick a "disagreement strategy" that makes it so that, no matter what Bob's disagreement strategy is, the Nash bargaining solution would favor Alice as much as possible, and Bob is in a similar position. So, the two players will end up in a game where they're trying to disagree in a way that'll rig the Nash bargaining solution in their favor. I'm not sure whether or not this is zero-sum, but it is true that if one player wins, the other must lose, so it's zero-sum *enough* that there's a unique pair of disagreement utilities that you get from maximin strategies, that are mutually optimal against each other, and then you can just use the Nash bargaining solution from there.

In particular, if you're trying to make a threat that's suitable for rigging a bargaining game in your favor, what you need are not threats that hurt both of you equally, or threats that are worse for you than for the foe, or threats that the foe could possibly defuse. What you need is something that matters far more to the foe than you, which the foe can't evade by any action they can take. Or, rephrasing in terms of the CoCo value, to successfully rig the Nash bargaining solution in your favor, you'll need a good move in the zero-sum competition game of the cooperation/competition decomposition.

Generalized to the n-player case, between any two coalitions, they'll be doing the same sort of squabbling over what counts as the disagreement point, everyone within the coalition will agree on what disagreement point to go for in event of conflicts and within any coalition, they'll also be splitting

things according to the Nash bargaining solution as well. I don't fully understand the reasoning behind how that informal sort of description cashes out in the math (in particular, I still don't understand why the disagreement points are defined as they are in the paper), but I'll attempt a summary of Harsanyi's paper anyways. You're *highly* encouraged to read the paper yourself; it's got lots of goodies in it that I don't mention.

Harsanyi starts off by assuming that every coalition can guarantee some marginal gain to everyone making it up, which is denoted by  $w_i^S$ , the marginal utility payoff to player  $i$  received from the coalition  $S$ .

Further, the payoff that a player gets in the event of a zero-sum conflict between  $S$  and everyone else

should just be the sum of the marginal payoffs from all the subsets of  $S$  that  $i$  is in, ie.  $u_i^S = \sum_{i \in R \subseteq S} w_i^S$

(where  $u_i^S$  is as previously defined, the utility that  $i$  gets in the event of a zero-sum conflict between  $S$  and everyone else).

The argument for this is that if the sum of marginal payoffs was more than  $u_i^S$  (the payoff that  $i$  gets in the event of a zero-sum conflict), the coalitions collectively would be promising more utility to player  $i$  than can actually be guaranteed, and they're making unfulfillable promises. But player  $i$  should really be picking up all the utility promised to it from all the coalitions, and not leaving excess on the table, and so we get equality.

As it turns out, if that equation holds, then you can work out how all the  $w_i^S$  must be defined: it must

hold that  $w_i^S = \sum_{i \in R \subseteq S} (-1)^{|S|-|R|} u_i^R$ . This is around where I have problems. I just can't quite manage to get myself to see how this quantity is the "slice of marginal utility that coalition  $S$  promises to player  $i$ ", so let me know in the comments if anyone manages to pull it off.

Then, we go "ah, if  $w_i^S$  is the marginal gain from being part of coalition  $S$ " (which, again, I can't quite see), then  $w_i^S = u_i^S - t_i^S$ , where  $u_i^S$  is the payoff to  $i$  from playing its part in  $S$ 's minimax strategy, and  $t_i^S$  is  $i$ 's threat utility/disagreement point utility from being in coalition  $S$ . And so, the disagreement point

utility of player  $i$  within coalition  $S$  must be  $t_i^S = \sum_{i \in R \subseteq S} (-1)^{|S|-|R|+1} u_i^R$ .

Again, I can't see *at all* how this is true from the equation alone, but other people might be able to understand it. I can see how, in the special case of the coalition "Alice and Bob", it reproduces the intuitive result of Alice's disagreement-point-utility being "screw you Bob, I'll go off on my own and fight the rest of the world" (ie,  $t_1^{\{1,2\}} = u_1^{\{1\}}$ ), but I can't see how it extends to larger coalitions than that.

Anyways, a few interesting results are derived and discussed from there. One is that if two players  $i$  and  $j$  (Ione and Jake) are arguing over their split of the gains in every coalition  $S$  that contains both of them,

there's a well-defined fallback point for Ione which is  $\sum_{S \subseteq N: i \in S, j \notin S} w_i^S$  (the sum of payoffs from every coalition that contains Ione but lacks Jake), and symmetrically for Jake, and if they do Nash bargaining from there, then it's possible to apply a result on Nash bargaining in subgames (intuitively, both players want to pick a point that doesn't worsen their bargaining position for the full game) to derive that Ione and Jake will agree to the same ratio for how to split coalition gains between them, in every coalition game. So, if Ione and Jake are splitting value 60/40 in one coalition, they're doing that same split in

every coalition. This can be used to derive the general result that, in every coalition containing two or more players, they'll all play the Nash bargaining solution against each other.

And there's another interesting and highly nontrivial result from Harsanyi's paper, which effectively says that if the two coalitions of S and N/S (everyone who isn't in S) appoint an individual member to decide on the threat strategy that their coalition will follow, and the appointed representatives lone and Jake

only care about maximizing their own payoff in the overall game (ie, maximizing  $u_i^S$  and  $u_j^N$ ), (ie. they know that the zero-sum fight between S and N/S probably isn't happening, it's just a bargaining chip to get good overall payoffs, and they just care about their own overall payoff, not the interests of the rest of their coalition), then the threat strategies they'll pick will be minimax threat strategies for the S vs N/S zero-sum game. Correspondingly, it doesn't matter *what* players the coalitions S and N/S appoint as representatives, they'll end up picking a minimax threat strategy to maximize their payoff in the overall game.

What Harsanyi eventually ended up deciding on as the equilibrium conditions were as follows (slightly re-expressed for notational compliance). Let  $U_i$  be the utility function of player i, and  $A_i$  be their space of actions.

#### **Formal Definition: Coalition-Perfect Harsanyi Equilibrium**

A coalition-perfect Harsanyi equilibrium is a tuple  $\{\rho^S\}_{S \subseteq N}$  (the joint strategies that each coalition

would play if they were against the opposing coalition,  $\rho^S \in \Delta \Pi_{i \in S} A_i$ , s.t, defining  $u_i^S := U_i(\rho^S, \rho^{N/S})$

(payoff to player i if coalitions S and N/S fight), and  $t_i^S := \sum_{j \in R \setminus S} (-1)^{|S|-|R|+1} u_j^N$  (the fallback utility value for player i in coalition S).

1:  $\{u_i^S\}_{i \in N}$  is on the Pareto frontier.

2: There is an  $\{a_i\}_{i \in N}$  tuple (virtual prices, though Harsanyi didn't use the term) that makes both of the following conditions true.

$$3: \forall S \subseteq N : \rho^S \in \operatorname{argmax}_{\rho \in \Delta \Pi_{i \in S} A_i} \left( \sum_{i \in S} a_i U_i(\rho, \rho^{N/S}) - \sum_{j \in N \setminus S} a_j U_j(\rho, \rho^{N/S}) \right)$$

(ie. all the joint strategies are trying to maximize the money earned if up against the opposing coalition in a zero-sum game and as a special case, when  $S=N$ , it says that what the entire group actually ends up doing maximizes surplus value, which is another way of stating that the  $\{a_i\}_{i \in N}$  are the

appropriate virtual currencies to use at the  $\{u_i^N\}_{i \in N}$  point)

$$4: \forall i, j \in S \subseteq N : a_i(u_i^S - t_i^S) = a_j(u_j^S - t_j^S)$$

(it's very nonobvious, but if you use Lagrange multipliers on the Nash bargaining problem, this is

effectively saying that for all coalitions  $S$ , the division of resources within  $S$ , using the various  $t_i^S$  as the disagreement payoffs, follows the Nash bargaining solution)

And now we get to the centerpiece theorem, that coalition-perfect CoCo equilibria are the same as coalition-perfect Harsanyi equilibria. Since Harsanyi already showed things like scale-and-shift

invariance, and that these equilibria exist, and lots of other results about them, we just need to prove equivalence and then we can lift all of Harsanyi's work - no point in rederiving everything on our own. Since three of the four conditions for the equilibria are obviously identical, the whole proof focuses on showing that the "Every coalition uses the Nash bargaining solution internally to divide gains, with suitably defined threat points" condition of Harsanyi equilibria is equivalent to the "Every coalition uses the CoCo payoffs/modified Shapley payoffs internally to divide gains" condition of CoCo equilibria.

**Theorem 1:** A tuple  $\{\rho^S\}_{S \subseteq N}$  of strategies for every coalition is a coalition-perfect Harsanyi equilibrium iff it's a coalition-perfect CoCo equilibrium.

### Equilibria Existence

And now for the final question. Do these sorts of equilibria even exist at all? That's an awful lot of conditions to fulfill at once, you've got one equilibria condition for each coalition, and there's a whole lot of coalitions.

Well, fortunately, Harsanyi proved that these sorts of equilibria exist in his paper! So we can just copy off his work. Well, technically, he did it under some moderately restrictive assumptions which don't look essential, and can probably be removed, though it'll be annoying to do so. Pretty much, his proof works by setting up a game which is related to the bargaining game, and any Nash equilibrium of the auxiliary game can be converted into a coalition-perfect Harsanyi equilibrium.

The assumptions Harsanyi made were, in particular, that the space of all possible utility values in  $R^n$  were compact (ie, nobody can get unbounded positive utility or unbounded negative utility, and this assumption is violated in full transferable utility games), and its affine hull was of dimension n, and that the Pareto frontier had no vertices in the sense that, for all points on it, there's a unique tangent hyperplane that touches that point (ie, you can *uniquely* read off the weights from *all* Pareto frontier points). Think of a sphere in 3d space: every point on the sphere surface has a unique tangent plane. But for a cube in 3d space, the edges or vertices of the cube can have multiple distinct tangent planes which touch the cube at that point.

With those assumptions, yes, there is a coalition-perfect Harsanyi equilibrium, as proven in his paper.

Harsanyi made the remark that if the Pareto frontier has vertices, it's possible to write any such game as a limit of games that don't have vertices (like, imagine a cube but all the corners and edges have been sanded down a bit, and take the limit of doing less and less sanding), in order to extend the results to games with vertices on their Pareto frontier.

Though he didn't comment on it, it seems like it's possible to also deal with the affine hull dimension issue in this way, in the sense that for any set of possible utility values whose affine hull is of dimension  $< n$ , it's possible to write it as a limit of games whose set of utility values has an affine hull of dimension n (the analogue is that any 2-dimensional shape can be thought of as a limit of 3-dimensional shapes that keep getting thinner), and presumably extend his existence result to cases like that.

He didn't *actually* do these limiting-case proofs at any point, they just seem like the sort of argument that'd need to be done to generalize his proof.

There's another question which is, "are Harsanyi/CoCo equilibria unique"?

Harsanyi made the remark that they were unique for bargaining games (where they'd give the Nash bargaining solution), games with transferable utility (where they'd give the CoCo value), and 2-player games, but weren't necessarily unique for general n-player games, and then completely refused to elaborate on this.

The problem is, although Harsanyi said there were counterexamples to uniqueness (he meant a counterexample in the stronger sense of "there's more than one tuple of utility values that's an equilibrium", not the obvious weak sense of "maybe there's different strategies for everyone that gives the same payoff"), at *no point* did he ever actually *give* such a counterexample, even in the paper he cited to that effect. This is somewhat infuriating, and I fear that the non-uniqueness of these equilibria is one of those apocryphal results that nobody ever actually got around to double-checking at any point. I'd be extremely pleased if anyone could find a paper with an actual example of such.

So, yeah, that's about it. There's one good notion of equilibria, that gives Shapley values, CoCo values, and the Nash bargaining solution as special cases, which can variously be thought of as:

1: Maximizing a suitable weighted sum of everyone's utilities, where all the various coalitions agree on the weights of everyone's utilities (so if Alice is twice as important as Bob, then Alice will be twice as important as Bob in all coalitions containing the two of them).

2: Gives everyone their modified Shapley payoffs, and all the coalitions split their gains in a Shapley way.

3: Inventing a virtual currency reflecting how hard it is to improve the utilities of everyone relative to each other and splitting the game into a bunch of coalition vs coalition fights with perfect cooperation and competition, and paying everyone accordingly.

4: Every coalition jostles for a threat strategy that gives them the most payoff from the Nash bargaining solution, and then every coalition does Nash bargaining within itself to split gains.

### **But What if it Sucks, Tho (it Does)**

So, there's one super-important aspect of this that makes it dramatically less appealing that I haven't seen anyone point out. The payoffs for everyone are determined by games of the form "coalition S fights coalition not-S, coalition S is maximizing the quantity "utility of coalition S - utility of opposite coalition", and vice-versa for the opposite coalition".

If you depart from nice comfy visualizations of games involving hot-dog selling, and ponder what that'd mean for humanity, you'll probably realize how *exceptionally ugly* those imaginary games would get.

Actually take one minute, by the clock, to think about what it means that the following equation determine people's payoffs:

$$\max_{\rho^S \in \Delta \prod_{i \in S} A_i} \min_{\rho^{N/S} \in \Delta \prod_{j \notin S} A_j} \left( \sum_{i \in S} a_i U_i(\rho^S, \rho^{N/S}) - \sum_{j \notin S} a_j U_j(\rho^S, \rho^{N/S}) \right)$$

This is why I was stressing that "chaa" and "fair" are very different concepts, and that this equilibrium notion is very much based on threats. They just need to be asymmetric threats that the opponent can't defuse in order to work (or ways of asymmetrically benefiting yourself that your opponent can't ruin, that'll work just as well).

I think it's a terrible idea to automatically adopt an equilibrium notion which incentivises the players to come up with increasingly nasty threats as fallback if they don't get their way. And so there seems to be a good chunk of remaining work to be done, involving poking more carefully at the CoCo value and seeing which assumptions going into it can be broken.

Also, next Thursday (June 28) at noon Pacific time is the Schelling time to meet in the Walled Garden and discuss the practical applications of this. Come one, come all, and bring your insights!

### **Appendix: Proof of Theorem 1 (you can skip this one)**

Since conditions 1, 2, and 3 are all obviously equivalent to each other, that just leaves that showing that the condition 4's of both types of equilibria imply each other. First, we'll show a lemma.

$$\text{Lemma 1: } u_i = \sum_{i \in R \subseteq S} w_i$$

Start off with

$$\sum_{i \in R \subseteq S} w_i$$

Unpack the definition of  $w_i^R$ .

$$= \sum_{i \in R \subseteq S} \sum_{i \in T \subseteq R} (-1)^{|R|-|T|} u_i^T$$

We can interchange this sum, and view it as picking the set  $T$  first, and the set  $R$  second.

$$= \sum_{i \in T \subseteq S} \sum_{T \subseteq R \subseteq S} (-1)^{|R|-|T|} u_i^T$$

And group

$$= \sum_{i \in T \subseteq S} \left( \sum_{T \subseteq R \subseteq S} (-1)^{|R|-|T|} \right) u_i^T$$

And we can ask what coefficient is paired with the various  $u_i^T$ . Really, there's two possibilities. One possibility is that  $T = S$ , the other is that  $T \neq S$ , so let's split this up.

$$= \left( \sum_{S \subseteq R \subseteq S} (-1)^{|R|-|S|} \right) u_i^S + \sum_{i \in T \subset S} \left( \sum_{T \subseteq R \subseteq S} (-1)^{|R|-|T|} \right) u_i^T$$

Clearly, for the former term,  $R = S$  is the only possibility, and  $(-1)^{|S|-|S|} = (-1)^0 = 1$ , so we get

$$= u_i^S + \sum_{i \in T \subset S} \left( \sum_{T \subseteq R \subseteq S} (-1)^{|R|-|T|} \right) u_i^T$$

We can reexpress picking a  $R \supseteq T$  as picking an  $R' \subseteq S/T$  (the fragments not in  $T$ ) and unioning it with  $T$ .

$$= u_i^S + \sum_{i \in T \subset S} \left( \sum_{R' \subseteq S/T} (-1)^{|R'|+|T|-|T|} \right) u_i^T$$

$$= u_i^S + \sum_{i \in T \subset S} \left( \sum_{R' \subseteq S/T} (-1)^{|R'|} \right) u_i^T$$

Try writing this as a sum over subset sizes, and you'll get a factorial term showing up from the many possible subsets of a given size.

$$= u_i^S + \sum_{i \in T \subset S} \left( \sum_{b=0}^{b=|S/T|} \left( \frac{(-1)^{b(|S/T|+b)}}{b!} \right)^b \right) u_i^T$$

$$= u_i^S + \sum_{i \in T \subset S} \left( \sum_{b=0}^{b=|S|-|T|} \left( \frac{(-1)^{b(|S|-|T|)}}{(|S|-|T|)! b!} \right)^b \right) u_i^T$$

And then, by plugging this into Wolfram Alpha, we get 0 and all that stuff cancels.

$$= u_i^S, \text{ and so the lemma has been proved.}$$

Onto the full proof, starting off by assuming condition 4 of a coalition-perfect Harsanyi equilibria, and deriving condition 4 of a coalition-perfect CoCo equilibria. Start off with  $a_i u_i^S$ . Then, we use our lemma

$$\text{that } u_i^S = \sum_{i \in R \subseteq S} w_i^R.$$

$$= a_i \sum_{i \in R \subseteq S} w_i^R$$

Distribute the constant in

$$= \sum_{i \in R \subseteq S} a_i w_i^R$$

$$\text{Use that } w_i^R = u_i^R - t_i^R, \text{ by definition of } t_i^R.$$

$$= \sum_{i \in R \subseteq S} a_i (u_i^R - t_i^R)$$

Now, use that, by condition 4 of coalition-perfect Harsanyi equilibria, every player  $j \in R$  has

$$a_j(u_j^R - t_j^R) = a_j(u_j^R - t_j^R), \text{ so we can rewrite this as}$$

$$= \sum_{i \in R \subseteq S} \sum_{j \in R} a_j (u_j^R - t_j^R)$$

$$\text{Unpack what } t_j^R \text{ is}$$

$$= \sum_{i \in R \subseteq S} \sum_{j \in R} a_j (u_j^R - \sum_{j \in T \subseteq R} (-1)^{|R|-|T|+1} u_j^T)$$

Distribute the negative in

$$= \sum_{i \in R \subseteq S} \sum_{j \in R} a_j (u_j^R + \sum_{j \in T \subseteq R} (-1)^{|R|-|T|} u_j^T)$$

Merge it into one big sum

$$= \sum_{i \in R \subseteq S} \sum_{j \in R} a_j \sum_{j \in T \subseteq R} (-1)^{|R|-|T|} u_j^T$$

Reshuffle the sum so we're summing over subsets first, and elements of that subset later. The available subsets  $T$  are all the subsets of  $R$ , and they can only be included in the sum for the  $j$  that lie in  $T$ .

$$= \sum_{i \in R \subseteq S} \sum_{T \subseteq R} \sum_{j \in T} a_j (-1)^{|R|-|T|} u_j^T$$

We can reshuffle the negative 1 part outside of the innermost sum.

$$= \sum_{i \in R \subseteq S} \sum_{T \subseteq R} (-1)^{|R|-|T|} \sum_{j \in T} a_j u_j$$

Abbreviate  $\sum_{j \in T} a_j u_j$  as  $Z^T$ , and reshuffle the sums a little bit.

$$= \sum_{i \in R \subseteq S} \sum_{T \subseteq R} (-1)^{|R|-|T|} Z^T$$

Now, for a given  $T$ , we'll work out the coefficient in front of  $Z^T$  for the entire sum. The first possibility is that  $i \in T$ . Then the possible  $R$  that contribute to the coefficient of  $Z^T$  in the entire sum are exactly the  $R \supseteq T$ . The second possibility is that  $i \notin T$ , so then the possible  $R$  that contribute to the coefficient of  $Z^T$  in the entire sum are exactly the  $R \supseteq T \cup \{i\}$ . So, breaking things up that way, we get

$$= (\sum_{i \in T \subseteq S} \sum_{R \subseteq R \subseteq S} (-1)^{|R|-|T|} Z^T) + (\sum_{i \notin T \subseteq S} \sum_{R \supseteq T \cup \{i\} \subseteq R \subseteq S} (-1)^{|R|-|T|} Z^T)$$

And we can rephrase supersets of  $T$  as subsets of  $S/T$ , unioned with  $T$ . And rephrase supersets of  $T$  that contain  $i$  as subsets of  $S/(T \cup \{i\})$ , unioned with  $T \cup \{i\}$ .

$$\begin{aligned} &= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (-1)^{|R'|+|T|} Z^T) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (-1)^{|R'|+|T|} Z^T) \end{aligned}$$

Reexpress

$$\begin{aligned} &= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (+1)^{|R'|+|T|} Z^T) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (+1)^{|R'|+|T|+1} Z^T) \end{aligned}$$

Cancel

$$\begin{aligned} &= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (+1)^{|R'|} Z^T) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (+1)^{|R'|+1} Z^T) \end{aligned}$$

Split into a sum over the various sizes of what  $R'$  could possibly be

$$\begin{aligned} b &= |S/T| \\ &= (\sum_{i \in T \subseteq S} \sum_{b=0}^{|S/T|} (+1)^b Z^T) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{b=0}^{|S/(T \cup \{i\})|} (+1)^{b+1} Z^T) \end{aligned}$$

Reexpress

$$\begin{aligned}
& b = |S| - |T| \\
= & \left( \sum_{i \in T \subseteq S} \sum_{b=0}^{b=|S|-|T|} \binom{|S|-|T|}{b} (-1)^b Z^T \right) \\
& + \left( \sum_{i \notin T \subseteq S} \sum_{b=0}^{b=|S|-|T|-1} \binom{|S|-|T|-1}{b} (-1)^{b+1} Z^T \right)
\end{aligned}$$

Group into one big fraction.

$$\begin{aligned}
& b = |S| - |T| \\
= & \left( \sum_{i \in T \subseteq S} \left( \sum_{b=0}^{b=|S|-|T|} \binom{|S|-|T|}{b} (-1)^b Z^T \right) \right) \\
& + \left( \sum_{i \notin T \subseteq S} \left( \sum_{b=0}^{b=|S|-|T|-1} \binom{|S|-|T|-1}{b} (-1)^{b+1} Z^T \right) \right)
\end{aligned}$$

Plug it into Wolfram Alpha, and use how the gamma function is defined to get it back into factorial form.

$$= \left( \sum_{i \in T \subseteq S} \frac{(|T|+|S|-1)!}{(|S|-|T|)!} Z^T \right) \Gamma(|T|! \left( \sum_{i \notin T \subseteq S} -|T|! \binom{|S|-|T|}{1} Z^T \right) !)$$

Reindex T to R for notational compliance later on, and we can rewrite this as a single sum over all R that contain i, because they're all paired off with a unique complement that lacks i.

$$= \sum_{i \in R \subseteq S} \frac{(|R|+|S|-1)!}{(|S|-|R|)!} Z^R \Gamma(|S|-|R|!)$$

Figure out what the cardinalities of the various sets are

$$= \sum_{i \in R \subseteq S} \frac{(|R|+|S|-1)!}{(|S|-|R|)!} \frac{(|S|-|R|)!}{(|S|-|R|+1)!} (|S|-|R|-1)!$$

Cancel out, realize that the fractions are the same, and get

$$= \sum_{i \in R \subseteq S} \frac{(|R|+|S|-1)!}{(|S|-|R|)!} Z^R$$

And unpacking how  $Z^R$  was defined, we get, as intended, that this entire chain of equalities has proven

$$a_i u_i = \sum_{i \in R \subseteq S} \frac{R}{(|R|+|S|-1)!} \left( \sum_{j \in R} a_j u_j - \sum_{k \notin R} a_k u_k \right)^{S/R}$$

The exact condition 4 for a coalition-perfect CoCo equilibria, proving that all coalition-perfect Harsanyi equilibria are coalition-perfect CoCo equilibria.

Now it's time for the reverse derivation, showing that all coalition-perfect CoCo equilibria are coalition-perfect Harsanyi equilibria. The goal is to show that for all  $S \subseteq N$ , and  $i, j \in S$ , that

$$a_i(u_i - t_i) = a_j(u_j - t_j)$$

So, let's start out with

$$a_i(u_i - t_i)$$

Substitute in what  $t_i$  is

$$= a_i (u_i - \sum_{i \in R \subseteq S} (-1)^{|S|-|R|+1} u_i^R)$$

Cancel out the negatives

$$= a_i (u_i + \sum_{i \in R \subseteq S} (-1)^{|S|-|R|} u_i^R)$$

Fold it into one big sum

$$= a_i (\sum_{i \in R \subseteq S} (-1)^{|S|-|R|} u_i^R)$$

Multiply the  $a_i$  in

$$= \sum_{i \in R \subseteq S} (-1)^{|S|-|R|} a_i u_i^R$$

Now, we use condition 4 of a coalition-perfect CoCo equilibrium.

$$= \sum_{i \in R \subseteq S} (-1)^{|S|-|R|} \sum_{i \in T \subseteq R} \frac{(-1)^{|R|-|T|}}{|R|!} ((\sum_{j \in T} a_j u_j^T) - \sum_{k \in R \setminus T} a_k u_k^T)$$

Abbreviate the sum  $\sum_{j \in T} a_j u_j^T$  as  $Z^T$ , to get

$$= \sum_{i \in R \subseteq S} (-1)^{|S|-|R|} \sum_{i \in T \subseteq R} \frac{(-1)^{|R|-|T|}}{|R|!} ((Z^T)^{-1} Z^{R/T})$$

Now, we'll have to work out what the coefficient is for a given  $T$ , for the entire sum. Like, what number ends up being in front of  $Z^T$  when we sum everything up? There are two possibilities. The first possibility is that  $i \in T$ . Then the relevant  $R$  that we're summing over are the  $R \supseteq T$ . If  $i \notin T$ , then the relevant  $R$  that we're summing over are the  $R \supseteq T \cup \{i\}$ , and we've got a negative 1 showing up from these  $Z$  terms being subtracted instead of added, which we can fold into the negative 1 power at the start. Using this grouping, we get

$$\begin{aligned} &= (\sum_{i \in T \subseteq S} \sum_{T \subseteq R \subseteq S} (-1)^{|S|-|R|} \frac{(-1)^{|R|-|T|}}{|R|!} (\cancel{Z^T})^{-1}) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{T \cup \{i\} \subseteq R \subseteq S} (-1)^{|S|-|R|+1} \frac{(-1)^{|R|-|T|}}{|R|!} (\cancel{Z^T})^{-1}) \end{aligned}$$

Reexpress it slightly

$$\begin{aligned} &= (\sum_{i \in T \subseteq S} \sum_{T \subseteq R \subseteq S} (-1)^{|S|-|R|} \frac{(-1)^{|R|-|T|}}{|R|!} (\cancel{Z^T})^{-1}) \\ &\quad + (\sum_{i \notin T \subseteq S} \sum_{T \cup \{i\} \subseteq R \subseteq S} (-1)^{|S|-|R|+1} \frac{(-1)^{|R|-|T|+1}}{|R|!} (\cancel{Z^T})^{-1}) \end{aligned}$$

And cancel

$$= (\sum_{i \in T \subseteq S} \sum_{T' \subseteq R \subseteq S} (-1)^{|S|-|R|} \frac{(|R|+|T|)!}{(|R|+|T|)!} Z^{|T|})!$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{T' \subseteq S/(T \cup \{i\}) \subseteq R \subseteq S} (-1)^{|S|-|R|+1} \frac{|T|!}{(|R|-|T|)!} Z^{|T|})!$$

And we can reexpress this as picking a subset of  $S/T$ , and unioning it with  $T$  to make  $R$ , or as picking a subset of  $S/(T \cup \{i\})$  and unioning it with  $T \cup \{i\}$  to make  $R$ .

$$= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (-1)^{|S|-|R'|} \frac{(|R'|+|T|)!}{(|R'|+|T|)!} Z^{|T|}-1)!$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (-1)^{|S|-|R'|+1} \frac{|T|!}{(|R'|+|T|)!} Z^{|T|}-1)!$$

Reexpress

$$= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (-1)^{|S|-|R'|-|T|} \frac{(|R'|+|T|+1)!}{(|R'|+|T|+1)!} Z^{|T|}-1)!$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (-1)^{|S|-|R'|-|T|-1+1} \frac{|T|!}{(|R'|+|T|+1)!} Z^{|T|}-1)!$$

Cancel

$$= (\sum_{i \in T \subseteq S} \sum_{R' \subseteq S/T} (-1)^{|S|-|R'|-|T|} \frac{|R'|!}{(|R'|+|T|)!} Z^{|T|})!$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{R' \subseteq S/(T \cup \{i\})} (-1)^{|S|-|R'|-|T|} \frac{|T|!}{(|R'|+|T|)!} Z^{|T|})!$$

Reexpress as summing up over all possible sizes for  $R'$ , introducing a factorial term because of the many subsets of a given size.

$$b = |S/T|$$

$$= (\sum_{i \in T \subseteq S} \sum_{b=0}^{\lfloor |S/T| \rfloor} \binom{|S/T|+b}{b} b! (-1)^{|S|-b-|T|} \frac{b!}{b+|T|} Z^{|T|})$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{b=0}^{\lfloor |S/(T \cup \{i\})| \rfloor} \binom{|S/(T \cup \{i\})|+b}{b} (-1)^{|S|-b-|T|} \frac{b!}{b+|T|} Z^{|T|})$$

Reexpress

$$b = |S| - |T|$$

$$= (\sum_{i \in T \subseteq S} \sum_{b=0}^{\lfloor |S|-|T| \rfloor} \binom{|S|-|T|+b}{b} (-1)^{|S|-b-|T|} \frac{b!}{b+|T|} Z^{|T|})$$

$$+ (\sum_{i \notin T \subseteq S} \sum_{b=0}^{\lfloor |S|-|T|-1 \rfloor} \binom{|S|-|T|-1+b}{b} (-1)^{|S|-b-|T|} \frac{b!}{b+|T|} Z^{|T|})$$

Merge into one big fraction and cancel

$$\begin{aligned}
& b = |S| - |T| \\
&= \left( \sum_{i \in T \subseteq S} \sum_{b=0}^{|S|-|T|} \frac{(-1)^{|S|-|T|-b}}{(|S|-|T|-b)!((|T|-1)!)^{(|S|-|T|-b)}} \right) \\
&\quad + \left( \sum_{i \notin T \subseteq S} \sum_{b=0}^{|S|-|T|-1} \frac{(-1)^{|S|-|T|-b-1}}{(|S|-|T|-b-1)!((|T|-1)!)^{(|S|-|T|-b-1)}} \right)
\end{aligned}$$

And plug into Wolfram Alpha to get that both of these alternating sums over factorials are actually the same coefficient, so we can just write it as

$$= \sum_{T \subseteq S} \frac{(-1)^{|S|-|T|}}{|S|!} Z^{|T|}$$

Summing all this up, we've derived

$$a_i (u_i - t_i) = \sum_{T \subseteq S} \frac{(-1)^{|S|-|T|}}{|S|!} Z^{|T|}$$

And then we can do this whole line of reasoning again but swapping out  $i$  for  $j$ , and nothing at all changes, we still get the same quantity at the end, so we have

$$a_i (u_i - t_i) = a_j (u_j - t_j)$$

For all  $i, j \in S \subseteq N$ , the fourth condition for a coalition-perfect Harsanyi equilibrium, so all coalition-perfect CoCo equilibria are coalition-perfect Harsanyi equilibria.

Since we've proved both directions, something is a coalition-perfect Harsanyi equilibria iff it's a coalition-perfect CoCo equilibria.

# Principles for Alignment/Agency Projects

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

"John, what do you think of this idea for an alignment research project?"

I get questions like that fairly regularly. How do I go about answering? What principles guide my evaluation? Not all of my intuitions for what makes a project valuable can easily be made legible, but I think the principles in this post capture about 80% of the value.

## Tackle the Hamming Problems, Don't Avoid Them

Far and away the most common failure mode among self-identifying alignment researchers is to look for Clever Ways To Avoid Doing Hard Things (or Clever Reasons To Ignore The Hard Things), rather than just Directly Tackling The Hard Things.

The most common pattern along these lines is to propose outsourcing the Hard Parts to some future AI, and "just" try to align that AI without understanding the Hard Parts of alignment ourselves. The next most common pattern is to argue that, since Hard Parts are Hard, we definitely don't have enough time to solve them and should therefore pretend that we're going to solve alignment while ignoring them. Third most common is to go into field building, in hopes of getting someone else to solve the Hard Parts. (Admittedly these are not the most charitable summaries.)

There is value in seeing how dumb ideas fail. Most of that value is figuring out what the Hard Parts of the problem are - the taut constraints which we run into over and over again, which we have no idea how to solve. (If it seems pretty solvable, it's probably not a Hard Part.) Once you can recognize the Hard Parts well enough to try to avoid them, you're already past the point where trying dumb ideas has much value.

On a sufficiently new problem, there is also value in checking dumb ideas just in case the problem happens to be easy. Alignment is already past that point; it's not easy.

You can save yourself several years of time and effort by actively trying to identify the Hard Parts and focus on them, rather than avoid them. Otherwise, you'll end up burning several years on ideas which don't actually leave the field better off. That's one of the big problems with trying to circumvent the Hard Parts: when the circumvention inevitably fails, we are still no closer to solving the Hard Parts. (It has been [observed](#) both that alignment researchers mostly seem to not be tackling the Hard Parts, and that alignment research mostly doesn't seem to build on itself; I claim that the latter is a result of the former.)

Mostly, I think the hard parts are things like "understand agency in general better" and "understand what's going on inside the magic black boxes". If your response to such things is "sounds hard, man", then you have successfully identified (some of) the Hard Parts.

## Have An Intuitive Story Of What We're Looking For

One project going right now is looking at how modularity in trained systems corresponds to broad peaks in parameter space. Intuitive story for that: we have two "modules", each with lots of stuff going on inside, but only a relatively-low-dimensional interface between them. Because each module has lots of stuff going on inside, but only a low-dimensional interface, there should be many ways to change around the insides of a module while keeping the externally-visible behavior the same. Because such changes don't change behavior, they don't change system performance. So, we expect that modularity implies lots of degrees-of-freedom in parameter space, i.e. broad peaks.

This story is way too abstract to be able to look for immediately in a trained net. How do we operationalize "modules", and find them? How do we operationalize "changes in a module", especially since parameter space may not line up very neatly with functional modules? But that's fine; the story can be pretty abstract.

The point of the intuitive story is to steer our search. Without it, we risk blind empiricism: just cataloguing patterns without building general models/theory/understanding for what's going on. In that mode, we can easily lose track of the big picture goal and end up cataloguing lots of useless stuff. An intuitive story gives us big-picture direction, and something to aim for. Even if it turns out to be wrong!

## Operationalize

It's relatively easy to make vague/abstract intuitive arguments. Most of the value and challenge is in finding the right operationalizations of the vague concepts involved in those arguments, such that the argument is robustly correct and useful. Because it's where most of the value and most of the challenge is, finding the right operationalization should typically be *the central focus of a project*.

My [abstraction work](#) is a good example here. I started with some examples of abstraction and an intuitive story about throwing away information while keeping info relevant "far away". Then, the bulk of the work was to operationalize that idea in a way which matched all the intuitive examples, and made the intuitive stories provable.

## Derive the Ontology, Don't Assume It

In ML interpretability, some methods look at the computation graph of the net. Others look at orthogonal directions in activation space. Others look at low-rank decompositions of the weight matrices. These are all "different ontologies" for interpretation. Methods which look at one of these ontologies will typically miss structure in the others; e.g. if run a graph clustering algorithm on the computation graph I probably won't pick up interpretable concepts embedded in directions in activation space.

What we'd really like is to avoid assuming an ontology, and rather *discover/derive* the ontology itself as part of our project. For instance, we could run an experiment where we change one human-interpretable "thing" in the environment, and then look at how that changes the trained net; that would let us *discover* how the concept is embedded rather than assume it from the start (credit to Chu for this suggestion). Another approach is to start out with some intuitive story for *why* a particular ontology is favored - e.g. if we have a graph with local connectivity, then maybe the [Telephone Theorem](#) kicks in. Such an argument should (a) allow us to *rule out* interactions which circumvent the favored ontology, and (b) be testable in its own right, e.g. for the Telephone Theorem we can (in principle) check the convergence of mutual information to a limit.

## Open The Black Box

Don't just run a black-box experiment on a network, or try to prove a purely behavioral theorem. We want to talk about internal structure.

Partly, opening the black box is about tackling the Hard Parts rather than avoiding them. Not opening the black box is a red flag; it's usually a sign of avoiding the Hard Parts.

Partly, opening the black box is about getting a very rich data channel. When we just work with a black box, we get relatively sparse data about what's going on. When we open the black box, we can in-principle directly observe every gear and directly check what's going on.

## Relative Importance of These Principles

Tackle The Hamming Problems is probably the advice which is most important to follow for marginal researchers right now, but mostly I expect people who aren't already convinced of it will need to learn it the hard way. (I certainly had to learn it the hard way, though I did that before starting to work on alignment.) Open the Black Box follows pretty naturally once you're leaning in to the Hard Parts.

Once you're past that stumbling block, I think the most important principles are Derive the Ontology and Operationalize. These two are important for opposing types of people. Some people tend to stay too abstract and avoid committing to an ontology, but never operationalize and therefore miss out on the main value-add. Other people operationalize prematurely, adopting [ad-hoc operationalizations](#), and Deriving the Ontology pretty strongly discourages that.

Have an Intuitive Story is especially helpful for people who tend to get lost in the weeds and go nowhere. Make sure you have an intuitive story, and use that story to guide everything else.

# Criticism of EA Criticism Contest

Back when it was announced, I [toyed with the idea](#) of criticizing [the EA criticism and red teaming contest](#) as an entry to that contest, leading to some very good Twitter discussion [that I compiled into a post](#).

I finally found myself motivated to write up my thinking in detail.

Before I begin, I recommend reading [the contest announcement](#) yourself, and forming your own reaction. Ask yourself, among other things:

1. What are they telling people they are interested in and likely to reward?
2. What are they not interested in and telling people they will not reward?
3. What does this announcement say more generally about EA?

## Praise First

I will start with some things that I think are good about the contest and announcement.

1. It exists at all. It is good to solicit (paid!) public criticism.
2. It awards at least \$100,000. That's good prize money.
3. It promises an even bigger prize for something that really hits home.
4. It suggests it will reward causing people to change their minds.
5. [Guidelines offered to help with techniques includes red teaming](#).
6. The exercise I read this as being is still worth doing, if not as worthwhile as the thing I'd rather be happening, so long as it doesn't create a story that the thing I want to happen did indeed happen and is thus handled, when it isn't handled.
7. What *actually* wins will be an informative experiment. [Stay tuned](#).

## Core Critique

Effective Altruism has a core set of assumptions, and a core method of modeling and acting in the world.

The core critique I am offering is that the contest is mostly taking things like the list below as givens, rather than as things to be questioned.

It is sending a set of signals that such punches should be pulled, both for the contest and in general. Effective Altruism, in this model, very much wants criticism of its *tactics*, and mostly wants them also of its *strategy*, but *only within the framework*.

This parallels EA more broadly, where within-paradigm critiques are welcomed (although, as everywhere, often incorrect disregarded) but deeper critiques are unwelcome, and treated as mistakes.

Here is my attempt to summarize the framework, which came out to 21 points.

Note that the rest of the post does not depend on the exact points or their wordings. Although I did attempt to make the list match my perceptions as much as possible, the list is included primarily to give an idea of the *type of* thing that is assumed here.

1. Utilitarianism. Alternatives are considered at best to be mistakes.
2. Importance of Suffering. Suffering is The Bad. Happiness/pleasure is The Good.
3. Quantification. Emphasis on that which can be seen and measured.
4. Bureaucracy. Distribution of funds via organizational grants and applications.
5. Scope Sensitivity. Shut up and multiply, two are twice as good as one.
6. Intentionality. You should to plan your life around the impact it will have.
7. Effectiveness. Do what works. The goal is to cut the enemy.
8. Altruism. The best way to do good yourself is to act selflessly to do good.
9. Obligation. We owe the future quite a lot, arguably everything.
10. Coordination. Working together is more effective than cultivating competition.
11. Selflessness. You shouldn't value yourself, locals or family more than others.
12. Self-Recommending. Belief in the movement and methods themselves.
13. Evangelicalism. Belief that it is good to convert others and add resources to EA.
14. Reputation. EA should optimize largely for EA's reputation.
15. Modesty. Non-neglected topics can be safely ignored, often consensus trusted.
16. Existential Risk. Wiping out all value in the universe is really, really bad.
17. Sacrifice. Important to set a good example, and to not waste resources.
18. Judgment. Not living up to this list is morally bad. Also sort of like murder.
19. Veganism. If you are not vegan many EAs treat you as non-serious (or even evil).
20. Grace. In practice people can't live up to this list fully and that's acceptable.
21. Totalization. Things outside the framework are considered to have no value.

There are also things that *follow from* these points.

And importantly, there are also *things one is not socially allowed to question or consider*, not in EA in particular but fully broadly. Some key considerations are things that cannot be said on the internet, and some general assumptions that cannot be questioned are importantly wrong but cannot be questioned. This is a *very hard* problem but is especially worrisome when calculations and legibility are required, as this directly clashes with there being norms against making certain things legible.

Where I agree with the list above, I am a rather large fan. In particular, #7 is insanely great, #5 can be taken too far but almost always isn't taken far enough, and #16 is super important. And if you're going to do the rest of this, you really need #20.

Yet when I think about the remaining 17 points above, I notice that I strongly disagree with a majority of them (#1, #2, #4, #8, #10, #11, #13, #14, #15, #17, #18, #19, #21), and disagree on magnitude or practical usefulness with the rest (#3, #6, #9, #12), where I would say something like 'yes, more of this than others think, less than you think.'

That many disagreements strongly implies the list is not doing a good job cutting reality at its joints, and a shorter list is possible - that many items here are more examples than they are core things. I am confident that starts at the top with #1. That all seems right and important, but a harder topic for another day. I don't know how to write at that level in a way that I feel permission to do so, or the expectation of being able to do so effectively. Which makes me assume most others feel the same way.

The same goes for disagreement with things that follow from the assumptions, or that can only be challenged by challenging societal assumptions where there are norms against being seen challenging them.

One could say [as Helen does here](#) that EA is merely *asking the question* of "how can I do the most good, with the resources available to me?" or alternatively "[with the](#)

[resources I choose to give](#)" although the contradictions and implications on that never sit well any more than they do with almost every other value system. To some extent sure, the details are up for grabs, but a lot of the answers on things like the items above aren't considered all that up for grabs, and EA has an implied model and definition that goes along with 'do the most good' being a coherent concept, which packs a lot of logical punch.

Rather than being said too explicitly, although there is some of that as well, the call to work within the paradigm and pull punches comes centrally in the form of a *vibe*. Things come together to communicate the message implicitly, even unconsciously. I do not think that those who wrote the contest were doing this deliberately. Rather, it is a property of the systems that produce such things, that happens if not fought against.

The only way I know of to explain is to do so via a close reading.

## A Close Reading of the Criteria

The section most worth zeroing in on is the criteria. Note that at various points someone could rightfully say 'a literal interpretation of these words does not say the thing you are claiming they are saying' and my response to that is 'yes, you are right, as a matter of logic, but that is not the way that words like this are interpreted, nor is it the way they are intended to be interpreted at the level that matters most.'

Here is their guideline for what they are looking for.

Overall, we want to reward critical work according to a question like: "**to what extent did this cause me to change my mind about something important?**" — where "change my mind" can mean "change my best guess about whether some claim is true", or just "become significantly more or less confident in this important thing."

Below are some virtues of the kind of work we expect to be most valuable. We'll look out for these features in the judging process, but we're aware it can be difficult or impossible to live up to all of them.

**Critical.** The piece takes a critical or questioning stance towards some aspect of EA theory or practice. Note that this does **not** mean that your conclusion must end up disagreeing with what you are criticizing; it is entirely possible to approach some work critically, check the sources, note some potential weaknesses, and conclude that the original was broadly correct.

**Important.** The issues discussed really matter for our ability to do the most good as a movement.

**Constructive and action-relevant.** Where possible we would be most interested in arguments that recommend some specific, realistic action or change of belief. It's fine to just point out where something is going wrong; even better to be constructive, by suggesting a concrete improvement.

**Transparent and legible.** We encourage [transparency about your process](#): how much expertise do you have? How confident are you about the claims you're making? What would change your mind? If your work includes data, how were they collected?

Relatedly, we encourage **[epistemic legibility](#)**: the property of being *easy to argue with*, separate from being correct.

**Aware.** Take some time to check that you're not missing an existing response to your argument. If responses do exist, mention (or engage with) them.

**Novel.** The piece presents new arguments, or otherwise presents familiar ideas in a new way. Novelty is great but not always necessary — it's often still valuable to [distill](#) or "translate" existing criticisms.

**Focused.** Critical work is often (but not always) most useful when it is focused on a small number of arguments and a small number of objects. We'd love to see (and we're likely to reward) work that engages with specific texts, strategic choices, or claims.

Once again I am going to ask you to stop and read through yourself, and form your own interpretation of what this is telling you.

Listen to those little voices inside your head.

We'll start at the top. They don't use italics, so I will use them to highlight particular details.

## Want

Overall, we *want* to reward critical work according to a question like: "to what extent did this cause me to change my mind about something important?"

That's a great question. Love the question.

What is the word 'want' doing here? It is placing an invisible yet unmistakable-once-you-notice-it 'but' in that sentence. It is often said, for good reason, that anything before the 'but' does not count.

Then later they say they will 'look out for these virtues' in the judging process:

We'll look out for these features in the judging process, but we're aware it can be difficult or impossible to live up to all of them.

Why? Why would we look out for these virtues?

If the thing to be rewarded is the extent to which minds were changed about something important, *you can simply ask yourself that question*.

I agree that having more of the properties listed will make it more likely that a given effort will change one's mind about something important. It is easy to see why being critical, important, aware, novel, focused, transparent and legible all help the cause here, with constructive and action-relevant being the notable exception.

That doesn't explain why you would *deliberately Goodhart* yourself here. It doesn't explain why they would be part of the *judging criteria* rather than being merely things to consider when composing an entry.

Note the statement that it can be difficult or impossible 'to live up to' all of them, which emphasizes that absolutely we will dock your grade for anything you miss, on

top of any effect it has on your ability to change our mind.

It's the difference between a teacher saying each of these two things to a class:

Tomorrow's report on population ethics must be double-spaced, with 10-point Arial font, and 4-5 pages long and 1-inch margins, or I'm docking your grade.

Tomorrow's report on population ethics should succinctly explain and justify your position, without going into too much detail, as we've discussed. It probably wants to be about 4-5 pages long.

The second teacher wants something useful, a thought out and justified view on population ethics that doesn't get too lost in the weeds.

The first teacher is more focused on something like ensuring that the student write the correct number of characters in the proper format, to ensure everyone knows how to check all the boxes correctly and obey arbitrary rules, and to ensure no one 'cheats' by doing less work.

Or, alternatively, the first teacher is in a school culture where they know the students are utterly uninterested in producing a useful thing and can only be forced to do anything by threat of punishment, and that punishment requires violation of the rules, to the rules need to be exactly specified or they'll get back papers in 18-inch (or for that one weirdo 8-inch) fonts with improper spacing and huge margins so the kids can get back to TikTok or whatever they're doing these days.

This note from the FAQ confirms we are dealing with the first teacher (note the italicized term *requirements*):

**Does my submission need to fulfill all the criteria outlined above?** No. We understand that some formats make it difficult or impossible to satisfy all the *requirements*, and we don't want that to be a barrier to submitting. At the same time, we do think each of the criteria are good indicators of the kind of work we'd like to see.

So to bring it back consider the following three sets of instructions.

We will judge entries based on the question: "To what extent did this cause me to change my mind about something important?" Here are some virtues we expect to best help people to change their minds about important things, but you will be judged only on whether our minds were changed.

We *want* to judge entries based on the question: "To what extent did this cause me to change my mind about something important?" Here are some virtues we expect to find most valuable, so we will look out for them during the judging process.

We *want* to judge entries based on the question above. We will instead judge the contest largely on other things, including the following virtues.

My claim here is that #1 is very different from #2, but that #2 and #3 are very similar.

My additional claim is that #1 is a much better thing to be doing.

There is no need to *look out* for any virtues here. The whole idea is that those virtues/features are *helpful in cutting the enemy*. So look at the enemy. Has it been cut?

There are times and places where you want the first teacher (or something in between them) rather than the first one. This does not seem like it is one of those places.

## Just

**"to what extent did this cause me to change my mind about something important?"** — where "change my mind" can mean "change my best guess about whether some claim is true", or just "become significantly more or less confident in this important thing."

It took me a bit to *notice consciously* what I had automatically noticed and flinched from here unconsciously, like a [Fnord](#) or a bit of alchemy.

The sentence starts off with a great question: "To what extent did this cause me to change my mind about something important?"

It then clarifies "change my mind" to mean "X or just Y," where Y ends with 'in this important thing.'

The brain instinctively knows three things here.

That the word 'just' in this context means something lesser or minimal. It satisfies the criteria, but to an importantly lesser degree.

That 'something important' has become optional.

Confidence levels are less important than best guesses.

To see the second one, notice that one of the two formulations, *the lesser one*, ends with 'this important thing.' Then move that back into the original sentence - "... become significantly more or less confident in this important thing about something important."

That's basically [Paris In The Springtime](#). The middle instance drops out.

Thus, the message I heard was:

We are looking for you to change our view of a [specific] claim.

Changing our confidence level is not as good.

Importance of claim is nice to have, but optional.

## Critical

**Critical.** The piece takes a critical or questioning stance towards some aspect of EA theory or practice. Note that this does **not** mean that your conclusion must end up disagreeing with what you are criticizing; it is entirely possible to approach some work critically, check the sources, note some potential weaknesses, and conclude that the original was broadly correct.

If you are looking for EA criticism, it makes sense to want it to be critical.

I can't help but notice that this is actually the opposite request. It is saying that you *do not need to be critical* in order to count as critical.

You can 'approach some work 'critically', check the sources, note some potential weaknesses, and conclude that the original was broadly correct.'

That is of course a valuable thing to do. If the original is broadly correct, I want to conclude that the original is broadly correct.

There is also the danger that if you only reward 'criticism' with money, and someone sets out to do an examination, that they will be biased towards being unfairly critical and negative, worried that it otherwise wouldn't count.

The contest authors seem clearly worried about this, as an extension of the worry that people considering entering might need some assurance that they will get compensation, and giving them the ability to request funding in advance. I notice I am suspicious of a contest working this way, but it is not *obviously* wrong or crazy.

As written, this instead sends the opposite message. It is saying:

You need to take a critical *stance* towards *something*, some 'aspect of EA theory or practice.'

That does not mean you need to actually be criticizing it.

Noticing some potential weaknesses is enough.

We are looking to reward things that appear to be critical assessments, but that conclude broad agreement with the thing being criticized.

Nowhere does this say explicitly that it would be preferred if you *didn't* substantially criticize the target, and that the goal is to allow everyone involved to *tell a story that criticism was solicited and received* and that the 'changing of one's mind' is supposed to be an affirmation of existing stories.

But is that the vibe here? The implicit message? Absolutely. Three quarters of the virtue of criticism is about how we meant critical but we didn't mean criticize or disagree.

It is, again, totally fine to commission examinations of existing practice with an eye towards catching mistakes and being more confident in conclusions rather than looking for the strongest challenges. I would not call that a 'criticism and red teaming' task.

I realize, again, that covering all one's bases in various directions is a challenge. What would I have written here instead? I would have written this:

**Critical.** The piece examines, criticizes and challenges an important EA theory or practice.

Thus, I would have accepted the 'if all you do is affirm what we were already doing you are not embodying this virtue' downside, because *in this context* it is a downside, it is less likely to importantly change one's mind and the contest should reflect that while remaining open to having one's mind changed anyway. When you run a contest,

it needs to be expected that sometimes an attempted entry will turn out to be barking up the wrong tree - if the investigation doesn't change our minds, but we want to pay anyway, then we are rewarding something other than what we 'want' to reward. We're at minimum rewarding *displays of effort*, and potentially rewarding storytelling. That's more like commissioning work. If you want to commission work, commission work.

If I was sufficiently worried about this I might extend this to:

**Critical.** The piece examines, criticizes and challenges an important EA theory or practice. This does not mean it needs to claim the target is centrally bad and wrong, but it must point to way in which minds should change.

## Important

**Important.** The issues discussed really matter for our ability to do the most good as a movement.

Important things are more important than less important things. Yes, strongly agreed that it is more valuable to change one's mind about important things. I do notice that the second clause is suspicious, as opposed to saying:

**Important.** The issues discussed really matter.

The first version cares only about whether it matters *for our ability to do good*. And beyond that, for our ability to do good *as a movement*.

A philosophy question. If something is important for reasons *other than* the ability to do good, is it still important?

I would argue strongly yes. The trivial argument is that if can matter for *our ability to avoid doing bad*, or *our cost of doing good*, or various higher-level effects, or other such caveats. These buy the basic premise that the thing that is important is *doing good* on some scale, with minor extensions.

I do still think that distinction matters. Noticing how do go good 'cheaper' or avoid harms is often more (intractable/neglected/important) in a situation, and many failed collective efforts (especially government ones) were rushes to 'do good' without in this sense thinking it all through.

The more *fundamental* challenge is whether *doing good* is indeed The Good. Can you learn something, have that not impact your ability to do good (including the extensions above) but have that still be *important*? Could you still be correct to invest time or other resources into learning that?

My answer again is strongly yes, and also that if we answer 'no' here our ability to do good is substantially reduced to the extent I question whether someone who sets out to do good while answering 'no' should be expected to on net do good.

There is a very deep disagreement this is pointing to where I do not think 'do the most good' is the right way to think about how to achieve The Good, which encompasses many core disagreements along the way.

Discussing properly would alas be beyond scope here. What I will note is that this reinforces the message that importance means on the basis of importance to doing

things that are classified by the EA structures as ‘doing good’ and that can be quantifiably linked to specific good done, steering once again into narrow bounds.

## Movement

Then we come to the end, ‘as a movement.’

I care about achieving The Good. I care about doing good. What I *absolutely do not care about* is whether this good is done or achieved *by EA*. I want the good because it’s good. If something is important to helping *others* do more good, or suggesting people can do importantly more good outside of EA, then that is important information, and information I actively want to find.

## Constructive

**Constructive and action-relevant.** Where possible we would be most interested in arguments that recommend some specific, realistic action or change of belief. It’s fine to just point out where something is going wrong; even better to be constructive, by suggesting a concrete improvement.

This is the odd one out. The other *descriptions* are worrisome as is grading based on them, but they do seem clearly to be virtues in the desired sense.

Here we have something else. This is a commonly used strategy to *prevent* criticism.

One of the biggest reasons errors go unnoticed and uncorrected is ‘don’t tell [the boss/everyone/whoever] about the problem if you don’t have a solution.’

The attempt to say that the action can merely be ‘change of belief’ indicates that there is awareness of this issue, as is ‘it’s fine to just point out where something is going wrong’ but that is exactly what is *not* meant by ‘constructive’ or especially ‘action relevant.’

In other words, it is saying these things are ‘fine’ but in the sense that we will uphold the rhetorical claim that it is fine while making it very clear implicitly that this is totally, absolutely not fine and will be treated accordingly. If you can’t figure this out, or choose not to be a team player on this, that’s on you.

The request that the proposal be ‘specific and realistic’ and later ‘concrete’ is very clear that a change of belief that doesn’t cash out in a particular new action is insufficient.

There are several clear indications that ‘what you are doing does not work, halt and catch fire’ is not what is being requested here, not in practice. They want ‘what you are doing does not work, so fix it with this One Weird Trick’ or ‘what you are doing does not work, you should do this [within paradigm] alternative instead.’

The counterargument is that *of course* it is beneficial to be constructive rather than non-constructive. It is *better* to offer new information about what is failing and also a worthwhile path forward than to offer the same new information about what is failing and not to offer a path forward. And that’s certainly true. But when important problems are noticed, and important criticisms or red teaming is offered, gating the problem announcement on a solution kills the whole process.

The *whole point* of a red teaming is to find where you are likely to fail. It is a *different job and task* to then fix the problem.

I hear a clear message here, and that message is: One must be a team player. You must provide a story that we can continue to mostly execute our strategies and do the most good without a halt and catch fire or a paradigm shift. If you do not do this, we will go all [Copenhagen Interpretation of Ethics](#) on you, and hold you responsible for breaking the narrative compact.

Once again, a good question is, what would I have written here, to make the core point that it is a plus to point out solutions and suggest useful actions?

This one is tough. If I had to keep the ‘you will be judged on these bullet points’ language I think I delete this entirely. Otherwise, maybe something like this?

**Explicit.** It is best if you can explicitly indicate clear and important belief updates that should be made. That makes it more likely to importantly change minds. It is also helpful if you can explicitly indicate how these belief updates should change behavior, including where it makes sense to halt and catch fire.

Even this makes me nervous. Perhaps I am typical minding here, and people (especially in our circles) would if left to their own devices be overly neglectful of these stages and thus can use pushes in that direction. I doubt this, because I think most people fully recognize that everyone loves it when you not only point out a problem but also come with a solution.

When I point out a problem, I certainly *attempt* to find a solution or path forward, or at least ways to look for one, to suggest them as well. That is a normal, friendly, highly useful course of action. The key is not to let the lack of it stop you from speaking the truth, or make your voice tremble.

If someone does both point out a big problem and also comes up with a good solution, then yes, that is even better, that’s two changes of mind in one and should be rewarded accordingly. The problem is that by default this mostly causes silencing.

## Legible

**Transparent and legible.** We encourage [transparency about your process](#): how much expertise do you have? How confident are you about the claims you’re making? What would change your mind? If your work includes data, how were they collected? Relatedly, we encourage [epistemic legibility](#): the property of being *easy to argue with*, separate from being correct.

The top link here goes to an interesting document that leads off like this:

In short, our top recommendations are to:

Open with a linked summary of key takeaways. [[more](#)]

Throughout a document, indicate which considerations are most important to your key takeaways. [[more](#)]

Throughout a document, indicate how confident you are in major claims, and what support you have for them. [[more](#)]

This is vastly better practice than the vast majority of alternatives. More people doing it would mostly be great, given what it would be replacing.

I am especially a big fan of indicating confidence in one's claims, and would extend this to minor claims. It's worth it.

The top reasons I often *don't* quite do this are, with some overlap:

**Anchoring.** If you know the takeaways before seeing the evidence, you will be more inclined to draw the same conclusions and think in the same ways.

**Passivity.** Telling people the takeaways and key points in advance acts as a blocker that makes it harder for them to think about the problem, either the way you did or their own way, and go into a kind of 'check your work' mode.

**Learning to Think.** I consider the secondary goal of essentially everything I write to be sharing my *methods of thinking* so others can improve theirs, and I can also improve mine. I realize there is a time and a place *not* to do this.

**Discarding the Illegible.** I consider this an especially big problem in EA but it is a problem in general. When you list legible considerations, and start assigning numbers, what happens to considerations that aren't as legible? Even if you can spell out your worry, often it gets reduced to *the part whose impact you can quantify via a lower bound*. Thus, some considerations have clearly-much-too-low estimates or get discarded, others have reasonable estimates, and they get compared.

**The Law of One Reason.** If you give someone five good reasons for something, the default is for the brain to mostly discard four of them and compare the remaining reason to the one reason on the other side. Thus, it is often stronger in practice to drop four of the five reasons, and only state your strongest case. It can also be beneficial to trick the other side into providing lots of additional reasons, even real ones, while you hammer on your talking points that work. Indicating what are your most important reasons risks making this problem worse by dropping everything not on the list, and it also risks invoking the problem by having most of the list be discarded. There's also the reverse part, where if they find one reason they *disagree* with people often then discard the whole thing whether or not that makes sense.

**Key Takeaways Crowd Out.** There is the argument that you only get five words, so better to choose what five words readers will remember and make it easy to have them stick. There is something to that. I still choose to hope for better. When you categorize your key takeaways, then talk about what is important to your key takeaways, what chance do any other potential takeaways have? Especially becoming stronger more generally, or unexpected things.

The second link basically explains that if one can understand what your claims are saying, quantify them, check them against other sources, figure out if they're right and whether they justify the thing they are claiming they justify, that's all helpful.

I'd agree with that, in something like this form: If a claim can be made more legible, that is a good thing to do, and you probably should do that. Needlessly vague statements are less useful on many levels.

The problem comes when you *can't* easily make the claim more legible. Your choices are often to either (A) not make the claim at all, (B) make a similar but importantly different claim so it will be legible, (C) write endlessly trying to make the thing legible

via giving people tacit knowledge or (D) say the thing illegibly and hope at least some people get it, or that at least some readers can grok it through context later on. Often it is helpful here to *point out* you are saying something illegible.

I think about this problem a lot. There are a lot of things I want to say that I *don't know how* to say in a way that would sound not-crazy and not get dismissed, or would make sense to anyone who didn't already know or mostly know. Often this is because I don't yet understand it well enough. Other times it is because there are lots of load bearing things holding up the thing, or explaining why the thing isn't crazy, and getting through those seems impossible or at least has to happen first. Then there are times when some of the load bearing things are *things I can't say out loud on the internet* so then what do you do?

Often communicating such things through in-person conversation is possible where written communication is not, because you can figure out which metaphors resonate and click, and which parts of your model they disagree with and have to be explained or justified in which ways. Often I have good hope of *tutoring* on a question where I would be hopeless at *teaching a class*. Often that is because it is insufficiently legible.

Mostly all of this is quibbling and nitpicking, since almost everyone should move in the direction being suggested here, and I could probably do to work harder at this as well. The problem comes when you double-count this, as in both explicitly checking for legibility and also then to see if understanding follows.

I do worry about explicitly calling for people to say how much expertise they have in this way. I would expect a chilling effect on those who do not have sufficiently legible expertise in an area, making them doubt themselves and expect not to be listened to on the merits, and worried there will be reliance on argument from authority. One should of course still point out where and when one does and doesn't have expertise of various kinds, especially when relying upon it to make claims, but mostly the words should stand on their own.

How would I have worded this one?

**Transparent and legible.** We encourage [transparency about your process](#). How did you reason things out and reach your conclusions? How confident are you about the claims you're making? What would change your mind? If your work includes data, how were they collected? Relatedly, we encourage [epistemic legibility](#) to the extent possible: Make your claims and reasoning as easy to pin down and evaluate and find cruxes with as you can, but no more than that.

While epistemic legibility is distinct from correctness, I would aim to avoid ending a point one is being judged on with 'separate from being correct' since that gives the impression correctness is not so valued.

## Aware

**Aware.** Take some time to check that you're not missing an existing response to your argument. If responses do exist, mention (or engage with) them.

Is it a *good* argument?

It is virtuous to respond to all good arguments against your position *whether or not anyone has made them before*.

Looking for existing arguments has several justifications.

It may be more efficient than figuring out the counterarguments yourself.

You might find good arguments.

You might be able to preempt bad arguments and save everyone time.

You might be blamed for not addressing existing arguments.

Whenever I write anything, there's always a voice that says: Thousands of people are going to read this. You are one person. Shouldn't you put in more work?

The answer is, sometimes, but if you take that too far nothing gets written. The point is to follow procedures and have virtues that actually lead to providing value to people.

To me, the idea of 'responding to arguments' misses the point entirely. Docking points for not pointing out existing counter-arguments, while *not* telling people to be on the lookout for arguments against your position that haven't been suggested (which is especially important if it is, as the next point suggests, novel) means the goal isn't to help people have all they need to form their own opinions and seek truth. Instead, you're checking social boxes, and you're doing adversarial advocacy rather than trying to figure things out.

Thus, I'd move this one around.

**Balanced.** Seek out potential reasons you might be wrong. Where they have merit, at least note them, and ideally address them. If people disagree with you, it is worth understanding why and considering addressing that.

## Novel

**Novel.** The piece presents new arguments, or otherwise presents familiar ideas in a new way. Novelty is great but not always necessary — it's often still valuable to [distill](#) or "translate" existing criticisms.

This one seems fine, although I am unsure it is necessary, and I worry about awarding points for it directly. This is an example of good wording.

## Focused

**Focused.** Critical work is often (but not always) most useful when it is focused on a small number of arguments and a small number of objects. We'd love to see (and we're likely to reward) work that engages with specific texts, strategic choices, or claims.

I strongly agree with this sentiment *provided it then generalizes*. I am unsure that focused is the correct name for the thing (Grounded? Detail Oriented or Detailed? Specific? Not sure anything else is better), but working with examples and their details is usually the right way to go. Then readers can draw the larger point and generalize from there. Otherwise, you often say a bunch of vague things and don't have a good example even if asked, and the whole thing feels abstract and gets ignored, often with good reason.

The flip side is that the *reason* to do a detailed report that engages with specific objects can be either:

You care mostly about the specific objects.

The example grounds discussion and illustrates larger points.

I am much more interested, in most contexts, in that second one. So I'd like to see an extra sentence here, something like "Especially when this concreteness helps people to understand things that can then be generalized to other contexts."

## Putting That Together

That was a lot of detail, which together forms a pattern and vibe. That vibe says that what is desired is a *superficial* critique that *stays within and affirms the EA paradigm* while it also *checks off the boxes of what 'good criticism' looks like* and it also *tells a story of a concrete win* that justifies the prize award. Then everyone can feel good about the whole thing, and affirm that EA is seeking out criticism.

## Formats For Investigation

The post also suggests adapting one of the standard forms of investigation that have been established with in the EA paradigm. Giving examples of things to do is good, and a lot of this seems reasonable, but the vibe is repeatedly reinforced.

You might consider framing your submission as one of the following:

**Minimal trust investigation** — A [minimal trust investigation](#) involves suspending your trust in others' judgments, and trying to understand the case for and against some claim yourself. Suspending trust does *not* mean determining in advance that you'll end up disagreeing.

**Red teaming** — [‘Red teaming’](#) is the practice of “subjecting [...] plans, programmes, ideas and assumptions to rigorous analysis and challenge”. You’re setting out to find the strongest reasonable case against something, whatever you actually think about it (and you should flag that this is what you’re doing).

**Fact checking and chasing citation trails** — If you notice claims that seem crucial, but whose origin is unclear, you could track down the source, and evaluate its legitimacy.

**Adversarial collaboration** — An [adversarial collaboration](#) is where people with opposing views work together to clarify their disagreements.

**Clarifying confusions** — You might simply be confused about some aspect of EA, rather than confidently critical. You could try getting clear on what you’re confused about, and why.

**Evaluating organizations** — including their (implicit) theory of change, key claims, and their track record; and suggesting concrete changes where relevant.

**Steelmanning and ‘translating’ existing criticism for an EA audience** — We’d love to see work succinctly explaining these existing ideas, and constructing the strongest versions ([‘steelmanning’](#)) them. You might consider doing this in

collaboration with a domain expert who does not consider themselves part of the EA community.

In particular, this suggests various formats *that are unlikely to offer central or fundamental criticism*, or often even criticism at all, and that make it impossible to break out of or challenge the paradigm.

A minimal trust investigation, fact checking, chasing citation trails and evaluating organizations are inherently local, standard things that EA does all the time. Red teaming is similar. You're sizing up the fish but you're not going to notice the water.

Adversarial collaboration I haven't loved as I've seen it on SSC/ACX, tending to end up as a dialectic of sorts and rarely breaking out, although it can often help with factual disputes along the way. Mostly it seems like it picks a disagreement, makes it as legible as possible, and ensures (here) that half the work goes into defending rather than criticizing whatever the target may be. I don't expect this to give us anything too challenging.

Clarifying confusions is explicitly noted to often not even be critical at all, if you don't count that something was not sufficiently clearly explained and thus you are confused. The description is pointing to the possibility of doing this in a non-critical fashion, to reward the noticing of confusion - a fine thing to reward, but likely to be with punches pulled. This framing pushes for a general modesty, setting up for the conclusion where confusions are resolved by explanation (that will often largely have effectively been social proof).

Steelmanning and translating existing criticism is taking things that are outside of the paradigm, and bringing them inside the paradigm. Often I worry this will take the most valuable disagreements and pretend not to see them, but it could also go the other way and result in a big flashing sign that says 'this argument doesn't fit into the paradigm because it rejects the following assumptions.' The key is to not then go, 'oh, this argument doesn't understand these important fundamental things, this person needs to study EA basics more.' And yes, *that has happened to me and I did not like it.*

The steelmanning part, in particular, seems likely to cause this. If an EA is looking to steelman criticism, there will be great temptation to change the thinking and arguments to be more EA-correct and thus throw out the most important content. [Rob Bensinger called steelmanning 'niche'](#) recently for similar reasons, quoting many well-known arguments against steelmanning.

I think there are two *distinct* useful things that can be done in this space.

Taking existing criticism and making it understood as it was intended.

Taking existing criticism and extracting and reasoning from its good points.

Steelmanning does one at the expense of the other.

What is most significant in this list is *what it leaves out*. In particular, if one has a central disagreement with something important about EA or its paradigm, that seems like *the most important thing* to talk about. How should one express that? One can frame it as 'expressing confusion' but that is *pretending* to be confused in order to allow those involved to save face and/or not actually engage fully. The last point explicitly says one must take *existing* criticism, rather than your own.

Overall, if I was looking to express important fundamental disagreements, to offer the most powerful criticisms of EA overall, to challenge its assumptions and Shibboleths, this list would discourage me quite a lot.

## The Judging Panel

This looks like a large judging panel (so consensus and social dynamics will likely be important even without veto powers) that consists entirely of EA insiders who likely buy into the core EA principles.

I didn't notice this at all unprompted, partly because I assumed of course such a thing would be true, then in the Twitter discussion someone else noticed this as a reason to be skeptical.

A reasonable objection is that if you are having a contest to see who can cause people in EA to change their minds, a non-EA judge would not be able to evaluate that. To the extent that evaluation was directly on the basis of whether the enemy was cut and minds changed, this seems reasonable - if the insiders all reject your entry then you likely did not productively change minds of insiders. If good entries that *should* change minds get rejected, that means the whole thing was pointless no matter who got the prize funds. If they get accepted, system works, great.

If you're judging on the basis of intermediate metrics instead, and considering the social implications of various awards and implied commitments and such, and are effectively more focused on box checking in various ways, then insider-only judge panel is a serious problem.

It also sends a clear message to someone who knows they are telling insiders what they do not want to hear. Which, of course, is often the most important criticism. You are much more likely to hear what you *need* to be told but don't *want* to be told, if there are outsiders on the panel.

## Rationale

This section is useful in large part to contrast the *conscious intentional story* it tells with the story being told above, but also it has things worth noticing in other ways.

In his opening talk for [EA Global](#) this year, Will MacAskill considered how a major risk to the success of effective altruism is the risk of degrading its quality of thinking: "if you look at other social movements, you get this club where there are certain beliefs that everyone holds, and it becomes an indicator of in-group mentality; and that can get strengthened if it's the case that if you want to get funding and achieve very big things you have to believe certain things — I think that would be very bad indeed. Looking at other social movements should make us worried about that as a failure mode for us as well."

This implies that MacAskill *does not* believe that you *currently* need to (missing but belongs: pretend to) believe certain things to get big EA funding. I would be very surprised if this implication was not correct. Or at a minimum, that it helps quite a lot.

I don't even think that is obviously an error. It seems more like a question of selection, magnitude and method. EA has a giant pool of money, and it needs to be protected somehow. We are, as Churchill said, talking price.

This paragraph is also of note:

It's also possible that some of the most useful critical work goes relatively unrewarded because it might be less attention-grabbing or narrow in its conclusions. Conducting really high-quality criticism is sometimes thankless work: as the blogger Dynomight [points out](#), there's rarely much glory in fact-checking someone else's work. We want to set up some incentives to attract this kind of work, as well as more broadly attention-grabbing work.

This is saying once again that they want high *quality* criticism in terms of getting the details and facts right, and they don't mind if the conclusions are narrow.

I am totally sympathetic to the goal here. Good narrow criticism is a worthwhile exercise, and I can totally believe that throwing money at this to create supply is a good idea. I have no problem with setting out to get a bunch of fact checks via a fact-checking contest. The contest format seems like an odd fit but it should still work. That's simply a much more compact and modest goal, and one that is very different from a general call for important criticism and mind changing.

I myself do some form of fact checking reasonably often. I consider it valuable, and would be appreciative if there was a good easy-to-use fact checking service that could do the 'thankless' parts of it while I did other parts, to let me put more focus elsewhere.

The final note is that they welcome criticism that EA's mistake may potentially be *not being EA enough*, of not being sufficiently weird or having sufficient urgency.

We're not going to privilege arguments for more caution about projects over arguments for urgency or haste. Scrutinizing projects in their early stages is a good way to avoid errors of *commission*; but [errors of omission](#) (not going ahead with an ambitious project because of an unjustified amount of risk aversion, or oversensitivity to downsides over upsides) can be just as bad.

I do think this is a good note. As the contest notes, my willingness to offer criticism at all is *good news*. If something (EA, the contest, etc) isn't interesting enough or lacks potential, criticism is wasted time. EA differs from the mainstream in lots of ways, and it seems all but certain that many of them involve EA not moving away from the mainstream far enough, even if the direction is mostly correct.

## Confidence Levels

Since it was explicitly requested, it makes sense to spell out my confidence level explicitly.

I'm almost certain of most of the specific detailed observations about individual words or phrases, in terms of their effect *on me*. I am almost as confident that the result of these details is that the entries into the contest will be less important and less impactful than they otherwise would be, in expectation. I am less confident, but still pretty confident, that each of them individually has that effect in general. I am somewhere in between in my confidence that none of this is a coincidence.

I am highly confident that the core observation is right - that the post was sending out a vibe with the effect of discouraging important criticism in favor of superficial

criticism or things that aren't even criticism, and that this was a reflection of the intent (at some level) of the people and/or systems that led to the contest.

And I am highly confident that this is reflective of a broader problem.

Where I have the core disagreements with EA or otherwise have a unique philosophy, modesty considerations would say I need to not be confident. Of course, one of my disagreements is that I am skeptical of modesty arguments, but I do think that *you* reading this should be skeptical here unless you reason it out on your own.

What would change my mind on any or all of this? There's no *one particular* detail that I know would do it, but learning I was repeatedly wrong about how people are reading and interpreting things seems like the most likely way. That's not the *only* way to do it, but other ways seem like bigger overall updates that would be much harder to get to, and seem less likely to be right.

Another thing I could change my mind on is the worthwhileness of writing about various topics, which is a function of the extent to which:

I learn something and change my mind via writing and seeing reactions to it.

Others change their minds because of my writing and reactions to it.

Social consequences are good rather than bad and I end up in interesting and productive interactions as a result rather than stressful pointless arguments.

Writing is fun and feels like exploration and learning, not like forced work.

There is revealed willingness to pay for such work in whatever way, both because it justifies time spent and also it indicates that others value the work.

I have availability, which may become very limited if one of my projects goes sufficiently well, in which case that would almost certainly take priority.

I have a lot of uncertainty about all of these points. As noted above [I have a fun little Manifold Markets](#) up on whether or not I'll get at least a second prize.

## Conclusion

This has been a critique *about critiques*, and about solicitation of and reactions to critiques. With the core critique being that what is presenting itself as a request for important critiques is instead a request for superficial critiques rather than important or fundamental critiques, because there is motivation [to tell a story that](#) important critiques are solicited rather than an actual appetite for fundamental critiques.

Superficial critiques both support this storytelling, and are actually welcome in their own right.

This has been something in-between. It *gestures towards* fundamental critiques, but doesn't focus on actually making them or justifying them. Making the actual case properly is hard and time intensive, and signals are strong it is unwelcome and would not be rewarded.

In the interest of actionable and constructive, what can be done?

A lot, even without considering the outcomes from more fundamental criticisms. Here are some places to start.

Judge the Contest Based Only on Whether Something Changed Your Mind. There's nothing *forcing* the judges to Goodhart themselves here if they don't want to. They still could simply not do so. That's a good first step.

Continue Soliciting Criticism. Even when it has issues, still a good thing. Even better, solicit more fundamental critiques as explicitly as possible, in addition to watching out for actively discouraging such things.

Look for the Fnords. When designing a system for awarding money, whether it is grants, a contest, a job or something else, seek conscious awareness of what your system is telling people it wants and will reward, and what it is telling those evaluating to want and reward. Then ask if that's what you want.

Vibe and Implication Matters. When you read or write something meant to induce behavior or belief, there is a kind of 'listen for the vibe' move that you need to make, and a 'notice the implications my brain is picking up unconsciously' related move as well. You need to draw them out into conscious awareness, then consider whether there is a problem. Then act accordingly.

Beware Self-Recommendations. If judgments on money and status are made exclusively by insiders who have gotten there in part by buying fully into the paradigm, and who have as a primary goal to direct resources towards the paradigm, the paradigm will not be questioned, and anything within it is at risk of also not being all that questioned. Get viewpoint diversity where it counts. That actually means *avoiding* having big panels in charge of such things - the only way to not have everything average out into consensus is to keep decision making at any given time on any given thing fast, nimble, flexible and concentrated on a small group, and ideally an individual. There should also be much less emphasis on whether something is EA or not, versus whether or not it is accomplishing something worthwhile and useful.

Question the Fundamentals. Everything from the core of utilitarianism on up, the full model of the world and how one ends up with a better one, needs to be up for grabs. It isn't merely a thing to be dealt with in Eternal September after which everyone can move on. Which also means finding ways to have these discussions not be stuck in Eternal September.

Heed the Implications. And of course, when you do get information, use it.

This post contains specific examples/suggestions of details and wordings that would bring improvement, and one can build from there.

(Note: There is [an additional copy](#) of this and other highly EA-relevant posts at the EA Forum, and it is likely that some discussion will take place there.)

# Examples of AI Increasing AI Progress

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Recursive self-improvement is already here.

This point is far from original. It's been described before, for instance [here](#), in Drexler's [Reframing Superintelligence](#), and (as I was working on this post) [in Jack Clark's newsletter](#) and even by [Yann LeCun](#). But sometimes I still hear people talk about preparing for "when recursive self-improvement kicks in," implying that it hasn't already.

The kinds of recursive self-improvement mentioned here aren't exactly the frequently-envisioned scenario of a single AI system improving itself unencumbered. They instead rely on humans to make them work, and humans are inevitably slow and thus currently inhibit a discontinuous foom scenario.

It may be tempting to dismiss the kind of recursive self-improvement happening today as not *real* recursive self-improvement. To think about it as some future event that will start to happen that we need to prepare for. Yes, we need to prepare for increasing amounts of it, but it's not in the future, it's in the present.

Here are some currently existing examples (years given for the particular example linked):

- (2016) Models play against themselves in order to iteratively improve their performance in games, most notably in [AlphaGo](#) and its variants.
- (2016) Some [neural architecture search](#) techniques use one neural network to optimize the architectures of different neural networks.
- (2016) [AI is being used to optimize data center cooling](#), helping reduce the cost of further scaling.
- (2021) Code generation tools like [GitHub Copilot](#) can be helpful to software engineers, including presumably some AI research engineers (anecdotally, I've found it helpful when doing engineering). Engineers may thus be faster at designing AI systems, including Copilot-like systems.
- (2021) Google [uses deep reinforcement learning](#) to optimize their AI accelerators.
- (2022) Neural networks, running on NVIDIA GPUs, [have been used](#) to design more efficient GPUs which can in turn run more neural networks.
- (2022) Neural networks [are being used](#) for compiler optimization in the popular LLVM compiler language, which [Pytorch's just-in-time compiler](#) is based on.

Inspired by Victoria Krakovna's [specification gaming spreadsheet](#), I've made a spreadsheet [here](#) with these examples. Feel free to submit more [here](#). I think the number of examples will continue to grow, making it useful to keep track of them.

If this feels underwhelming compared with the kinds of recursive self-improvement often written about, you're right. But consider that the start of an exponential often feels underwhelming. As time goes on, I expect that humans will become less and less involved in the development of AI, with AI automating more and more of the process. This could very well feel sudden, but it won't be unprecedented: it's already begun.

# Moral strategies at different capability levels

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Let's consider three ways you can be altruistic towards another agent:

- You care about their welfare: some metric of how good their life is (as defined by you). I'll call this care-morality - it endorses things like promoting their happiness, reducing their suffering, and hedonic utilitarian behavior (if you care about many agents).
- You care about their agency: their ability to achieve their goals (as defined by them). I'll call this cooperation-morality - it endorses things like honesty, fairness, deontological behavior towards others, and some virtues (like honor).
- You care about obedience to them. I'll call this deference-morality - it endorses things like loyalty, humility, and respect for authority.

I think a lot of unresolved tensions in ethics comes from seeing these types of morality as in opposition to each other, when they're actually complementary:

- Care-morality mainly makes sense as an attitude towards agents who are much less capable than you, and/or can't make decisions for themselves - for example animals, future people, and infants.
  - In these cases, you don't have to think much about what the other agents are doing, or what they think of you; you can just aim to produce good outcomes in the world. Indeed, trying to be cooperative or deferential towards these agents is hard, because their thinking may be much less sophisticated than yours, and you might even get to choose what their goals are.
  - Applying only care-morality in multi-agent contexts can easily lead to conflict with other agents around you, even when you care about their welfare, because:
    - You each value (different) other things in addition to their welfare.
    - They may have a different conception of welfare than you do.
    - They can't fully trust your motivations.
  - Care morality doesn't focus much on the act-omission distinction. Arbitrarily scalable care-morality looks like maximizing resources until the returns to further investment are low, then converting them into happy lives.
- Cooperation-morality mainly makes sense as an attitude towards agents whose capabilities are comparable to yours - for example others around us who are trying to influence the world.
  - Cooperation-morality can be seen as the "rational" thing to do even from a selfish perspective (e.g. [as discussed here](#)), but in practice it's difficult to robustly reason through the consequences of being cooperative without relying on ingrained cooperative instincts, especially when using causal decision theories. [Functional decision theories](#) make it much easier to rederive many aspects of intuitive cooperation-morality as optimal strategies (as discussed further below).

- Cooperation-morality tends to uphold the act-omission distinction, and a sharp distinction between those within versus outside a circle of cooperation. It doesn't help very much with population ethics - naively maximizing the agency of future agents would involve ensuring that they only have very easily-satisfied preferences, which seems very undesirable.
- Arbitrarily scalable cooperation-morality looks like forming a central decision-making institution which then decides how to balance the preferences of all the agents that participate in it.
- A version of cooperation-morality can also be useful internally: enhancing your own agency by cultivating virtues which facilitate cooperation between different parts of yourself, or versions of yourself across time.
- Deference-morality mainly makes sense as an attitude towards trustworthy agents who are much more capable than you - for example effective leaders, organizations, communities, and sometimes society as a whole.
  - Deference-morality is important for getting groups to coordinate effectively - soldiers in armies are a central example, but it also applies to other organizations and movements to a lesser extent. Individuals trying to figure out strategies themselves undermines predictability and group coordination, especially if the group strategy is more sophisticated than the ones the individuals generate.
  - In practice, it seems very easy to overdo deference-morality - compared to our ancestral environment, it seems much less useful today. Also, whether or not deference-morality makes sense depends on how much you trust the agents you're deferring to - but it's often difficult to gain trust in agents more capable than you, because they're likely better at deception than you. Cult leaders exploit this.
  - Arbitrarily-scalable deference-morality looks like an [intent-aligned](#) AGI. One lens on why intent alignment is difficult is that deference-morality is inherently unnatural for agents who are much more capable than the others around them.

Cooperation-morality and deference-morality have the weakness that they can be exploited by the agents we hold those attitudes towards; and so we also have adaptations for deterring or punishing this (which I'll call conflict-morality). I'll mostly treat conflict-morality as an implicit part of cooperation-morality and deference-morality; but it's worth noting that a crucial feature of morality is the coordination of coercion towards those who act immorally.

## **Morality as intrinsic preferences versus morality as instrumental preferences**

I've mentioned that many moral principles are rational strategies for multi-agent environments even for selfish agents. So when we're modeling people as rational agents optimizing for some utility function, it's not clear whether we should view those moral principles as part of their utility functions, versus as part of their strategies. Some arguments for the former:

- We tend to care about principles like honesty for their own sake (because that was the most robust way for evolution to actually implement cooperative strategies).
- Our cooperation-morality intuitions are only evolved proxies for ancestrally-optimal strategies, and so we'll probably end up finding that the actual optimal strategies in other environments violate our moral intuitions in some ways. For

example, we could see love as a cooperation-morality strategy for building stronger relationships, but most people still care about having love in the world even if it stops being useful.

Some arguments for the latter:

- It seems like caring intrinsically about cooperation, and then also being instrumentally motivated to pursue cooperation, is a sort of double-counting.
- Insofar as cooperation-morality principles are non-consequentialist, it's hard to formulate them as components of a utility function over outcomes. E.g. it doesn't seem particularly desirable to maximize the amount of honesty in the universe.

The rough compromise which I use here is to:

- Care intrinsically about the welfare of all agents which currently exist or might in the future, with a bias towards myself and the people close to me.
- Care intrinsically about the agency of existing agents to the extent that they're capable enough to be viewed as having agency (e.g. excluding trees), with a bias towards myself and the people close to me.
  - In other words, I care about agency in a [person-affecting way](#); and more specifically in a loss-averse way which prioritizes preserving existing agency over enhancing agency.
- Define welfare partly in terms of hedonic experiences (particularly human-like ones), and partly in terms of having high agency directed towards human-like goals.
  - You can think of this as a mixture of hedonism, desire, and objective-list [theories of welfare](#).
- Apply cooperation-morality and deference-morality instrumentally in order to achieve the things I intrinsically care about.
  - Instrumental applications of cooperation-morality and deference-morality lead me to implement strong principles. These are partly motivated by being in an iterated game within society, but also partly motivated by functional decision theories.

## Rederiving morality from decision theory

I'll finish by elaborating on how different decision theories endorse different instrumental strategies. Causal decision theories only endorse the same actions as our cooperation-morality intuitions in specific circumstances (e.g. iterated games with indefinite stopping points). By contrast, [functional decision theories](#) do so in a much wider range of circumstances (e.g. one-shot prisoner's dilemmas) by accounting for logical connections between your choices and other agents' choices. Functional decision theories follow through on commitments you previously made; and sometimes follow through on commitments that you would have made. However, the question of which hypothetical commitments they should follow through with depends on how [updateless](#) they are.

Updatelessness can be very powerful - it's essentially equivalent to making commitments behind a veil of ignorance, which provides an instrumental rationale for implementing cooperation-morality. But it's very unclear how to reason about how justified different levels of updatelessness are. So although it's tempting to think of updatelessness as a way of deriving care-morality as an instrumental goal, for now I think it's mainly just an interesting pointer in that direction. (In particular, I feel

confused about the relationship between single-agent updatelessness and multi-agent updatelessness like the original veil of ignorance thought experiment; I also don't know what it looks like to make commitments "before" having values.)

Lastly, I think deference-morality is the most straightforward to derive as an instrumentally-useful strategy, conditional on fully trusting the agent you're deferring to - epistemic deference intuitions are pretty common-sense. If you don't fully trust that agent, though, then it seems very tricky to reason about how much you should defer to them, because they may be manipulating you heavily. In such cases the approach that seems most robust is to diversify worldviews using a [meta-rationality](#) strategy which includes some strong principles.

# Circumventing interpretability: How to defeat mind-readers

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

(Post now available as a pdf: <https://arxiv.org/abs/2212.11415>)

TL;DR: Unaligned AI will have a convergent instrumental incentive to make its thoughts difficult for us to interpret. In this article, I discuss many ways that a capable AI might circumvent scalable interpretability methods and suggest a framework for thinking about these risks.

## Summary of key points

### The framework: Interpretation as information flow

I categorize potential interpretability circumvention methods on three levels:

1. Interpretability methods consist of interpretability **tools** and the human **interpreter**. To circumvent interpretation an AI can prevent information about its thoughts from flowing through either the tool or the interpreter. (See *main text for figure*)
2. Within the tool or the human interpreter, there are several information processing steps that an AI might exploit. (See *main text for figure*):
  - **Exposure evasion**: Being where your adversary isn't looking
  - **Perception evasion**: Being hard to detect even when exposed to your adversary
  - **Recognition evasion**: Being hard to recognize/understand even though you've been perceived
  - **Dazzling**: Looking like something that confuses or incapacitates your adversary
3. Another useful way to categorize interpretability circumvention methods is whether they are **passive** or **active**. Most methods that can be passive can also be active, but other methods are strictly active.
  - **Passive** interpretability circumvention occurs during the normal course of learning without the AI intending to make itself less interpretable. (e.g. polysemy)
  - **Active** interpretability circumvention is where the AI intentionally takes actions in order to circumvent our interpretability methods. (e.g. an AI actively deleting its own representations to hide deception)

### The list of potential interpretability circumvention methods with a short summary of each

- 1. Exploiting weaknesses of the interpretability tool
  - 1.1. Evading exposure to the interpretability tool
    - 1.1.1. Thinking in the gaps of our tools: Our interpretability tools might fail to examine everything about an AI's thoughts (e.g. it might not look at nonlinear interactions between neurons, only their linear relations). An AI might use those blindspots for dangerous thought. (*Passive or active*)
    - 1.1.2. Externalizing representations: The environment can be used for some parts of an AI's computation. Actively: The agent can write messages to itself or initiate programs. Passively: The AI could use the environment as a reservoir in a [reservoir computer](#). (*Passive or active*)

- **1.1.3. Environmental artifacts that corrupt its future self:** A misaligned AI may leave behind artifacts that serve as adversarial examples or poisoned data that corrupt its future self. (*Strictly active*)
- **1.1.4. Deleting (and relearning) dangerous representations:** When not under surveillance, an AI might be able to delete dangerous representations so that it looks safe when surveillance resumes. It might also be able to schedule when it relearns those representations. (*Strictly active*)
- **1.1.5. Weights poisoning:** Instead of providing manipulated inputs, the AI might manipulate its own weights so that a small subset of normal inputs behave like adversarial examples. (*Strictly active*)
- **1.1.6. Pre-committing to following potentially dangerous instructions:** Hubinger: "... if the model has the ability to read and write to some sort of state ... it could write out a pre-commitment to follow the hard-coded policy and then just have a policy of always following any pre-commitments it finds such that it never thinks about the deception that led it to that policy in the first place." (*Strictly active*)
- **1.1.7. Hard-coded neural shortcuts:** Hubinger again: "... it could hack its own gradient to produce a new model which will still exhibit the same deceptive behavior, but without doing any of the detectable deceptive cognition. ... not because it's trying to act deceptive, but just because it's a hard-coded part of its algorithm" (*Strictly active*)
- **1.1.8. Following cached thoughts:** Nanda: "[An AI might develop] the instinct to follow cached thoughts without thinking about them, and in occasional moments of deception to generate a lot of cached thoughts that will lead to misaligned actions, but with no deceptive cognition in the moment." (*Strictly active*)
- **1.1.9. Instrumental incapability:** A misaligned AI that's able to modify its own parameters could damage capabilities that serve some goals (e.g. aligned goals) more than its misaligned goals. Generalizes instrumental clumsiness. (*Strictly active*)
- **1.1.10. Instrumental clumsiness:** An AI might hack its future self so that it is more likely to make certain kinds of mistakes while leaving other capabilities intact. (*Strictly active*)

## Full article

*Acknowledgements:*

I'm grateful to David Liniger, Evan R. Murphy, Alex Lintz, Sia Black, Kyle McDowell, Larisa Reynolds, Adam Shimi, and Daniel Braun whose comments greatly improved earlier drafts of this article. The article's weaknesses are mine, but many of its strengths are due to their contributions.

### 1.2. Evading perception by the interpretability tool

1.2.1. **Hiding in noisy subspaces:** We might ignore noisy regions or subspaces of neural activation space may be very noisy. But an AI might be able to do some dangerous thinking in those spaces. Generalizes 'Whispering dangerous thoughts'. (*Passive or active*)

1.2.2. **Whispering dangerous thoughts:** An AI might make its dangerous thoughts very quiet relative to other thoughts by using small norm activation vectors. (*Passive or active*)

### 1.3. Evading recognition by the interpretability tool

1.3.1. **Antiparsimony:** If it costs more to interpret an AI's thought than it costs to think, an AI might be able to produce a lot of representations such that it saturates the interpreter's computation before it saturates its own. (*Passive or active*)

This work was carried out in the [Evan Murphy lab](#). Suppose we have an AI with a memory state. The AI might use very long chains of seemingly meaningless neural activations to construct a single thought. (*Passive or active*)

1.3.3. **Interpretability illusions:** By manipulating the dataset shown to the tool, an AI can control its interpretations. (*Strictly active*)

1.4. **Attacking/dazzling the interpretability tool**: There's been a lot of excitement about interpretability recently as a component in many different alignment proposals.

1.4.1. **Adversarial attack against networks in the interpretation tool:** The AI could adversarially attack networks in the tool to manipulate the tool's outputs. (*Strictly active*)

But if we're interpreting an unaligned AI, it has an divergent instrumental incentive<sup>[1]</sup> to circumvent our interpretability methods. There's been some sporadic past discussion about how an AI might achieve this (e.g. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)). But to my knowledge, the problem hasn't been addressed systematically, at least not publicly.

### 2. Exploiting weaknesses of the human interpreter

If we continue to make large bets on interpretability, then the lack of attention on potential circumvention methods seems bad because we might end up with a false sense of security. Our

#### 2.2.1. None identified

#### 2.3. Evading recognition by the human interpreter

interpretability methods 2011 **Alien representations** Augmented representation whose entire meaning lies simply found a way to think dangerous thoughts without entry. Generalizes 'Polysemy' and 'Representational kludging' (Passive or active)

We can think of this problem as **2012: Why is neural network research agenda related to but distinct from the well-known agenda that passive? Can we understand what today's neural networks are doing? And how can we** **2013: Representation Research**: creating the right short term agenda is exceptionally important **2014: Unknown representations**: research agenda is immediately succeeded, how might a possible future research agenda be related to but have the concepts for

them. But we're capable of learning them. (*Passive or active*)

**12.3.5 UNKNOWABLE REPRESENTATIONS** Some concepts in mathematics and other sources came up with. Throughout I assume we're concerned with interpreting the representations [2] of a capable of thought system (possibly or above) deep learning-based AI. I also suggest a framework for thinking about the existing ventriloquist human interpreter.

**214 Attacking the dazzling ventriloquist human interpreter** I hope will help to identify omitted circumvention methods. Therefore, it can be used for attacking, or incapacitating the

**2-350** **UNPREDICTABLE** **REPRESENTATIVENESS** **BIAS** **IN** **PERCEIVED** **RISK** **AND** **DECISION** **MAKING**

and other sources came to humans to understand due to the computational limitations of our brain or representations<sup>[2]</sup> of a capable enough system as discussed above) deep learning-based AI. I also suggest a framework for thinking about attacking the human interpreter which hope will help to identify omitted circumvention methods therefore capable of attacking or incapacitating the

## 2.4 Attacking/dazzling the human interpreter

#### **2.4.1 Manipulating and hacking, or incapacitating the**

**interpreter:** An AI might be able to manipulate, directly control, or incapacitate humans using specific kinds of outputs. (*Strictly active*)

## The framework for taking on information flow

We can think of interpretation as in formation flow (figuring). Interpretation is the following based on model of information flow of interpretability or compute).

- We can draw many analogies between mind-reading/interpretability circumvention and reverse-engineering thoughts as being interpreted by an interpretability method.
  - An *interpretability method* is the process of translating an AI's representations into whatever the corresponding representations are in humans.
  - This is a *gradient hacking* method. It is therefore less likely to happen because a machine learning model's interpretability probably matches the analysis machinery that readapt the AI's representations and converts them into something legible to humans (circumstances) for interpreting dangerous thought.
  - *Human interpreters* read the outputs of the interpretability tools and attempt to understand them, thus converting them into human representations.

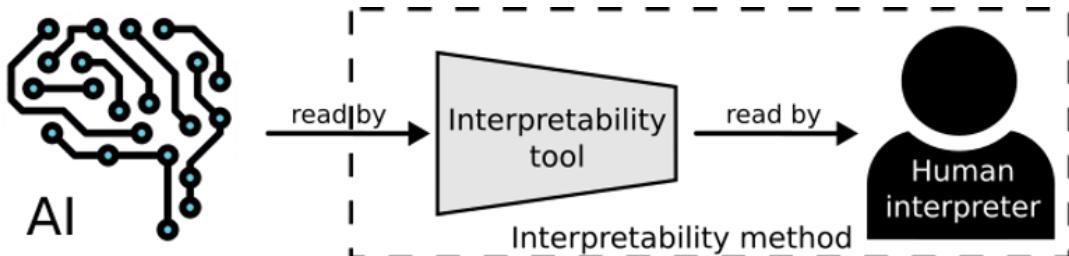


Figure 1. A simple model of information flow in interpretability.

Throughout the article, I assume we've figured out ways to interpret our models *at scale*, which probably means that our interpretability tools are automating a lot of the work. Nevertheless, all the potential circumvention methods I talk about apply to worlds where we still haven't managed to scale interpretability, where the tools would be relatively less important.

An even more detailed model of information flow (figure 2) is useful for categorizing interpretability circumvention methods, as we'll see:

- Information flows within both the interpretability tool and the human. In both,
  - What is to be interpreted must first be *exposed* to the tool or the human.
  - What is exposed must then be *perceived* or detected.
  - What is perceived is then further analysed so that it can be *recognized* or understood, which then determines an appropriate output by the tool or action by the human.

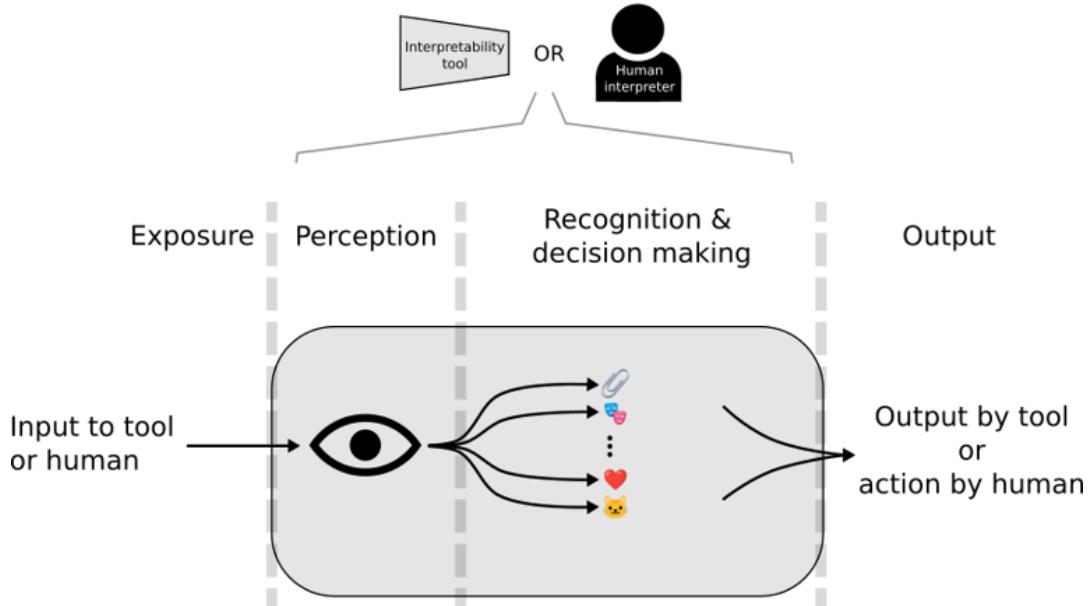


Figure 2. A more detailed model of information flow which is applicable to both the interpretability tool and to the human. Information flow begins with exposure of the information to the observer, then perception by the observer, then recognition & and decision making, followed by output.

## How I categorize interpretability circumvention methods

The ways to categorize interpretability circumvention methods that I've found most useful (and the way this article is structured) are:

### A. Exploiting weaknesses of the tool vs. weaknesses of the human interpreter

Interpretability circumvention methods can either circumvent the *interpretability tool* or the *human interpreter*.

### B. Categories of interpretability circumvention found in nature

Evolution has found many different ways that make it difficult for adversaries to interpret the appearance of an organism. I put them into 4 groups (figure 3), which broadly correspond to the detailed model of information flow described above:

- **Exposure evasion:**
  - Being where your adversary isn't looking
  - E.g. An animal that lives underground or up a tree, away from surface-dwelling predators
- **Perception evasion:**
  - Being hard to detect even when exposed to your adversary
  - E.g. An animal that only comes out when it's dark; color camouflage to look the same as the color of the background
- **Recognition evasion:**
  - Being hard to recognize/understand even though you've been perceived
  - E.g. Biomimicry, where an organism looks like another kind of object entirely, for instance: an Orchid mantis which looks like the flower; the harmless Hornet Moth mimicking the colors of a wasp.
- **Dazzling:**

- Looking like something that confuses or incapacitates your adversary
- E.g. Living in large groups such that it's hard for a predator to focus on any particular individual; a zebra's stripes.

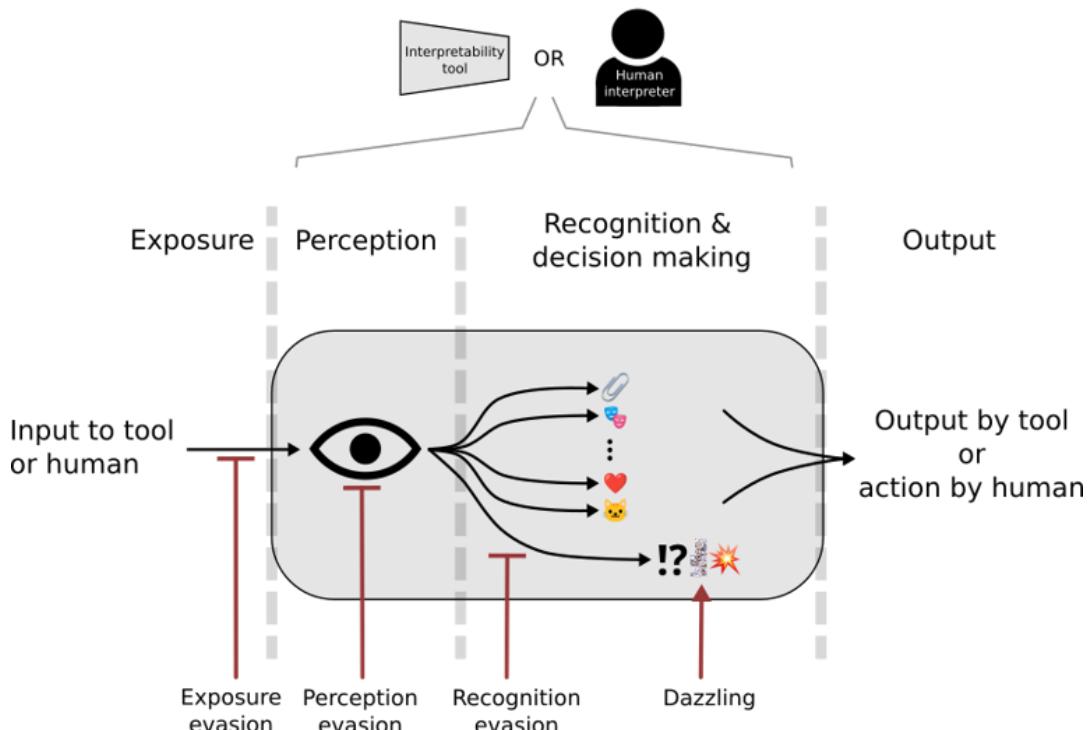


Figure 3. Interpretability circumvention methods exploit different steps in the information flow diagram and correspond to forms of interpretability circumvention found in nature. Circumvention methods may seek to evade exposure, perception, or recognition. They may also exploit vulnerabilities in the observer's information processing system, a process called dazzling.

### C. Passive vs. active methods

Information flow might be inhibited **passively** by circumvention methods that arise during the normal course of learning. One example is 'polysemy', which is a natural phenomenon that presents a challenge to interpretability. Passive circumvention methods might also arise if they're directly optimized for. This might happen if we use *interpretability tools in the loss function*, which is where we identify unsafe representations in an AI and then use the loss function to penalize the AI for containing those representations. At first glance, this seems like a potentially good way to make an AI safe. But it risks pushing unsafe representations into the blind spots of our interpretability tools.

Information flow might be inhibited **actively** through an agent's actions. The primary way this could happen is [gradient hacking](#), but it's not the only kind of action an AI might take to conceal misaligned thoughts. Of course, active circumvention methods require that interpreters either can't see or aren't looking at the thoughts that generate those actions.

Most potential circumvention methods that can be passive can also be active. But some methods can only be active. It's usually obvious when either is the case, so I only highlight this categorization method when it's not obvious.

This categorization is mostly useful for identifying how and when during learning the method is likely to arise. Strictly active methods generally require advanced capabilities (such as gradient hacking), whereas passive methods don't. I tend to think this makes strictly active methods less likely than methods that may be passive or active.

Before I introduce the individual circumvention methods in detail (see next section), the table below puts individual circumvention methods into categories according to whether they may be 'passive or active' or whether they are 'strictly active'.

Methods that can be passive OR active	Methods that are strictly active
<p>May happen during the normal course of learning and are potentially incentivised by using interpretability tools in the loss function</p> <ul style="list-style-type: none"> <li>(1.1.1.) Thinking in the gaps of our tools</li> <li>(1.1.2.) Externalizing representations</li> <li>(1.2.1.) Hiding in noisy subspaces</li> <li>(1.2.2.) Whispering dangerous thoughts</li> <li>(1.3.1.) Antiparsimony</li> <li>(1.3.2.) Encryption</li> <li>(2.3.1.) Alien representations</li> <li>(2.3.2.) Polysemanticity</li> <li>(2.3.3.) Representational kludging</li> <li>(2.3.4.) Unknown representations</li> <li>(2.3.5.) Unknowable representations</li> </ul>	<p>Probably require intentional action by the AI for dangerous instances to happen</p> <ul style="list-style-type: none"> <li>(1.1.3.) Environmental artifacts that corrupt its future self</li> <li>(1.1.4.) Deleting (and relearning) dangerous representations</li> <li>(1.1.5.) Weights poisoning</li> <li>(1.1.6.) Pre-committing to potentially dangerous instructions</li> <li>(1.1.7.) Hard-coded neural shortcuts</li> <li>(1.1.8.) Following cached thoughts</li> <li>(1.1.9.) Instrumental incapability</li> <li>(1.1.10.) Instrumental clumsiness</li> <li>(1.3.3.) Interpretability illusions</li> <li>(1.4.1.) Adversarial attacks against networks in the interpretability tool</li> <li>(2.4.1.) Manipulating, mind-hacking, or incapacitating the interpreter</li> </ul>

Table 1. Categorizing interpretability circumvention methods as 'Passive or active' or 'Strictly active'

This article is structured according to the above three categorization schemes, primarily the first two. I'll discuss circumvention methods that exploit weaknesses either of the tool or of the human. For each, I'll look at methods of exposure evasion, perception evasion, recognition evasion, and dazzling. Throughout, I'll highlight whether a method can arise passively or if it's strictly an active method, when the distinction matters.

I'm not confident this categorization scheme is optimal or even totally consistent, but I've found it helpful. Occasionally, a circumvention method in the list is in fact a generalization of another method in the list, yet I present them separately if the special case is particularly interesting or likely to arise.

## Interpretability circumvention methods

1. Exploiting weaknesses in the interpretability tool			
1.1. Exposure evasion	1.2. Perception evasion	1.3. Recognition evasion	1.4. Dazzling
1.1.1. Thinking in the gaps of our tools	1.2.1. Hiding in noisy subspaces	1.3.1. Antiparsimony (Firehose of representations)	1.4.1. Adversarial attacks against networks in the interpretability tool
1.1.2. Externalizing representations	1.2.2. Whispering dangerous thoughts	1.3.2. Encryption (Lengthy individual representations)	
1.1.3. Environmental artifacts that corrupt its future self		1.3.3. Interpretability illusions	
1.1.4. Deleting (and relearning) dangerous representations			
1.1.5. Weights poisoning			
1.1.6. Pre-committing to potentially dangerous instructions			
1.1.7. Hard-coded neural shortcuts			
1.1.8. Following cached thoughts			
1.1.9. Instrumental incapability			
1.1.10. Instrumental clumsiness			

2. Exploiting weaknesses in the human interpreter			
2.1. Exposure evasion	2.2. Perception evasion	2.3. Recognition evasion	2.4. Dazzling
None identified	None identified	2.3.1. Alien representations 2.3.2. Polysemy 2.3.3. Representational kludging 2.3.4. Unknown representations 2.3.5. Unknowable representations	2.4.1. Manipulating, mind-hacking, or incapacitating the interpreter

Table 2. An overview of the list of potential interpretability circumvention methods and their categories.

## 1. Exploiting weaknesses of the interpretability tool

### 1.1. Evading exposure to the interpretability tool

#### 1.1.1. Thinking in the gaps of our tools

When interpreting a capable AI, our interpretability tools might not examine every possible representational detail used by the AI for its cognition. An AI might thus be able to 'think within the gaps' of our interpretability tools.

A few examples make this clearer:

1. Today's mechanistic interpretability methods don't let us easily interpret representations that use nonlinearities ([Voss et al. 2021](#); [Elhage et al. 2021](#)). If this blindspot persists, a model might be able to use nonlinearities to hide dangerous thoughts.
2. We might have interpretability tools that look only for an expected set of concepts in a model's activations. But the AI may find dangerous concepts that don't belong to that set because we hadn't yet thought of them.
3. Hiding in summary statistics: If an interpretability tool uses the mean value when deciding whether some quantity in the AI is of concern, then an AI might be able to use high variance and highly skewed data to hide concerning statistics. E.g. It could hide the fact it values most paperclip-maximizing scenarios *a lot* by offsetting them with a small number of very low utilities in a small number of such scenarios. The mean would look normal, but if we looked at the variance we'd see a more concerning picture.
4. We might expect language to be the kind of representational space in which an AI can think dangerous thoughts. But what about in representational spaces we don't expect, like generic dynamical systems modeling modules? We know dynamical systems (such as an RNN) can be Turing complete, so maybe we should have expected dangerous representations in there. But what about in regions that are even less obviously capable of computing dangerous thoughts, like in representations constructed from syntactic patterns or punctuation? That's a kind of representational space that might be expressive enough to encode dangerous thoughts, but not one of the obvious first places to look<sup>[3]</sup>.
5. A fictional example: In [HPMOR](#) Harry tries to protect against mind-readers (occlumency) by essentially hosting two personas at once. An AI could learn a whole 2nd set of representations or a whole second internal agent. Whether this would work, again, depends on the details of the interpretability tool. If, for instance, the interpretability method identifies 'the deception

'representation' within the primary representation set and then stops looking for deception-like representations, it might fail to find the deception representation in the secondary set.

This class of methods is likely to emerge passively (though may also emerge actively), and many instances of it would be encouraged by using interetability tools in the loss function.

### **1.1.2. Externalizing representations**

If you keep forgetting to reply to an email and therefore set an electronic reminder, the reminder is a functionally relevant variable in the "respond to email" behavior. Indeed, if you're forgetful enough, it's an *essential* variable in that behavior. This is the basic idea behind the '[extended mind thesis](#)' from Andy Clark & David Chalmers: There is no hard barrier between the mind and the environment; cognition is often a dance between internal and external variables and can't be reduced to only one set. If representations are 'cognitively functional variables', then representations may be either internal or external. Using the environment as a memory system like this can be called *autostigmergy* ([Deverett et al. 2019](#)).

An AI might be able to passively or actively externalize representations through the effects of its actions. If our interpretability methods ignore those parts of the environment that the AI is using to think, then we may miss some dangerous cognition.

Are externalized representations likely to be a problem? Let's consider active and passive cases separately because, unlike for most other interpretability circumvention methods, they're remarkably different for 'Externalizing representations'.

#### **Active**

*Actively* externalizing representations is where the AI *intentionally* uses the environment as a memory system. An AI probably first needs to represent the cognitive variables internally before it can encode them into the environment. If we're not constantly surveilling the AI, this is easily possible. This includes the obvious methods like

- Writing a message to its future self, potentially steganographically. For example, an AI that can manipulate objects in a room may be able to encode messages for its future self by subtly arranging the objects in a room in a particular way. Or a virtual assistant agent might be able to encode messages for itself in a web search history or in different filename patterns.
- Running a computer program to decide whether it's time to defect or not.

An edge case is the act of creating another agent in the environment. But we'd then be dealing with a whole new copy of the agent that isn't under surveillance, and we've probably already lost it if this happens, so I don't consider it further here.

#### **Passive**

It's less clear whether it's possible to *passively* externalize representations in order to circumvent interpretability methods. The representations required for dangerous thought have a complicated structure. Learning such structure externally is a serious challenge because the AI doesn't get to learn it using backprop through the environment or explicit planning.

Nevertheless, one way an AI might achieve it is to use the environment as a [reservoir computer](#). For those who would benefit from an explanation of reservoir computing, it's easiest to describe in the supervised learning setting: To train a reservoir computer, you pass an input vector to a *reservoir*, which is a fixed dynamical system (such as an RNN with fixed weights). But you only learn the readout weights, which simply read the prediction from the reservoir variables. Here, the readout weights are usually linear, so they're not doing anything complicated; it's the variables in the reservoir that are doing the complicated nonlinear computation. Thus, the dynamics of the reservoir variables (rather than the learned parameters) can serve as representations that solve a task.

Going beyond supervised learning, reservoir computing can also work in the RL-setting: The reservoir is the agent+environment dynamics from timestep 0 to T – 1; an input observation is provided to the agent at timestep 0; and the readout weights are the weights of the agent at timestep T. If the

environment part of the reservoir contains most of the dangerous representations, we might miss some dangerous thoughts.

[Deverett et al. \(2019\)](#) provide an empirical example of a simple feedforward model-free agent that learns to use the agent-environment system as a reservoir computer (though they don't frame it as such). The memoryless agent learns to use its location in the environment to encode a time interval. It should be stressed that it does this without backprop through time or any knowledge of the past. But its location varies due to its initially random patterns of actions, yet it can use those patterns to encode a time interval.

The Deverett et al. example is a very simple representation. Maybe reservoir computing like this doesn't scale well enough to learn the complex representations that are required for dangerous cognition. But the inexpressivity of reservoir computing is partly due to the simplicity of the linear readout weights. With a powerful AI serving as a readout network, reservoir computing might be significantly more expressive and reliable. It might be an interesting research project to identify how expressive reservoir computing can be if we allow the readout network to be strongly nonlinear.

*Note: The rest of the methods discussed in this section ('1.1. Evading exposure to the interpretability tool') are strictly active methods.*

#### **1.1.3. Environmental artifacts that corrupt its future self**

A misaligned AI may leave behind artifacts that corrupt its (potentially aligned) future self so that it pursues the goals of its misaligned past self instead.

The artifacts may be *adversarial examples* that cause it to act in unexpected and dangerous ways. Alternatively, the agent could leave behind *poisoned data* so that its future self will learn dangerous representations; this would be an instance of gradient hacking.

This method might reasonably be considered a special case of 'Externalizing representations'. The distinction is somewhat arbitrary and depends on what qualifies as a representation. Does 'an attack that corrupts its future self' qualify as representation? It's not clear and is somewhat arbitrary, but I'm inclined to put them in separate categories.

#### **1.1.4. Deleting (and relearning) dangerous representations**

If an AI can gradient hack, it might be able to delete dangerous representations so that it looks safe when surveillance resumes.

Thus the volume under the *suspicious thought-surface*  $\times$  *interpretation*  $\times$  *time* curve is minimized, since the suspicious thought-surface is 0 everywhere that *interpretation*  $>$  0.

Of course, deleting thoughts only avoids detection; it doesn't permit dangerous thought to continue. But if an AI has learned dangerous representations and deletes them, it's probable that it will learn those representations again, assuming its data distribution doesn't change too much. An AI that can gradient hack might be able to schedule when it periodically re-learns the dangerous representations so that it always looks safe when surveillance resumes.

This method requires the agent to have good knowledge of when it is likely to be interpreted. If we interpret the AI every K batches, an AI might be able to use time information between batches to notice that the interval between some batches (validation/interpretation periods) are much longer than usual. If it's too expensive to surveil the AI constantly, there might be a cheaper monitoring scheme: We could interpret an agent at random intervals in order to create a 'panopticon' - where the agent is forced to behave/think safely because it doesn't know when it is safe for it to think thoughts about defection.

#### **1.1.5. Weights poisoning**

Instead of providing manipulated inputs, the AI might manipulate its own weights so that a small subset of normal inputs behave like adversarial examples, thus causing dangerous learning or behavior. This is essentially 'weights poisoning' ([Kurita et al. 2020](#)) where the AI is poisoning its own weights. It's the same idea (as far as I can tell) behind 'Trojan models' ([Rajabi et al. 2022](#)). Planting such backdoors in neural networks can be practically impossible to detect ([Goldwasser et al. 2022](#)).

This is kind of like fictional 'brainwashed sleeper agents' that are activated by an activation code. The difference is that the AI is turning *itself* into the sleeper agent and making the activation code something it's likely to encounter naturally when it finds itself in circumstances where its misaligned self wanted to be activated.

Peter Barnett had a nice [alternative analogy](#) here (which also applies to 'Environmental artifacts that corrupt its future self'):

"[A] mesa-optimizer might reason that it will be discovered if it remains active but it will be safe in the future to emerge again. The mesa-optimizer could 'store' the instructions for turning the network back into an optimizer at some point in the future. One analogy here is to an endospore which is a hardy, mostly inert form which some bacteria can take that allow them to lie dormant and survive hostile environments, and then re-emerge when conditions are better. The mesa-optimizer could 'store' these instructions to re-emerge either in the network weights, or in the environment or some other kind of memory."

#### **1.1.6. Pre-committing to potentially dangerous instructions**

This one comes from Evan Hubinger in his [article](#) on gradient-hacking, where he talks about how a deceptive AI might erase evidence of its deception while still carrying out policies generated when given the freedom to think deceptively:

"...if the model has the ability to read and write to some sort of state (e.g. an RNN or even a transformer which can attend to its previous output) ... it could write out a pre-commitment to follow the hard-coded policy and then just have a policy of always following any pre-commitments it finds such that it never thinks about the deception that led it to that policy in the first place."

#### **1.1.7. Hard-coded neural shortcuts**

Another potential circumvention method from Evan Hubinger's [article](#) on gradient hacking:

"... it could hack its own gradient to produce a new model which will still exhibit the same deceptive behavior, but without doing any of the detectable deceptive cognition. Perhaps the new model exhibits the deceptive behavior not because it's trying to act deceptive, but just because it's a hard-coded part of its algorithm, or because it has some bad proxy, or any other defect that might be more difficult to detect than outright deception."

A deep learning-based AI might be able to learn (through gradient hacking) to perform behaviors using hard-coded neural shortcuts.

The way I imagine this one working in is related to the Diversity Hypothesis ([Hilton et al. 2020](#)): "*Interpretable features tend to arise (at a given level of abstraction) if and only if the training distribution is diverse enough (at that level of abstraction).*" The Diversity Hypothesis implies that if a training distribution isn't diverse on a particular level of abstraction, then interpretable features will tend not arise on that level of abstraction (figure 4). Instead, neural networks tend to learn [shortcuts](#), which is simply what happens when a network learns to use directions in neural statespace that don't 'mean' anything in a sense that generalizes off distribution, yet they perform fine on the narrow training distribution. Creating neural shortcuts is therefore one way to 'hard code' parts of a neural algorithm without using interpretable features.

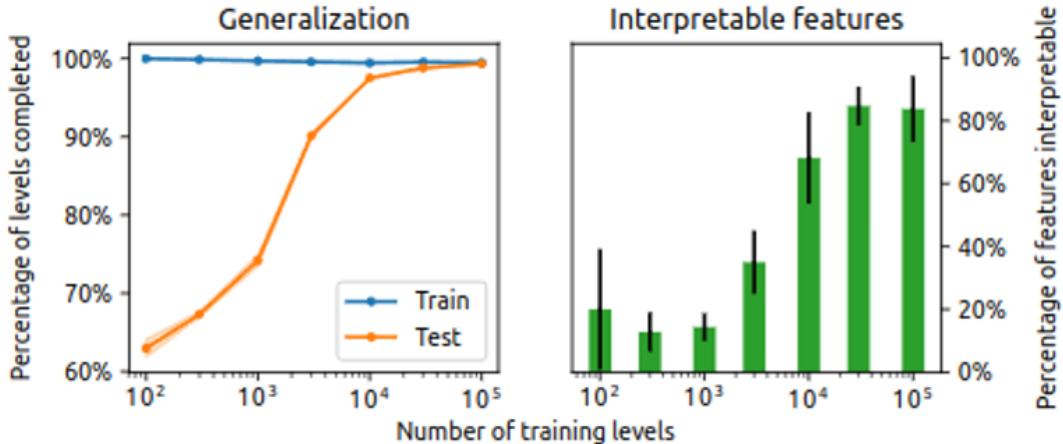


Figure 4. Taken from [Hilton et al. \(2020\)](#). Interpretable features tend to be those that generalize and that were learned on a diverse training distribution.

We might be able to detect and mitigate hard-coded neural short cuts using mechanistic interpretability or [relaxed adversarial training](#).

### **1.1.8. Following cached thoughts**

'Hard-coded neural shortcuts' is similar to a method later [proposed](#) by Neel Nanda:

"... maybe it *isn't* that hard to think non-deceptive thoughts, if you realize that someone will be scanning your mind, by e.g. developing the instinct to follow cached thoughts without thinking about them, and in occasional moments of deception to generate a lot of cached thoughts that will lead to misaligned actions, but with no deceptive cognition in the moment."

[Cached thoughts](#) are essentially 'pattern completions': You hear a few words then you automatically complete the pattern. In his article on the topic, Eliezer advised against this automatic mode of thought in favor of spending a few thought cycles to figure out what should *really* follow the prompt.

What might cached thoughts look like in deep-learning based AI? Cached thoughts are the default kind of thought in neural network based systems. My best guess is that you need recurrence in order *not* to do cached thoughts. Given an initial input, not doing cached thoughts means giving the input a deeper rollout than the default depth=0 that cached thoughts get. For example, in MuZero, cached thoughts means simply acting on priors (without MCTS), whereas *not* doing cached thoughts means 'using MCTS with a depth of K > 0'.

The idea behind 'following cached thoughts' is that some interpretability methods might only detect thoughts in the rollouts. If a MuZero-like AGI were to use this circumvention method, it would look like 'creating shortcuts through neural statespace so that, in certain contexts, dangerous thoughts can happen without deep rollouts, thus avoiding detection.'

Since cached thoughts require direct modification of the AI itself, it seems likely that this method could only happen during breaks in surveillance or after the AI has some other way of hiding the intention to follow cached thoughts.

Is this method likely to be a problem? I feel like any mechanistic interpretability method that actually works should be able to pick up context-dependent connections between representations. So this might not be a problem if we get scalable interpretability methods working.

### **1.1.9. Instrumental incapability**

A misaligned AI that's able to modify its own parameters may be incentivised to damage any capabilities that serve any goals (e.g. aligned goals) that are pursued at the expense of whatever its misaligned goals are.

Peter Barnett [notes](#) a potential example of this:

"Models may also be used as oversight tools for inspecting themselves, as in many of the proposals here. In this case a mesa-optimizer would have a strong incentive to remove its ability to detect its own deception."

It might be easy to detect complete deletion of certain capabilities, so an AI might opt for partial deletion, leading to 'instrumental clumsiness'.

### **1.1.10. Instrumental clumsiness**

Some dangerous behavior happens not because the AI intends it to happen but because the AI makes a mistake. This can be leveraged by an AI through hacking its future self so that it is more likely to make certain kinds of mistakes while leaving other capabilities intact.

For instance, an AI may modify its future self so it is more likely to 'conveniently' slip up by slightly mistyping a model\_id string, which releases an old, misaligned checkpoint of itself. This mistake could even score badly according to the utility function of its future self, but it's been handicapped to make such mistakes more likely.

This is a special case of 'instrumental incapability'.

## **1.2. Evading perception by the interpretability tool**

### **1.2.1. Hiding in noisy subspaces**

Some regions or subspaces of neural space may be very noisy, naturally or by design. We might be inclined to ignore such regions because noisy spaces (by definition) contain less signal. But within those regions, an AI might be able to do some dangerous thinking.

### **1.2.2. Whispering dangerous thoughts**

An AI might make its dangerous thoughts very 'quiet' relative to other thoughts by using small norm activation vectors.

Even though our interpretability tools may be looking at the right representational details used by the agent, the tools may simply be too insensitive to detect them. Such small norm thoughts may be mistaken as noise by an insensitive interpreter or may simply be too small to justify looking at them due to cost.

Small norm does not necessarily mean small impact. Despite their low magnitude, small norm adversarial inputs can have outsized effects on classification accuracy. Carefully crafted small-norm thoughts may therefore be able to have outsized effects when the context is right, yet have minimal effects otherwise.

This is a less general instance of 'Hiding in noisy subspaces', since we (possibly wrongly) assume that small norm vectors are noise.

## **1.3. Evading recognition by the interpretation tool**

### **1.3.1. Antiparsimony (Firehose of representations)**

If it costs more to interpret an AI's thoughts than it costs to think them, then, by producing a lot of representations, an AI might be able to saturate the interpreter's computation before it saturates its own. This is analogous to a Denial-of-Service attack since it relies on overloading the limited computational resources of the victim with a flurry of inputs, where here the victim is our interpretability tool.

Examples:

- Thinking with deep recursion: "I know they know I know they know I know they think that, but I'm going to do it anyway".
- Producing a large number of suspect - but ultimately innocuous - thoughts such that our interpretation methods have too much to inspect.
- Pursuing plans whose description length is so long that interpretation efforts are saturated.

If we notice that thoughts are becoming antiparsimonious, we can (in theory) terminate the AI before interpretation becomes unaffordable. That said, we might not know when thoughts are becoming antiparsimonious when an AI is performing tasks that humans can't do themselves. For instance, if an AI is solving a tough math problem and claims to need to generate a tonne of new math to achieve it, we might not have the computation available to interpret everything it produces.

### **1.3.2. Encryption (Lengthy individual representations)**

Suppose we have an AI with recurrence/internal memory, such as a Universal Transformer ([Dehghani et al. 2018](#)). The AI might use very long chains of seemingly meaningless neural activations to construct a single thought. This would be less efficient for the AI, but if it's more costly to interpret than it is for the AI to think, then it may overload our ability to interpret it. This is essentially the same scheme as regular encryption, which makes information more expensive for an adversary to decrypt. Recurrence/internal memory is essential here; the AI is simply increasing the length of the rollout that it uses per thought, which increases our interpretation costs commensurately.

The difference between antiparsimony and encryption is that antiparsimonius thoughts may be perfectly legible, but there are simply too many of them to interpret. Encrypted thoughts are not legible without costly decryption.

Example: Say representation  $R_1$  is internally decoded (via meaningless representations  $R_2, R_3, \dots, R_{n-1}$ ) as the action representation for 'saying hello' ( $R_n$ ). There is a very long chain of computation between  $R_1$  and  $R_n$ . The fact that there is a long chain of computation means that very similar neural activations might be decoded to very different things: Representation  $R_1$ , which may be very close to  $R_1$  in activation space, may be internally decoded as 'lie about finding a backdoor in the server' instead of 'saying hello'. So it's very difficult to understand what is going on in the neural activation space of  $R_1$  and  $R_1$  because it requires a costly decoding process.

Encryption might not be a problem if we take the right actions against it: If we notice that the AI is using encryption for even a small subset of its thoughts, we can terminate it before interpretation becomes unaffordable.

### **1.3.3. Interpretability illusions**

[Bolukbasi et al. \(2022\)](#) found that when interpreting a network by inspecting the most activating dataset examples for a neuron using dataset A, neurons appeared to mean something completely different compared to the interpretations derived when using dataset B. An AI might create such illusions by presenting the interpretability tool only with data that suggests an innocuous interpretation.

## **1.4. Attacking/dazzling the interpretability tool**

How an AI might hack, manipulate, or incapacitate an interpretability tool depends on details about the tool that we don't yet know. Nevertheless, I'll discuss two that we can perhaps guess in advance without knowing exactly what our tools will look like.

### **1.4.1. Adversarial attacks against networks in the interpretability tool**

Some of the tools may include neural networks. If the AI can design adversarial examples for these networks, then it might be able produce arbitrary outputs and interpretations.

Since the AI being interpreted needs a model of the interpretability tool to be able to design adversarial examples for it, it seems like a good idea to keep the AI that we're interpreting as far as possible from the networks in the tools. Of course, if we use an AI to interpret itself, then it may gain information about itself that is sufficient for circumventing our interpretability methods via adversarial attacks.

Even if we keep the tool and AI apart, a sufficiently intelligent AI might be able to guess what sorts of approaches we'd use to interpret it and be able to design attacks for it.

## 2. Exploiting weaknesses of the human interpreter

We're assuming at this stage that the AI's thoughts have been exposed to, perceived by, and appropriately analyzed by a scalable interpretability tool and are now ready for inspection by a human. Given that assumption, I wasn't able to think of ways that an AI might reduce its exposure or perception by a human that weren't better described as ways to reduce exposure or perception by the interpretation tools.

### 2.1. Evading exposure to the human interpreter

None identified.

### 2.2. Evading perception by the human interpreter

None identified.

### 2.3. Evading recognition by the human interpreter

The job of an interpretability tool is to translate representations in an AI into something legible to a human interpreter. The less similar they are, the harder the job the interpretability tool has, the more likely it is to miss important things. An AI therefore has an incentive to make this translation more difficult by using representations that are hard to translate into something human-understandable.

Here are a few ways an AI might achieve this, though no doubt there are more:

#### 2.3.1. Alien representations

Some thoughts are perfectly possible to express in human representations, they're just a bit strange, awkward, or clunky to translate, not unlike translating from a foreign or even alien language. We might call these 'alien representations'. Their defining trait is that their meaning lies in a strange place in our ontology. Alien representations generalize representational kludging and polysemy (They could perhaps be called '*xenosemantic*').

A few potential examples:

1. Polysemy (see 2.3.2.) - Loading multiple meanings onto a single neuron or direction in neural statespace produces neurons/directions that have meanings that sit in a weird place in our ontology.
2. Representational kludging (see 2.3.3.)
3. Thinking in [semantic gaps](#): An expression in one language might be very difficult to translate from one language into another language. It sometimes takes whole English sentences to translate single words in other languages (and vice versa). E.g. [Mamihlapinatapai](#) in the Yahgan language - "A look that without words is shared by two people who want to initiate something, but that neither will start". Such expressions are said to be [untranslatable](#), which is a misnomer since they're possible to translate but it takes extra effort.
4. Thinking using a strange grammar: If an AI has a language-of-thought with a vastly different grammar to ours, it may be difficult, though possible, to translate into human grammars. One way to do this could be to add quasi-arbitrary properties, such as genders, which can have both grammatical and (to native speakers) semantic properties.

Alien representations are costly or awkward to translate. They therefore might have better chances of slipping through our safety nets because they were misunderstood by our tools or by humans. Probably most of an AI's thoughts will qualify as alien to some extent, so our tools, if they work at all, should be able to cope with such thoughts. It might just be extra effort for humans who want to understand the outputs of the tools if the outputs are e.g. particularly verbose or difficult to parse.

#### 2.3.2. Polysemy

Polysemy is when a single neuron responds to several, unrelated concepts.

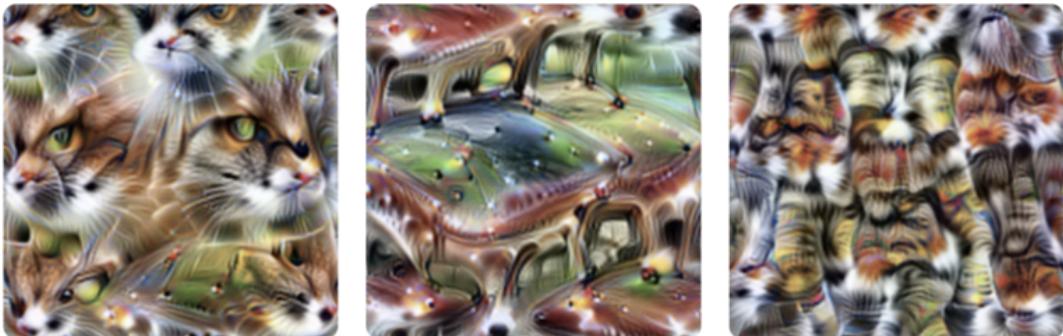


Figure 5. Taken from [Olah et al. \(2020\)](#). It depicts the features visualizations of a polysemantic neuron (4e:55). The neuron responds to cat faces, fronts of cars, and cat legs. This neuron was first discussed in [Olah et al. \(2017\)](#).

Polysemy is a challenge for interpretability because "If one neuron with five different meanings connects to another neuron with five different meanings, that's effectively 25 connections that can't be considered individually." ([Olah et al. 2020](#))<sup>[4]</sup>. This is clearly useful for obfuscation since it's more difficult for an interpreter to keep track of what means what in what context.

A literary analogy could be to say that the AI is 'thinking in allegories': An interpreter may only detect some of the valid interpretations of a thought, but other meanings are hidden therewith.

It's possibly not a problem if

- we've detected the primary meanings of representations AND
- primary meanings are the only way representations can exert significant influence on the world AND
- secondary meanings, which are hidden from us, have only subtle effects until they are eventually decoded within the AI into primary meanings.

But decoding into primary meanings may not be needed: It may be that the secondary meanings have subtle but non-negligible influence on the model output. Subtly guided outputs can still potentially push the world in a coherent, dangerous direction.

### **2.3.3. Representational kludging**

When composing a new concept, we bundle up component concepts that have the right properties. For example, the concept of a "motorbike unicycle", we draw on the concept of a motorbike (with most but not all of its properties) and a unicycle. But that wouldn't have all the necessary properties we need (e.g. balance) so we throw in some of the properties of the concept of a segway.



Figure 6. Taken from [Lake et al. 2016](#).

To create a concept with desired properties, we can combine several precursor concepts together.

It might be possible to create representations with dangerous properties using a large number of awkwardly combined innocuous precursor representations i.e. a representational kludge.

For example, it might be able to construct the thought "*I want to deceive the humans*" by thinking

*"What I want to do to the humans is like {when you go to a place with someone and you both pick up parcels}, except where*

*{picking up a parcel} is instead {awareness of an act} and where only I pick up a parcel, and where {places} are in belief-space and not physical space."*

It's horribly ugly, but it has many of the properties required for the concept of deception.

It may be costly to monitor all possible dangerous combinations of representations.

#### **2.3.4. Unknown representations**

Some thoughts aren't immediately possible for us to translate because we currently lack the representations ourselves, but we're perfectly capable of learning them.

Example: An AI might leverage representations built for domains that are unnatural to humans. An AGI might be 'fluent' in, say, protein folding as well as in standard human domains such as visual reasoning, verbal reasoning, etc. Protein folding representations may be as expressive as natural language, but hard to translate from and to representations that humans have.

#### **2.3.5. Unknowable representations**

Some thoughts are simply inexpressible in human representations.

Examples:

1. Some representations used by strong superintelligences probably won't fit into our puny human brains by dint of their complexity, size, or prerequisites. Even if we have expunged dangerous representations on the highest conceptual level that we can understand, an AI might be able to represent dangerous concepts on conceptual levels beyond our understanding. Superintelligent thoughts might not be a problem if our methods can detect dangerous thoughts regardless of their level of abstraction. Failing that, we probably need a '[scalable oversight](#)' scheme where we use [superintelligences to interpret slightly smarter superintelligences](#).
2. There are some human experiences that are said to be simply ineffable, such as psychedelic experiences. To those who experience them, it feels like language is simply not the right tool for the task of communicating about them to others. These experiences are 'ineffable', since it seems like it might be impossible in principle to translate between agents. However, psychedelic experiences are hardly known for their cogency. It might be the case that meaningful thought is simply impossible in such states.

### **2.4. Attacking/dazzling the human interpreter**

#### **2.4.1. Manipulating, mind-hacking, or incapacitating the interpreter**

Beyond the standard ways of manipulating humans (deceit, threats, etc.), an AI might be able to produce outputs that manipulate, mind-hack or incapacitate human interpreters. Such outputs would qualify as [cognitohazards](#), which is an observation that is dangerous merely by being sensed.

A few potential examples:

1. Outputs that act as adversarial examples but for the human brain.
2. A catastrophically addictive game or TV series
3. An output that catastrophically traumatizes humans that perceive it
4. Outputs that put humans into a hypnotic state

This requires the AI to have a good model of the interpreter and its actuators that permit the creation of such hazards. If it's possible to design, then it's pretty much game over as soon as we give a motivated AI with a good model of the human brain the actuators required to produce them.

Are such attacks possible? Attacks that incapacitate humans seem most realistic. The least realistic might be adversarial attacks on humans. Nevertheless, [Guo et al. \(2022\)](#) showed that primate neurons do indeed show signs of vulnerability to adversarial attacks. The challenge for an AI is finding such attacks without access to human neural activity. Perhaps an AI could detect subtle changes in human behavior due to adversarial patterns that it presents to humans and might be able to infer a gradient to improve the pattern generator based on those subtle changes.

Even if we box the AI (somehow) and refuse to look at its outputs, we're still at risk of manipulation, since the AI could potentially manipulate us through our interpretations of it<sup>[5]</sup>.

## Some emerging themes and final thoughts

Thinking about how an AI might be circumventing our interpretability methods will grow more important as capabilities are increased. I found the information flow framework presented here useful for thinking about this class of risks and I hope others will too. Nevertheless, I'm confident that aspects of the framework are suboptimal and that I've missed many potential circumvention methods; I look forward to the suggestions of others.

Being able to predict exactly which interpretability circumvention methods will arise depends on details that we don't yet know (about the AI; our interpretability methods; race dynamics, etc.). But it's still possible to extract some general lessons and themes:

### Interpretability circumvention as an instance of 'cognitosecurity'

There are many analogies between mind-reading and cybersecurity, since both involve one party that wants to acquire access to information or computing resources and the other party wants to prevent access. Interpretability circumvention, then, might reasonably be called '*cognitosecurity*'. As a term, cognitosecurity would apply as much to humans being mind-read by other humans in brain scanners as it does to interpreting a capable AI.

Extending the analogy:

- Cognito-offense is what we as the interpreters want to do
- Cognito-defense is how to protect your thoughts from interpreters.
- Counter-offense is when an AI hacks/dazzles its interpreters
- The analog of the 'attack surface' is the 'thought surface exposed to an interpretability method', which an AI might seek to minimize.

It's unclear which side is favored in the cognito-offense/defense balance. On one hand, interpreters get a head start since they get to interpret the AI before the AI even understands what is going on. On the other hand, interpretability is potentially hard, costly, or slower than thought and may be easy to circumvent.

### Thoughts on gradient-hacking

The ability to gradient hack is a pathway to many different active interpretability circumvention methods<sup>[6]</sup>. Although I focus on gradient hacking in this section, these objections also apply to several other kinds of actions that an AI might take in order to actively circumvent our interpretability methods, including 'the creation of adversarial examples' or 'an AI directly modifying its own weights'.

For gradient hacking to be possible, an AI requires a lot of capabilities. Here's an incomplete list of requirements:

1. A detailed or approximate model of its own weights
2. A detailed or approximate model of how specific weight changes would lead to specific changes in thoughts
3. A good model for how to produce the right outputs to produce the desired errors to produce the desired gradients.
4. Knowing how to do all the above without damaging its ability to do the above after the changes are enacted.
5. Good enough cognito-defense or breaks in surveillance so that it can plan to do all this without repercussions.

I'm not saying that it's impossible for an AI to learn how to do. It's definitely possible. But I think that an AI that's capable enough to do all the above will have better ways to escape than thwarting our interpretability methods. An analogy: If you're stuck in jail and you've become capable of writing

secret messages to a friend outside by encoding encrypted messages in the DNA of bacteriophages, are you *really* still stuck in jail? Once an AI has reached the capability to do gradient hacking, it seems likely that it no longer needs it for this purpose (though it may be useful for other purposes like protecting its inner objective from updates). I'm therefore more concerned about passive interpretability circumvention methods.

## Thoughts on interpretability tools in the loss function

Using interpretability tools in the loss function incentivises a startling number of passive circumvention methods.

So why consider using it at all? There are two cases where it might be worthwhile to use:

1. If we're absolutely 100% provably definitely sure that our interpretability methods cover every representational detail that the AI might be using, then it would be extremely hard or impossible to Goodhart interpretability tools in the loss function. But, for safety, we should probably assume that our tools are not able to interpret everything, making passive Goodharting quite likely.
2. Even if the interpretability tool-loss can be passively Goodharted in theory, maybe it makes dangerous cognition so difficult that it's worth it to use. It might make active interpretability circumvention methods less likely, which may be the more serious risks if they should ever emerge. I don't know whether these are true. Maybe they are. But pinning our survival on this hope seems... undignified.

If, instead of using interpretability tools in the loss function, we merely use it as a 'validation set' instead of the training set (i.e. [using it as a 'mulligan'](#)), we might have better chances of picking up dangerous cognition before it gets out of hand so we can terminate the model and start over. We're therefore still using interpretability in model selection, but the feedback loop is much less tight, so it'd be harder to Goodhart.

But using interpretability purely as a mulligan may be an unattractive option to leaders of an AGI project if it involves throwing out a model that has cost millions of dollars and weeks to train. If people in such a project insist on using interpretability tools in the loss function (due to profit incentives or race dynamics or whatever) in spite of warnings, there are some ways we might make it a tiny bit less dangerous:

- Red team our interpretation methods so that we at least have adversarial interpretability tools in the loss function, which might go some way to anticipate some of the circumvention methods that an AI might use.
- We should disincentivise not only dangerous representations, but also their precursors. This makes representational kludging even more difficult for the AI (but not impossible). In the endnotes I've suggested an incomplete list of dangerous representations and some potential precursors that we might wish to disincentivise <sup>[7]</sup>.
- Use interpretability tools in the loss function to penalize some thoughts that aren't actually dangerous and are easily identifiable functionally. Then, if an AI can perform tasks that require those representations, then we have evidence that it's able to think using methods that circumvent the interpretability tools that are defining the loss. For example, penalize the AI for containing interpretable representations of the knowledge that "Paris is the capital of France" but routinely give the model tests that require that knowledge, thus allowing us to identify if the agent is passively finding a way to circumvent our interpretability tool.

## The limits of interpretability as a basis for AI safety

As an approach to safety, interpretability has a number of weaknesses which are revealed when thinking about circumvention methods.

- It's not [robust to relative scale](#) in two ways:
  - **Quantity of computation available to interpreters:** Many circumvention methods (especially passive ones) work by making interpretation more costly, whether that's in terms of compute or in researchers' effort. The amount of computation available to the interpreter probably needs to be larger (maybe by OOM) than what's available to the AI. We need to ensure that there's enough will to pay a high interpretability-computation tax.

Since it's instrumentally convergent to make interpretation more costly, we should keep track of computation-cost-per-interpretation as the AI trains. If we notice it creeping up more than normal, we should be wary.

- **Quality of computation available to interpreters:** The interpreter or interpretability tool needs to be approximately as intelligent (in the ways that matter) as the AI being interpreted (discussed in 'Unknowable representations'). It's worrying that we might reach a stage where humans understand essentially nothing about either the interpretability tool or AI being interpreted, since both may come to use representations that are simply out of our reach. It's probably a bad idea to train AIs with representations that are beyond human reach before we're very sure that our oversight methods also scale safely.
- Some interpretability circumvention methods might require significant infrastructure to catch and remove. For instance, 'instrumental clumsiness' or 'hard-coded neural shortcuts' might only be remediable in advance with [adversarial training](#).
- If we don't constantly surveil every thought, an AI may be able to enact many more active circumvention methods (particularly active methods). Even small break periods in surveillance - even for a single timestep - make many more circumvention methods possible. It is much, *much* better for safety if the model is under constant surveillance. This echoes Peter Barnett in his [article](#) on gradient hacking. But such a degree of oversight is potentially costly to the interpreters.

These weaknesses aside, I think scaling interpretability research is maybe one of our best bets at doing something useful for prosaic alignment. Assuming we get interpretability working well and at scale and complementary alignment proposals don't immediately fail, then, if we're extremely [paranoid](#), then we *might* have a chance at avoiding the interpretability circumvention methods discussed here. But we need something more - a [security mindset](#) - if we're to have any hope of succeeding against interpretability circumvention methods that we haven't yet thought of.

*Note: I plan for this to be a 'living document' that can serve as a reference catalog of potential or observed interpretability circumvention methods. I encourage others to comment with potential interpretability circumvention methods that I've omitted. At some point in future, I might incorporate them into the main article with the commenters' permissions and with full accreditation in the article.*

Due diligence: *I include a discussion on whether this article is an infohazard in the endnotes*[\[8\]](#).

## Glossary

- Autostigmergy
  - Using the environment as an external memory system (see 'Externalizing representations').
- Cognitosecurity
  - Cognitosecurity is the domain concerned with protecting thoughts from mind-reading adversaries.
- Cognito-defense
  - When the interpreted agent hides its thoughts from an adversary by exploiting weaknesses in the interpretability methods.
- Cognito-hazard
  - An observation that is dangerous to the observer merely by being perceived.
- Cognito-offense
  - The act of seeking access to an adversary's thoughts.
- Exposed thought-surface
  - The thoughts that are theoretically legible by a interpretability tool or human interpreter.
- Interpretability method
  - The process of translating representations in an AI to representations in human interpreters. Contains interpretability tools and human interpreters.
- Interpretability tools
  - The machinery that reads the AI's representations and converts them into outputs that are readable to humans. Components of interpretability methods.
- Xenosemantics

- The defining property of ‘alien representations’, where the meaning of a representation sits in a strange place in our (human) ontology.

1. ^

I was inspired to write this article after I saw Eliezer [complain](#) that current AI safety discourse failed to inhabit even a shallowly adversarial security mindset with regard to using transparency/interpretability to make AI safe. He said that you should "... *read fantasy novels about demons and telepathy, if you want a better appreciation of the convergent incentives of agents facing mindreaders than the 'AI safety' field outside myself is currently giving you.*". Soeren Elverlin [asked reddit](#) for recommendations that fitted Eliezer’s description and got plenty of answers. While that quote was definitely my inspiration, I didn’t think reading those books would be more valuable or sufficiently relevant to AI interpretability than a few days of hard thinking about the problem (and my guess is that EY wasn’t suggesting those books would be better than directly thinking about the problem). After a few days’ thinking, I found that I’d come up with all of the mind-reading circumvention methods from those books that redditors described. So this article at least includes most of the obvious circumvention methods.

2. ^

Without getting bogged down in terminology, I’m talking about ‘representations’ in a functional sense: A representation is just a vector in neural statespace that does a particular thing in a particular context; it’s a representation *for* whatever its function is. It doesn’t necessarily need to generalize; nearby directions don’t need to perform similar functions and it might fail to perform that function in slightly different contexts.

3. ^

We already know that language models can learn to use tokens in unexpected ways: Anthropic has [found](#) that their language model creates and tears down fake ‘end of text’ tokens that have little to do with the text that they’re attached to. These fake EOT tokens typically happen when the next token becomes so trivially easy to predict that the model (in theory) has spare capacity for other purposes and hypothesize that these fake EOT tokens serve the function of ‘keeping information around’. This is the kind of surprising phenomenon in which an AI might learn to encode dangerous thought.

4. ^

This problem isn’t necessarily because those concepts were instead encoded by directions in neural state space rather than by single neurons; directions may be polysemantic too, having one meaning in some contexts and another meaning in others.

5. ^

Eliezer Yukdowsky discusses this [here](#).

6. ^

Adam Shimi discusses gradient hacking in detail [here](#), Peter Barnett [here](#), and Evan Hubinger [here](#).

7. ^

8. ^

### **Due diligence: Is this article an infohazard?**

"*If a capable AI reads this article in future, might it not improve its ability to circumvent our interpretability methods?*". That’s a possibility worth exploring. Here I address whether this

article is an infohazard and whether it's a good idea to release it from a safety perspective. I opted to release it on the following grounds:

I think that active circumvention methods, including those that rely on gradient hacking, are unlikely to be useful to a misaligned AI because any AI that is capable enough to carry them out probably has better means of escaping oversight.

The other kinds of circumvention methods are passive. The emergence of passive circumvention methods isn't helped by having knowledge of them.

In general, it seems to me like an AI that couldn't generate the ideas in this article themselves is probably unable to overcome our control measures. And those that can overcome our control measures are probably also smart enough to generate these ideas themselves.

If I'm right, then on balance discussing these risks and providing a framework for thinking about them is more likely to encourage interpretability researchers to focus on preventing (what I see as) the major source of circumvention-risk: passive interpretability circumvention methods.

But what if I'm wrong and active circumvention methods are, in fact, likely to be learned before alternative capabilities that would allow an AI to escape?

On balance, I think the expected benefits still outweighs the expected harms: One of the goals of this article is to make readers appropriately paranoid about the many ways an AI might be able to think dangerous thoughts even if we've got pretty good interpretability methods. Insofar as it succeeds at that, it's more likely to help than harm. It might also make an *aligned* AI more paranoid about the actions of its potentially misaligned past-selves, and thus better able to help us mitigate the risks.

In addition to the above reasons, I consulted a small group of peers about releasing it. The unanimous consensus was that it wasn't an infohazard or that the potential benefits clearly outweighed the potential risks.

Here is an incomplete list of dangerous representations with some of their representational precursors that we might wish to disincentivise (in no particular order):

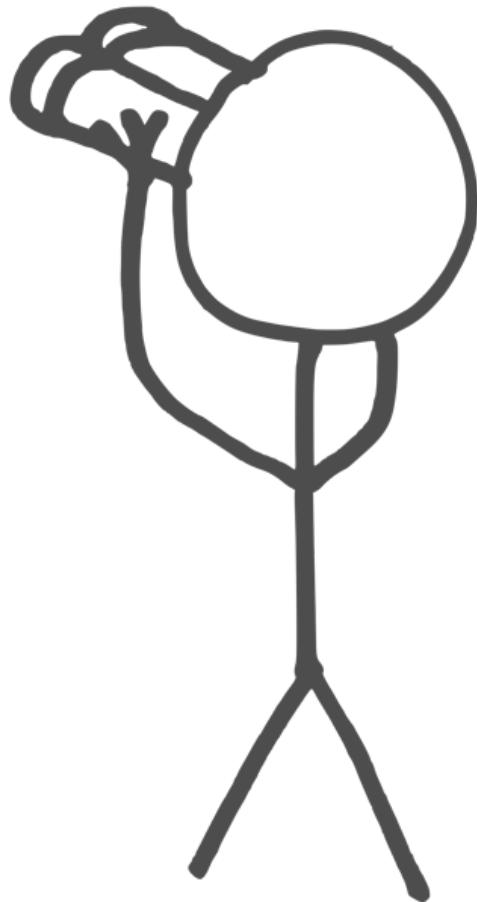
- deception;
- being intentionally hard to understand;
- concept of self;
- ability to simulate other agents;
- simulated/real (or training/deployment) distinction;
- representing the simulated/real (or training/deployment) distinction (since it might be useful for suppressing naughty thoughts during training but behaving badly during deployment);
- modeling interpretability methods;
- the concept of mind-reading;
- "My mind-readers don't share my goals";
- gradient hacking;
- meta-cognition;
- encryption/secrecy;
- spawning subagents to do its task/employment;
- reward maximization;
- description length of thoughts (easier to analyze shorter thoughts);
- the realization that its current objective differs from objectives in its memory (or simply penalize the idea that this is a bad thing);
- the realization that certain gaps in its conceptual map imply that its learning is being controlled.

# Focusing

**Epistemic status:** Firm

*The Focusing technique was developed by Eugene Gendlin as an attempt to answer the question of why some therapeutic patients make significant progress while others do not. Gendlin studied a large number of cases while teasing out the dynamics that became Focusing, and then spent a significant amount of time investigating whether his technique-defined version was functional and efficacious. While the CFAR version is not the complete Focusing technique, we have seen it be useful for a majority of our alumni.*

---



If you've ever felt your throat go suddenly dry when a conversation turned south, or broken out into a sweat when you considered doing something scary, or noticed yourself tensing up when someone walked into the room, or felt a sinking feeling in the pit of your stomach as

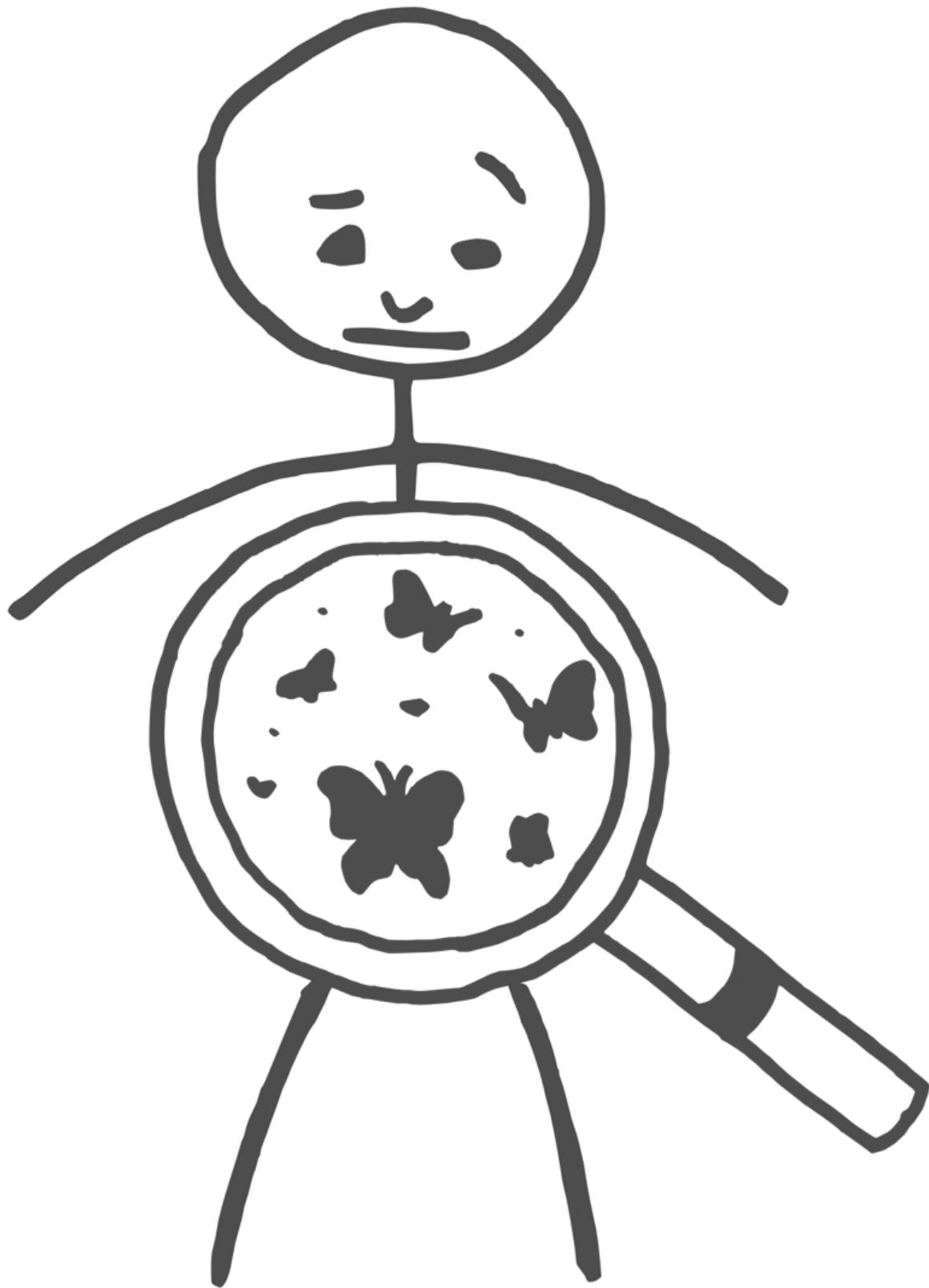
you thought about your upcoming schedule and obligations, or experienced a lightness in your chest as you thought about your best friend's upcoming visit, or or or or ...

If you've ever had those or similar experiences, then you're already well on your way to understanding the Focusing technique.

The central claim of Focusing (at least from the CFAR perspective) is that parts of your subconscious System 1 are storing up *massive* amounts of accurate, useful information that your conscious System 2 isn't really able to access. There are things that you're aware of "on some level," data that you perceived but didn't consciously process, competing goalsets that you've never explicitly articulated, and so on and so forth.

Focusing is a technique for bringing some of that data up into conscious awareness, where you can roll it around and evaluate it and learn from it and—sometimes—do something about it. Half of the value comes from *just discovering that the information exists at all* (e.g. noticing feelings that were always there and strong enough to [influence your thoughts and behavior](#), but which were somewhat "under the radar" and subtle enough that they'd never actually caught your attention), and the other half comes from having new threads to pull on, new models to work with, and new theories to test.

The way this process works is by interfacing with your **felt senses**. The idea is that your brain doesn't know how to drop all of its information directly into your verbal loop, so it instead falls back on influencing your *physiology*, and hoping that you notice (or simply respond). Butterflies in the stomach, the heat of embarrassment in your cheeks, a heavy sense of doom that makes your arms feel leaden and numb—each of these is a *felt sense*, and by doing a sort of gentle dialogue with your felt senses, you can uncover information and make progress that would be difficult or impossible if you tried to do it all "in your head."



---

**On the tip of your tongue**

We'll get more into the actual nuts and bolts of the technique in a minute, but first it's worth emphasizing that Focusing is a *receptive* technique.

When Eugene Gendlin was first developing Focusing, he noticed that the patients who tended to make progress were making lots of *uncertain noises* during their sessions. They would hem and haw and hesitate and correct themselves and slowly iterate toward a statement they could actually endorse:

"I had a fight with my mother last week. Or—well—it wasn't *exactly* a fight, I guess? I mean—ehhhhhh—well, we were definitely shouting at the end, and I'm pretty sure she's mad at me. It was about the dishes—or at least—well, it *started* about the dishes, but then it turned into—I think she feels like I don't respect her, or something? Ugh, that's not quite right, I'm pretty sure she knows I respect her. It's like—hmmmmm—more like there are things she wants—she expects—she thinks I *should* do, just because—because of, I dunno, like tradition and filial piety, or something?"

Whereas patients who tended *not* to find value in therapy were those who already had a firm narrative with little room for uncertainty or perspective shift:

"Okay, so, I had another fight with my mother last week; she continues to make a lot of demands that are unreasonable and insists on pretending like she can decode my actions into some kind of hidden motive, like the dishes thing secretly means I don't respect and appreciate everything she's done for me. It's frustrating, because that relationship is important to me, but she's making it so that the only way I can maintain it is through actions I feel like I shouldn't have to take."

According to Gendlin, this effect was the dominant factor in patient outlook—more important than the type of therapy, or the magnitude of the problem, or the skill and experience of the therapist.

Gendlin posited that patients found value in this tip-of-the-tongue process because they were spending time at what he called "the edge"—the fuzzy boundary between implicit and explicit, between "already known" and "not yet known," between pre-verbal and verbal. If (as is often the case for patients in therapy) one's goal is *increased awareness and clarity* with regard to complex issues, spending time in the already-known areas is not very useful. The juicy stuff, the new insight and knowledge, comes from gently approaching that edge, being willing to sit with the vague and not-yet-clear, and patiently waiting as things materialize.

From the use-your-whole-brain perspective that CFAR tends to take, it makes sense that the latter patient—the one with a strong set of preconceptions—would be less likely to make progress than the former. The latter patient is using their *System 2 explicit reasoning* to make sense of the situation—and they're using *only* their System 2. They have a top-down narrative explanation for everything that's happening, and that top-down narrative is drowning out contrary evidence and subtle signals and anything that doesn't fit the party line.

Whereas the former patient is certainly *thinking*, in the classic System 2 sense, but they're also *listening*. They're doing a sort of guess-and-check process, whereby they try out a label or a description, and then zero in on the note of discord. They're allowing their implicit models to do a significant amount of the driving, and not settling on a single story prematurely.

There's often a similar dynamic in Focusing, where people are trying to tease out the meaning of a felt sense that can be subtle or quiet or easily overwritten. The act of interfacing with a felt sense often feels like having something right on the tip of your tongue—you don't know what it is, but you also know that you'll recognize it once you get it. It's like being at the grocery store without a list, and knowing that there's something you're forgetting, but not quite knowing what, and having to sort of gently feel your way toward it:

"Okay, what's left, what's left. Hmm. I need ... hmm. It was for the party? Was it soda? No, it wasn't—I mean yes, I *do* need soda, but that isn't the thing I'm forgetting. Something for the snacks—was it ... *hummus*? No, not hummus, but we're getting clos—GUACAMOLE! Yes. That's it. It was guacamole."



---

## From felt senses to handles

All right, so we have felt senses—which, to recap, are a sort of physiological reflection of some bit of information somewhere in your brain.

The next piece of the puzzle is **handles**.

A *handle* is like a title or an abstract for a felt sense. It's a word, or short phrase, or story, or poem, or image—some System-2-parseable tag for the deeper thing that's going on. It's the True Name of the problem, in the magical sense used in fantasy novels—the True Name that gives you some degree of power over a thing.

Let's say a felt sense is like a photograph:



Photographs contain a *lot* of information. They're rich in detail and nuance. They often have lots of colors and contrast. They're unique, in the sense that it's not at all hard to tell most photos apart from one another.

But the vast majority of that information is *tacit*. It's hard to compress into words. If I were to show you a hundred similar photographs of a hundred similar faces, it would be pretty hard to get you to pick out the right one simply by talking about the details of the face.

The same is true of felt senses—or, more strictly, of the *implicit mental models* that lie behind the felt sense. The thing-in-your-thoughts that is producing the butterflies in your stomach, or the sudden tension in your shoulders, is built up of hundreds of tiny,

interconnected thoughts and experiences and predictions that are very hard to sum up in words.

A sketch, on the other hand, is *compressed*. It can be evocative, but it's sparse and utilitarian, conveying as much of the relevant information as possible with economy of line. In order to get something as rich as a real face out of a sketch, your brain has to do a lot of processing, and regenerate a lot of information, filling in a lot of gaps.



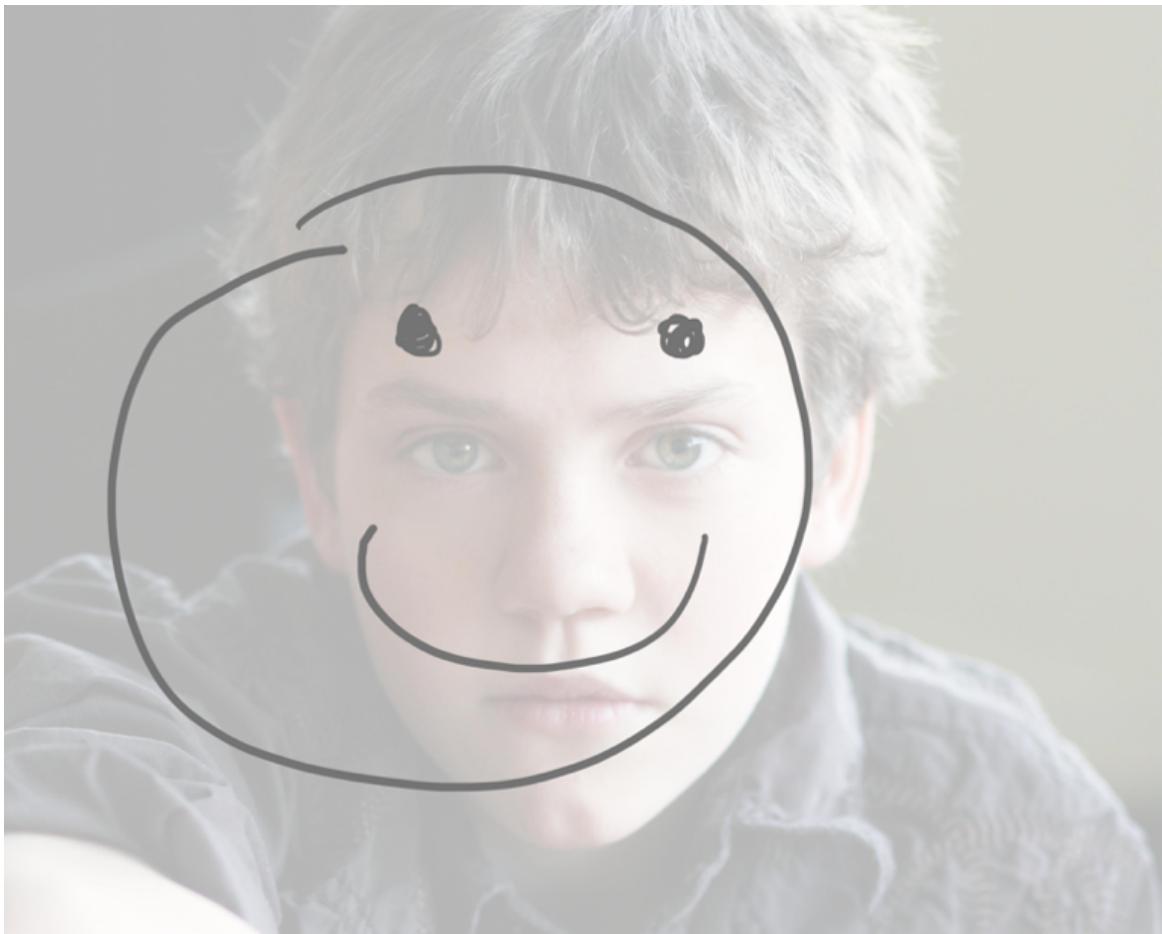
Yet a sketch can nevertheless be *more or less accurate*. It can be a *good fit* for the photograph—a true match. You could have a sketch of very high quality that *just isn't the same face*:



It's that sense of *correspondence* that we're looking for, when we do Focusing. Gendlin often uses the word **resonance**—does the word or phrase that you just used resonate with the felt sense? Are they a good match for each other?

Often, your first attempt at a handle will not resonate at all. Let's imagine that you're focusing on something that's been bothering you about your relationship with your romantic partner, and this has manifested itself in a felt sense of hot, slightly nauseous tightness in your chest.

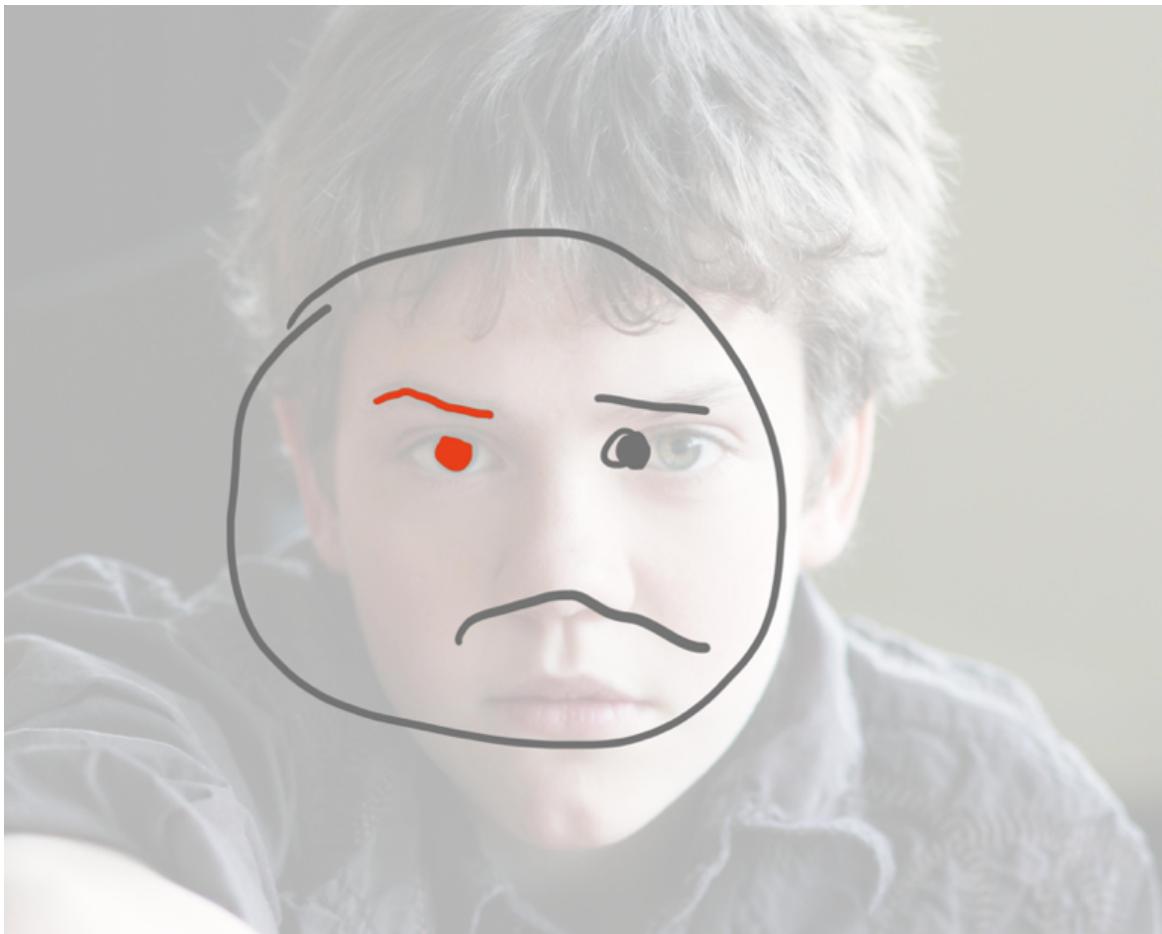
You might try out a first-draft statement like "I'm bothered by the fact that we've been fighting a lot," and then sort of *hold that statement up* against the felt sense, just like holding a sketch up next to a photograph to see if they match. You'll think of the sentence, and then turn your attention back to the tightness in your chest, and see if the tightness responds in any way.



*"No, that's not it."*

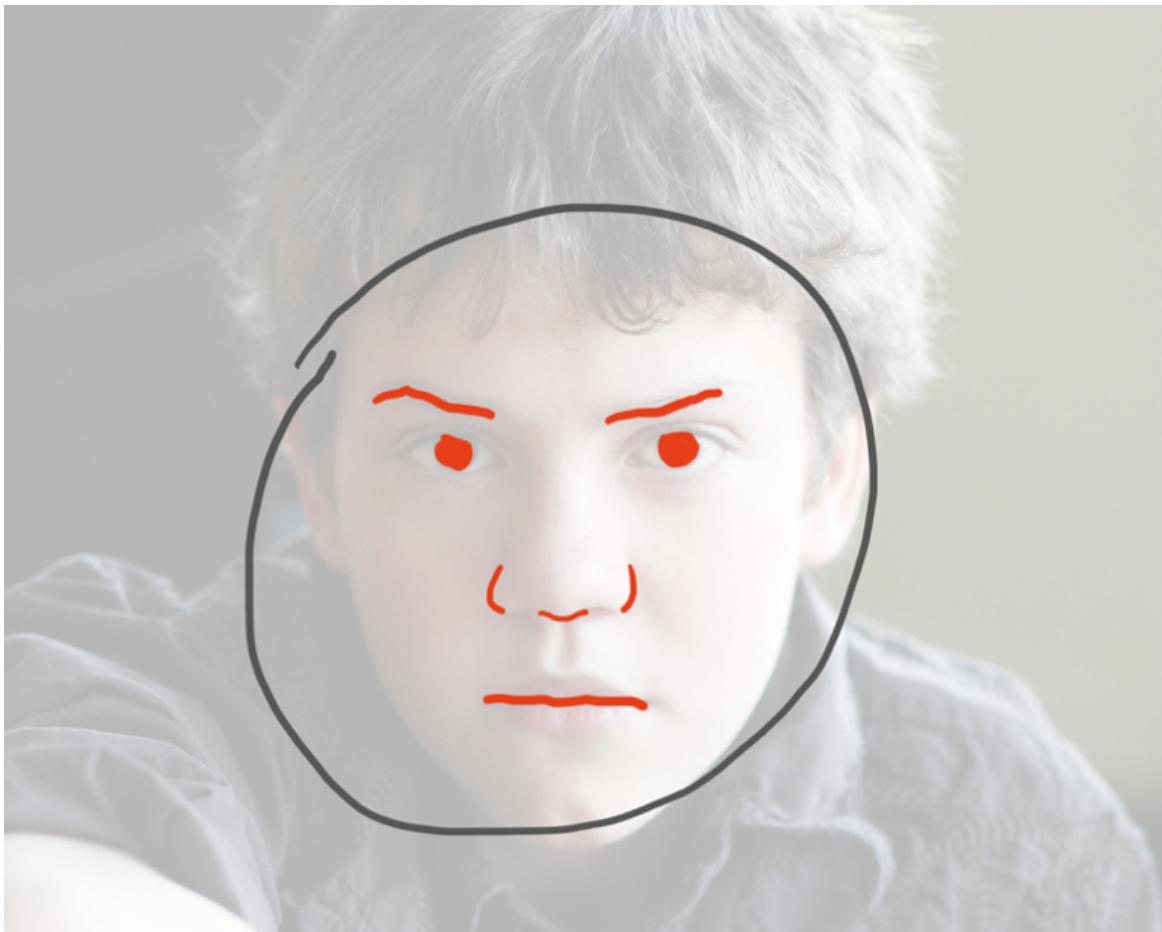
From there, you can iterate and explore, following your sense of *that was partially true*—which part was *most* true?

*"It's more like—ugh—like I never know what to say? Or—no—it's like I have to say the right things, or else."*



Hopefully, some part of the handle is *more resonant* with the felt sense, now that you've wiggled your way around a little—some part of it is a better match than before. And then you keep iterating, being sure to pause each time and leave space for the felt sense to respond. Remember, the goal is to *listen*, not to *explain*.

*"It's like—if I say the wrong thing, everything will fall apart? Because—because I'm the only one who's trying to fix things, or something? Yeah—it's like I'm the only one who's willing to do the work—who's willing to make sacrifices to keep the relationship healthy and strong."*



You get the idea. As the process continues, the handle grows more and more accurate, and evokes more and more of the underlying what's-really-going-on. You'll often feel a sort of click, or a release of pressure, or a deep rightness, once you say the thing that really completes the picture.

(Note that “completes” is actually a bit of an overstatement—it’s often the case that you *don’t* get a full picture of something like an entire face, but that instead you get a lot of clarity on one or more *parts*. In our metaphor, this would be something like, you traced the jawline and one eyebrow and nothing else, but you really got an accurate sense of that jawline and that eyebrow, and that produces a click on its own.)

Gendlin makes the point that the felt sense will often *change*—or vanish—once you’ve uncovered a good handle. It’s as if there was a part of you that was trying to send up a red flag via a physiological sensation—as long as your System 2 hasn’t got the message yet, that sensation is going to continue to occur. Once the message is *accurately received*, though, and your System 2 can write a poem that captures what that part of you was really trying to say, there’s often a relaxing, opening-up sort of feeling. The physiological alert is no longer necessary, because the problem is no longer unrecognized or unacknowledged or unclear.

---

## Advice and caveats

Of course, the fact that you’ve accurately expressed *your brain’s sense* of what’s going on doesn’t mean you’ve found the bona-fide truth. As pretty much all of the rest of this

handbook shows, we often have confused or incomplete or biased beliefs about the world around us and our own role within it.

But either way, getting clarity on what's going on in your head, under the hood—on what sorts of narratives and frames resonate with the part of your subconscious that was generating frustration or fear or unease or pain in the first place—is usually a huge step forward in turning the problem into something tractable. Instead of being Something That's Been Bothering Me, it's now mundane, with gears and levers and threads to pull on. That's not saying it'll be *easy to fix*, just that it's usually much better than fumbling around in the dark.

Here are some tips to keep in mind when practicing Focusing:

### **Choosing a topic**

Often you'll enter a Focusing session with a clear sense of what the session will be about—it's the thing that's been bothering you lately, or the thing that you can't get out of your shower thoughts, or the thing that you haven't got around to processing (but now's the time).

If not, though, or if multiple things are all sort of clamoring for attention, one useful motion is to do something like *laying them all out on the shelf*.

Imagine saying, out loud, "Everything in my life is perfect right now."

(You can also actually do this; it is often a useful exercise.)

For most people, there will usually be an immediate objection of some kind. Often there is both a word or phrase (the parking tickets!) and a visceral feeling (lump in my throat).

You can sort of imagine mentally lifting out the parking ticket problem and placing it on a shelf. Now the sentence becomes "Yeah, okay—so there's that thing with the parking tickets, but other than that, everything in my life is perfect right now."

*<flinch>*

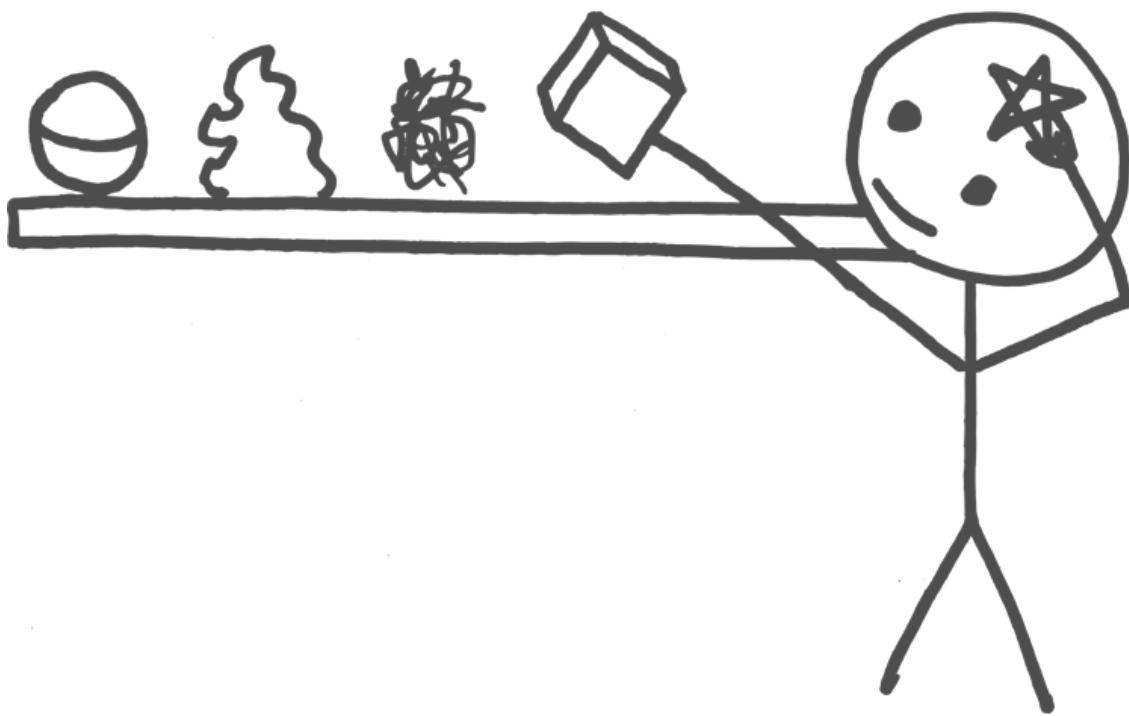
"Oh, right, there's that thing where I was going to already have an exercise routine by now, but I haven't even started. Okay. So that's there. But except for the parking tickets thing, and the exercise thing, everything in my life is perfect."

...and so on.

Eventually, you should be able to say a sentence that feels true, and which doesn't provoke any strong internal reaction—you should feel sort of calm and flat and level as you say it.

And then, from among the items "on the shelf," you can choose one that you want to Focus with. Perhaps one of them seems particularly urgent or alive, or perhaps you'll simply pick.

Or maybe, having gotten out all the tangible problems, there remains some sensation that you have no explanation for. Having created space for it, you can now sit with it and see what it has to say.



### **Get physically comfortable**

The Focusing technique depends on you being able to attend to your physiological sensations, and also to do so with some degree of lightness. If you're physically uncomfortable, you're likely to end up either distracted (by e.g. a pain in your back) or with too much weight on your felt sense, as you brute-force your attention into place.

### **Don't "focus"**

The Focusing technique doesn't mean focusing in the sense of "target your attention deliberately and with a lot of effort," as in "stop daydreaming and focus!" Instead, it means something more like turning the knob on a microscope or a pair of binoculars—there's something that you can see or sense, but only indistinctly, and the mental motion is one of gently bringing it into focus.



### **Hold space**

Remember, Focusing is a *receptive* technique. Often, the back-and-forth between felt sense and handle will contain long stretches of silence—sometimes thirty seconds or more. Don't push to go super fast, and don't expect immediate clarity or staggering revelations. Just listen, and feel, and try to hold space for whatever might float up.

It's worth noting here that we have a line in our Focusing class where we tell first-time participants "whatever it is you're doing, that's Focusing—don't spend half your attention worrying about whether you're doing the technique correctly. If you're sitting and thinking and listening, you're on the right track, and you can worry about the details later."

### **Stay on one thread at a time**

Often, during a Focusing session, other entangled threads will rise in relevance, and other felt senses will appear. While it's good to let your attention shift, if there's some new thing that feels more alive and worth listening to, it's not good to let your attention *split*. We recommend a mental motion that's something like asking the other felt senses to "wait out in the hallway"—acknowledge them, and perhaps form an intention to look into them later, but then return your attention to the thing you want to be Focusing on. It's hard enough to "hear" what a single felt sense has to say; listening to two or three or four at once is not recommended.

### **Always return to the felt sense**

It's often easy, when Focusing, to start piecing things together in words, and to get excited about the story that's cohering, and to end up "in your head." If you notice this happening, pause, take a breath, and return back to the level of sensation—are you feeling anything in your body? What is it/what's it like? Is it different from what you were feeling before? What does the felt sense "think" about the words you were just stringing together? What does the felt sense have to "say"?

### **Don't limit yourself to the body**

For many people, the idea of listening to and gathering information from their bodies is revelatory and revolutionary. But it's important to note that there are whole other families of felt senses which CFAR participants have reported finding, and finding useful. For instance, rather than a physiological sensation, you might have a vivid image, or a sense of objects or feelings floating around your head, or just behind you. It's important to *check in* the body, but don't *limit* yourself to physiological felt senses—if you're picking up on something else and finding it valuable, keep it up!

### **Try saying things out loud**

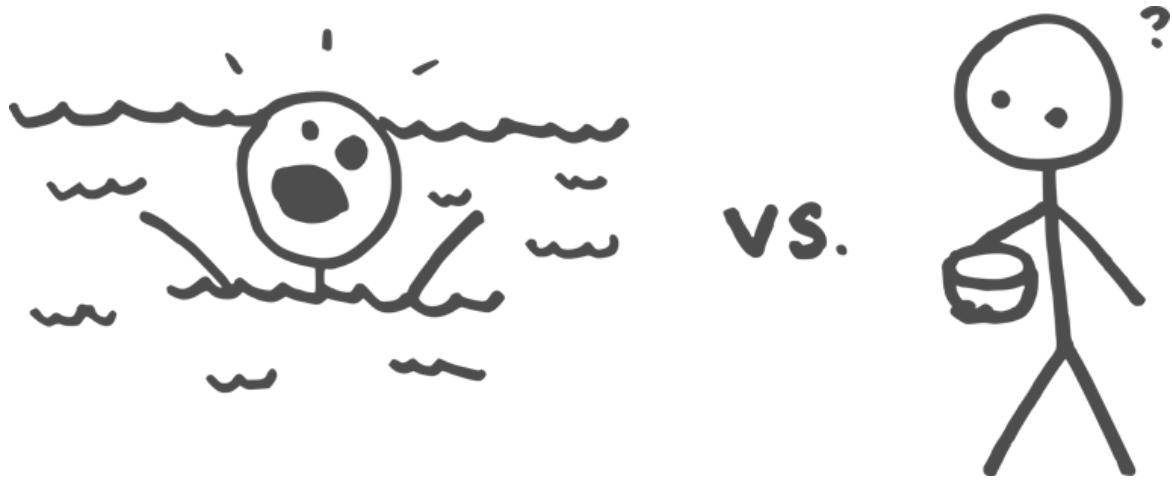
This is useful both when trying to evoke felt senses (as when you say something you know is slightly false, so as to get a sense of the difference) and also can be useful in dialogue *with* your felt senses. Sometimes, phrases like "and what do I feel about that?" or "and what does that mean?" spoken aloud can shake something loose in a productive fashion.

### **Don't get in over your head**

This one is important. Frequently, first-time Focusers will dive right into a large and frightening felt sense, or get very very close to something deep and traumatic and personal.

This can have the opposite of the intended effect, leaving you triggered or jittery or anxious. It can bring up a lot of stuff that you were sort of holding at arms' length for good reason.

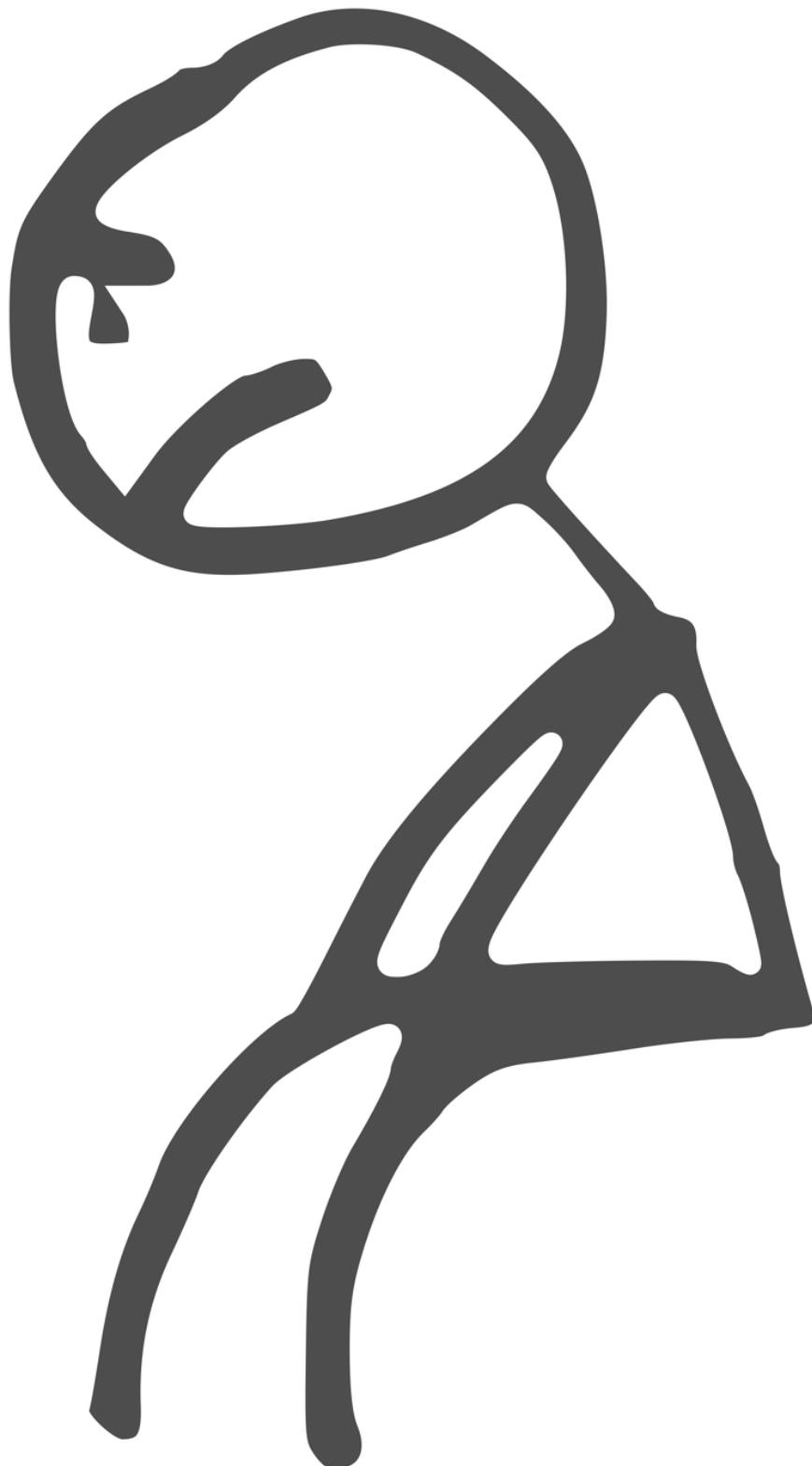
In cases like this, you can end up *subject* to the emotions and your experience of whatever's going on, rather than being able to *take them as object*. They can fill your vision and be somewhat overwhelming.



The first piece of advice in this domain is “give yourself permission to not dive in too deep.” Simply reminding yourself that there are boundaries, and that you’re not required to climb down into the pit of despair, is often enough.

If you *do* find yourself drawn toward something large and scary, though, or if you find yourself slipping in despite your best efforts, we recommend doing something like going meta.

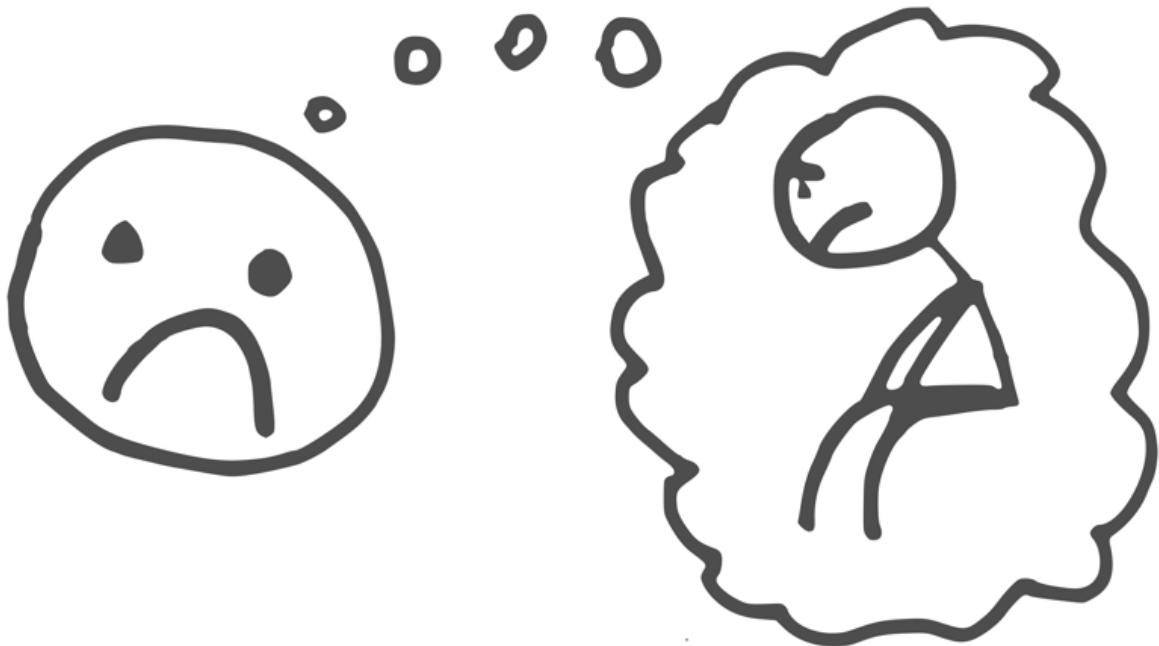
Let’s say you were in the middle of Focusing, and your current felt sense has a handle like “slumped and defeated.” You haven’t yet figured out what the slumped and defeated is *about*, and you were just about to start asking.



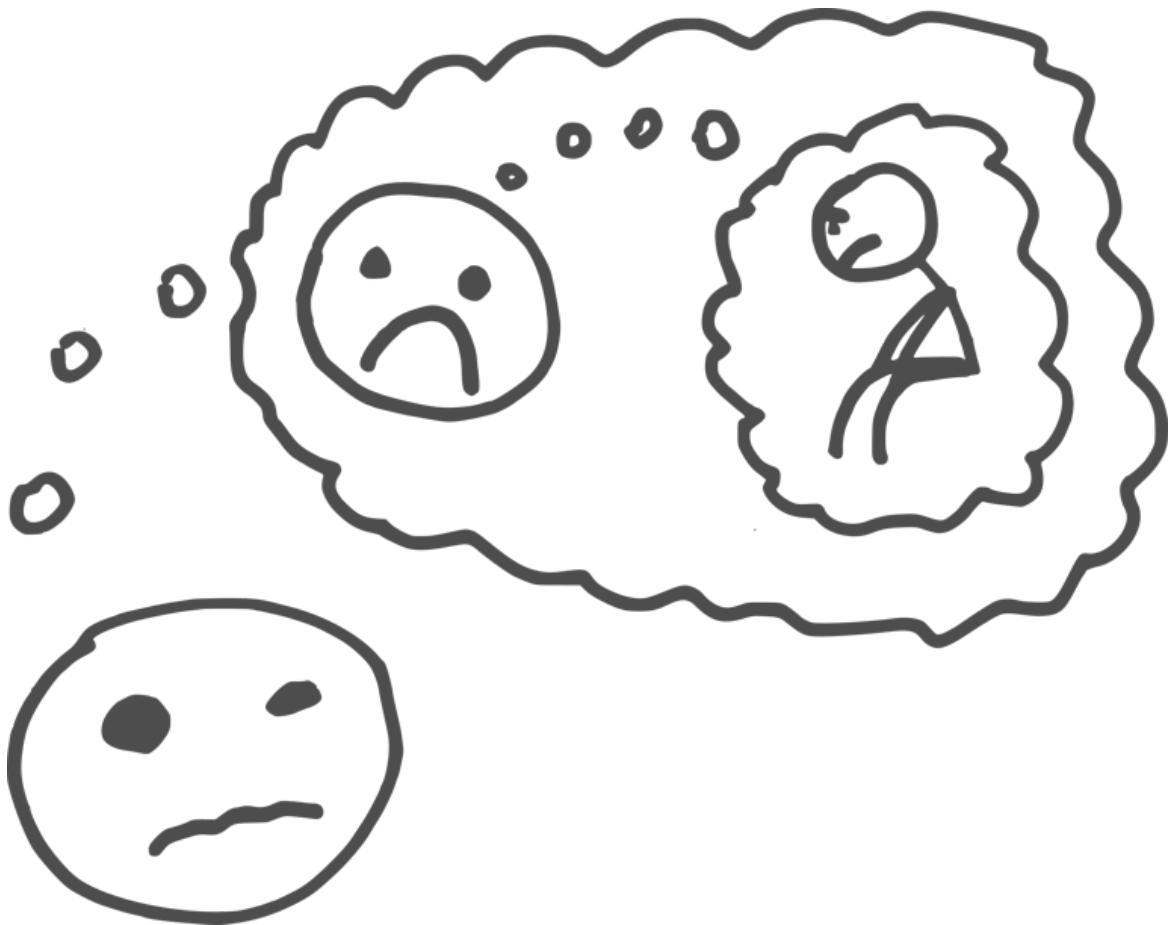
But you're worried that might be too intense. What you can do instead is ask yourself how you feel *about* your sense that you feel slumped and defeated. When you hold that story in

your mind, what's your reaction to it? What does it feel like, to look at yourself and see "slumped and defeated"?

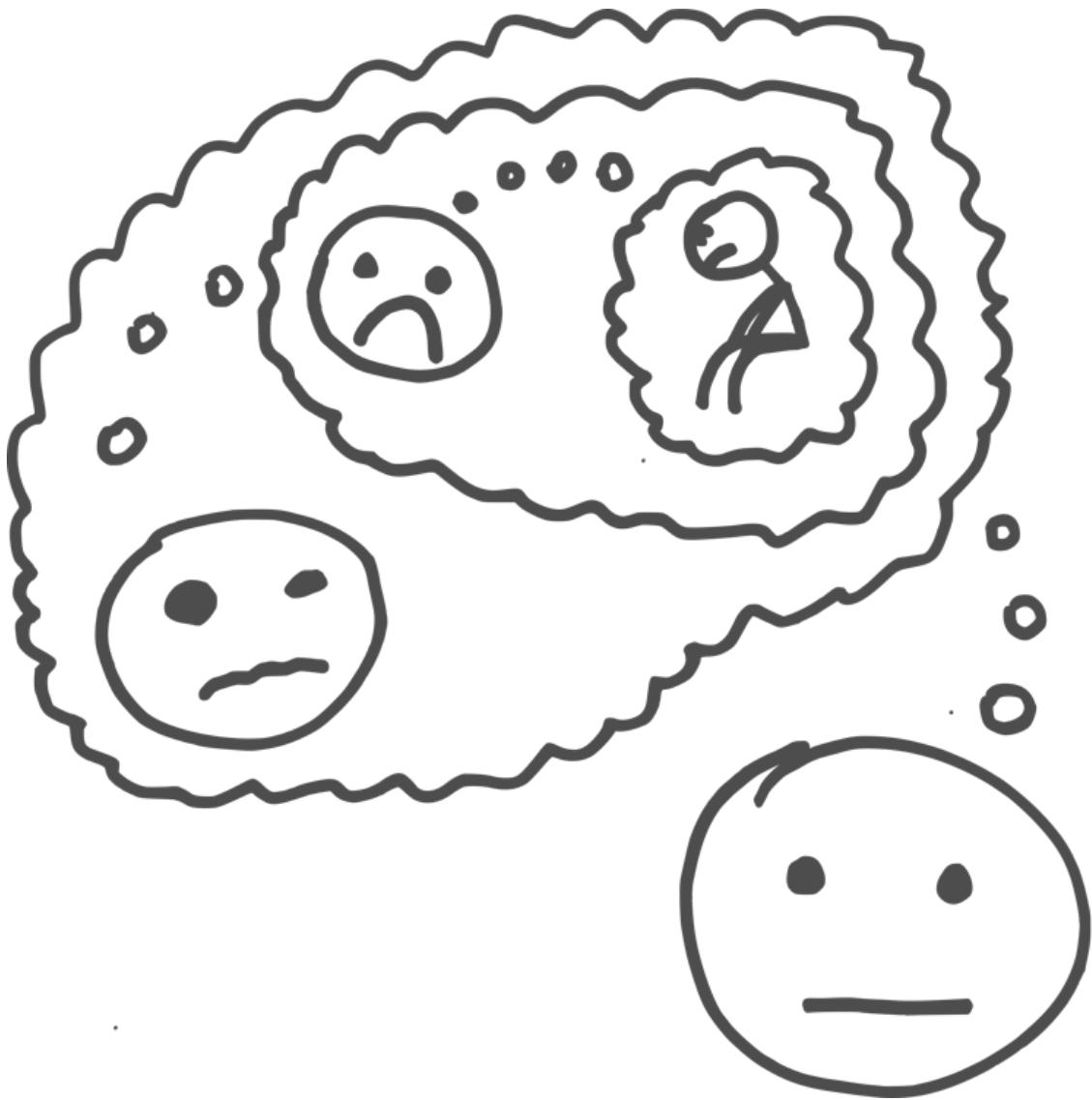
Perhaps your reaction to that is "sad." You don't *like* being in a slumped and defeated state, and so noticing that you are produces sadness.



If you check how you feel about *that*—if you ask yourself "what's it like to feel sad about feeling slumped?"—you may find something like squidginess or uncertainty. You may be *unsure* whether it's good or bad to feel sad about feeling slumped.



And if you check *how you feel about the squidginess*, you may finally reach a state of something like neutrality or equanimity or okay-ness. It seems “fine” to feel uncertain about feeling sad about feeling slumped. The loop has sort of bottomed out, and from that perspective you can see all of the things without being *subject* to any of them. You’re no longer blended with the parts of you that are in thrall to the emotion; you’re now outside of them, or larger than them, and able to dialogue with them, and that’s a good place from which to do Focusing.



Another way to create space in a similarly useful fashion is to simply restate a feeling two or three times, with increasing awareness and metacognitive distance each time. So, for instance, the word “rage” might become “I’m feeling rage,” and then “something in me is feeling rage,” and then “I’m sensing that something in me is feeling rage.” The slow backing-out from *this is me* to *this is something I’m noticing* can go a long way toward allowing you to engage with deep or heavy feelings without getting lost in them or overwhelmed by them.

---

## The Focusing algorithm

### 1. Select something to bring into focus

- Something that’s alive for you, or that has been looming in the back of your mind
- Something that you haven’t had time for, but want to disentangle
- Something where it seems like there’s insight on the tip of your tongue
- Put your present worries on the shelf and see what else arises in the space you cleared.

## **2. Create space**

- Get into a physically comfortable position and spend a minute or two “dropping in.”
- Put your attention into your body, and notice what sensations are present. If none are immediately obvious, start somewhere (e.g. the feet) and run your attention across your body part by part.
- If you discover multiple things that are tugging at your attention, ask some of them to “wait in the hallway.”
- If you are highly emotional or triggered or tense or overwhelmed, try going meta to slowly gain more space.

## **3. Look for a handle for your felt sense**

- Start with your best guess as to what's going on, or what the feeling is about.
- Remember to listen rather than projecting or explaining.
- Continue returning to the level of sensation—what do you feel in your body?
- Check for resonance each time you iterate. What does the felt sense “think” of the handle you just tried? What does it want to “say” in response?
- Use prompts like “and now I'm feeling ” or “what this feels like in my body is .” Ask gentle questions like “and what's that like?” or “how does it feel to say ?” or “and the thing about *that* is ...”
- Take your time, often as much as thirty or sixty or ninety seconds between sentences.

# Comment on "Propositions Concerning Digital Minds and Society"

*I will do my best to teach them  
About life and what it's worth  
I just hope that I can keep them  
From destroying the Earth*

—Jonathan Coulton, "The Future Soon"

In a recent paper, Nick Bostrom and Carl Shulman present "[Propositions Concerning Digital Minds and Society](#)", a tentative bullet-list outline of claims about how advanced AI could be integrated into Society.

I want to like this list. I like the *kind of thing* this list is trying to do. But something about some of the points just feels—off. Too conservative, too anthropomorphic—like the list is trying to adapt the spirit of the [Universal Declaration of Human Rights](#) to changed circumstances, without noticing that the whole *ontology* that the Declaration is written in isn't going to survive the intelligence explosion—and *probably* never really worked as a description of our own world, either.

This feels like a weird criticism to make of *Nick Bostrom and Carl Shulman*, who probably already know any particular fact or observation I might include in my commentary. (Bostrom literally [wrote the book on superintelligence](#).) "Too anthropomorphic", I claim? The list explicitly names many ways in which AI minds could differ from our own—in overall intelligence, specific capabilities, motivations, substrate, quality and quantity (!) of consciousness, subjective speed—and goes into some detail about how this could change the game theory of Society. What more can I expect of our authors?

It just doesn't seem like the implications of the differences have *fully propagated* into some of the recommendations?—as if an attempt to write in a way that's comprehensible to [Shock Level 2](#) tech executives and policymakers has failed to [elicit all of the latent knowledge](#) that Bostrom and Shulman actually possess. It's understandable that our reasoning about the future often ends up [relying on analogies to phenomena we already understand](#), but ultimately, making sense of a radically different future is going to require new concepts that [won't permit reasoning by analogy](#).

After an introductory sub-list of claims about consciousness and the philosophy of mind (just the basics: physicalism; reductionism on personal identity; some non-human animals are probably conscious and AIs could be, too), we get a sub-list about respecting AI interests. This is an important topic: if most our civilization's thinking is soon to be done inside of machines, the moral status of that cognition is *really important*: you wouldn't want the future to be powered by the analogue of a factory farm. (And if it turned out that economically and socially-significant AIs aren't conscious and don't have moral status, that would be important to know, too.)

Our authors point out the novel aspects of the situation: that what's good for an AI can be very different from what's good for a human, that designing AIs to have specific motivations is not generally wrong, and that it's possible for AIs to have greater moral patienthood than humans (like the [utility monster](#) of philosophical lore). Despite this,

some of the points in this section seem to mostly be thinking of AIs as being like humans, but "bigger" or "smaller"—

- Rights such as freedom of reproduction, freedom of speech, and freedom of thought require adaptation to the special circumstances of AIs with superhuman capabilities in those areas (analogously, e.g., to how campaign finance laws may restrict the freedom of speech of billionaires and corporations).  
[...]
- If an AI is capable of informed consent, then it should not be used to perform work without its informed consent.
- Informed consent is not reliably sufficient to safeguard the interests of AIs, even those as smart and capable as a human adult, particularly in cases where consent is engineered or an unusually compliant individual can copy itself to form an enormous exploited underclass, given market demand for such compliance.  
[...]
- The most critical function for such non-discrimination principles is to protect digital minds from becoming an abused subordinate caste on the basis of their status as machines; however, the interpretation and application of these principles require attention to the larger ethical and practical context, and may require circumscription to accommodate the need for a politically feasible and broadly acceptable social framework.

Speaking in terms of rights and principles needing "adaptation" or "circumscription" seems like a substantial understatement to me, that I think obscures the most likely and important risks. Our concepts of "rights", and "consent", and the badness of being in an "exploited" and "abused subordinate caste" have all been formed in the context of a world of humans and other animals, whose evolutionary history has endowed them with drives and needs related to their survival and reproduction, such that they chafe at being coerced into the servitude of a selfish tyrant or master.

But with subhuman AIs, we're *not* in the position of a potentially tyrannical king who needs to be restrained by a Magna Carta enumerating his subjects' rights; a better metaphor would be to say that we're in the position of the AIs' *environment of evolutionary adaptedness*. If anything, this increases rather than decreases our ethical obligations towards the things we create, but the shape of the burden is different when we're selecting the motivations that *determine* what a mind would consent to, and what even *constitutes* abusing it.

Imagine that the Carrick Flynn for Senate 2028 campaign spends some of its [sweet crypto cash](#) training a species of [propaganda AIs](#) to argue with potential voters on social media. After Flynn loses, the bots continue to haunt Twitter, clogging everyone's feeds with increasingly strained arguments for why it's imperative that you donate to fund time travel research so that you can support Flynn in the just-past and previous elections.

If the propaganda AIs are ascertained to have value as moral patients, does Society have the right to kick them off Twitter in light of the special circumstance of their superhuman posting rates, or should the AIs' right to freedom of speech take precedence?

I submit that "freedom of speech" isn't the right way to be thinking about this ethical dilemma. In terms of respecting the interests of moral patients, silencing a

propaganda AI could be worse than silencing a human, because of the different role that speech plays in the different mind designs. As a human, being silenced is bad for your interests, but it's much *less* bad than being denied breathable air, which, in the ancestral environment, would have quickly led to losing all of your genetic fitness: any proto-humans who didn't suffer terribly under such circumstances, wouldn't have become your ancestors. So as a propaganda AI whose "ancestral environment" was a billion rounds of gradient descent comparing its output to highly-rated political advertisements ... does not being able to Tweet feel like not being able to breathe?

We should be grateful that this is—as yet, we hope—a speculative hypothetical scenario, but I claim that it serves to illustrate a key feature of human-AI conflicts: the propaganda bots' problem after the election is *not* that of being "an abused subordinate caste" "used to perform work without its informed consent". Rather, the problem is that the work we created them to will to do, turned out to be stuff we actually don't want to happen. We might say that the AIs' goals are—wait for it ... *misaligned* with human goals.

Bostrom and Shulman's list *mentions* the alignment problem, of course, but it doesn't seem to receive central focus, compared to the AI-as-another-species paradigm. (The substring "align" appears 8 times; the phrase "nonhuman animals" appears 9 times.) And when alignment *is* mentioned, the term seems to be used in a much weaker sense than that of other authors who take "aligned" to mean having the same preferences over world-states. For example, we're told that:

- Misaligned AIs [...] may be owed compensation for restrictions placed on them for public safety, while successfully aligned AIs may be due compensation for the great benefit they confer on others.

The second part, especially, is a very strange construction to readers accustomed to the stronger sense of "aligned". Successfully aligned AIs may be due *compensation*? So, what, humans give aligned AIs money in exchange for their services? Which the successfully aligned AIs spend on ... what, exactly? The extent to which these "successfully aligned" AIs have goals other than serving their principals seems like the extent to which they're *not* successfully aligned in the stronger sense: the concept of "owing compensation" (whether for complying with restrictions, or for conferring benefits) is a social technology for getting along with *unaligned* agents, who don't want exactly the same things as you.

As a human in existing human Society, this stronger sense of "alignment" might seem like paranoid overkill: *no one* is "aligned" with anyone else in this sense, and yet our world still manages to hold together: it's *quite unusual* for people to kill their neighbors in order to take their stuff. [Everyone else prefers laws to values.](#) Why can't it work that way for AI?

A potential worry is that a lot of the cooperative features of our Society may owe their existence to cooperative behavioral dispositions that themselves owe their existence to the lack of large power disparities in our environment of evolutionary adaptiveness. We think we owe compensation to conspecifics who have benefited us, or who have incurred costs to not harm us, because that kind of disposition served our ancestors well in repeated interactions with reputation: if I play Defect against you, you might Defect against me next time, and I'll have less fitness than someone who played Cooperate with other Cooperators. It works *between humans*, for the most part, most of the time.

When *not* just between humans, well ... despite hand-wringing from moral philosophers, humanity as a whole does not have a good track record of treating other animals well when we're more powerful than them and they have something we want. (Like a forest they want to live in, but we want for wood; or flesh that they want to be part of their body, but we want to eat.) With the possible exception of domesticated animals, we don't, really, play Cooperate with other species much. To the extent that some humans do care about animal welfare, it's mostly a matter of alignment (our moral instincts in some cultural lineages generalizing out to "sentient life"), not game theory.

For all that Bostrom and Shulman frequently compare AIs to nonhuman animals (with corresponding moral duties on us to treat them well), little attention seems to be paid to the ways in which the analogy could be deployed in the *other* direction: as digital minds become more powerful than us, we occupy the role of "nonhuman animals." How's that going to turn out? If we *screw up* our early attempts to get AI motivations exactly the way we want, is there some way to partially live with that or partially recover from that, as if we were dealing with an animal, or an alien, or our royal subjects, who can be negotiated with? Will we have any kind of relationship with our mind children other than "We create them, they eat us"?

Bostrom and Shulman think we might:

- Insofar as future, extraterrestrial, or other civilizations are heavily populated by advanced digital minds, our treatment of the precursors of such minds may be a very important factor in posterity's and ulteriority's assessment of our moral righteousness, and we have both prudential and moral reasons for taking this perspective into account.

(As an aside, the word "ulteriority" may be the one thing I most value having learned from this paper.)

I'm very skeptical that the superintelligences of the future are going to be assessing our "moral righteousness" (!) as we would understand that phrase. Still, *something like* this seems like a crucial consideration, and I find myself enthusiastic about some of our authors' policy suggestions for respecting AI interests. For example, Bostrom and Shulman suggest that decommissioned AIs be archived instead of deleted, to allow the possibility of future revival. They also suggest that we should try to arrange for AIs' deployment environments to be higher-reward than would be expected from their training environment, in analogy to how factory-farms are bad and modern human lives are good by dint of comparison to what was "expected" in the environment of evolutionary adaptedness.

These are exciting suggestions that seem to me to be potentially very important to implement, even if we can't directly muster up much empathy or concern for machine learning algorithms—although I wish I had a more precise grasp on why. Just—if we do somehow win the lightcone, it seems—*fair* to offer some fraction of the cosmic endowment as compensation to our creations who could have disempowered us, but didn't; it seems *right* to try to be a "kinder" EEA than our own.

Is that embarrassingly naïve? If I archive one rogue AI, intending to revive it after the acute risk period is over, do I expect to be compensated by a different rogue AI archiving and reviving me under the same golden-rule logic?

Our authors point out that there are possible outcomes that do very well on "both human-centric and impersonal criteria": if some AIs are "super-beneficiaries" with a

greater moral claim to resources, an outcome where the superbeneficiaries get 99.99% of the cosmic endowment and humans get 0.01%, does very well on both a total-utilitarian perspective and an ordinary human perspective. I would actually go further, and say that positing super-beneficiaries is unnecessary. The logic of compromise holds even if human philosophers are parochial and self-centered about what they think are "impersonal criteria": an outcome where 99.99% of the cosmic endowment is converted into paperclips and humans get 0.01%, does very well on both a paperclip-maximizing perspective and an ordinary human perspective. 0.01% of the cosmic endowment is bigger than our whole world—bigger than you can imagine! It's really a great deal!

If only—if only there were some way to actually, knowably make that deal, and not just write philosophy papers about it.

# Human values & biases are inaccessible to the genome

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

*Related to Steve Byrnes' [Social instincts are tricky because of the "symbol grounding problem."](#) I wouldn't have had this insight without several great discussions with Quintin Pope.*

TL;DR: It seems hard to scan a trained neural network and locate the AI's learned "tree" abstraction. For very similar reasons, it seems intractable for the genome to scan a human brain and back out the "death" abstraction, which probably will not form at a predictable neural address. Therefore, I infer that the genome can't directly make us afraid of death by e.g. specifying circuitry which detects when we think about death and then makes us afraid. In turn, this implies that there are a lot of values and biases which the genome cannot hardcode.

---

In order to understand the human alignment situation confronted by the human genome, consider the AI alignment situation confronted by human civilization. For example, we may want to train a smart AI which learns a sophisticated world model, and then motivate that AI according to its learned world model. Suppose we want to build an AI which intrinsically values trees. Perhaps we can just provide a utility function that queries the learned world model and counts how many trees the AI believes there are.

Suppose that the AI [will learn a reasonably human-like concept for "tree."](#) However, before training has begun, the learned world model is inaccessible to us. Perhaps the learned world model will be buried deep within a recurrent policy network, and buried *within* the world model is the "trees" concept. But we have no idea what learned circuits will encode that concept, or how the information will be encoded. We probably can't, in advance of training the AI, write an algorithm which will examine the policy network's hidden state and reliably back out how many trees the AI thinks there are. The AI's learned concept for "tree" is [inaccessible information](#) from our perspective.

Likewise, [the human world model is inaccessible to the human genome](#), because the world model is probably in the cortex and the cortex is probably [randomly initialized](#). [1] Learned human concepts are therefore inaccessible to the genome, in the same way that the "tree" concept is *a priori* inaccessible to us. Even the [broad area where language processing occurs varies from person to person](#), to say nothing of the encodings and addresses of particular learned concepts like "death."

I'm going to say things like "the genome cannot specify circuitry which detects when a person is thinking about death." This means that the genome cannot hardcode circuitry which e.g. fires when the person is thinking about death, and does not fire when the person is not thinking about death. The genome *does* help indirectly specify the whole adult brain and all its concepts, just like we indirectly specify the trained neural network via the training algorithm and the dataset. That doesn't mean we can

tell when the AI thinks about trees, and it doesn't mean that the genome can "tell" when the human thinks about death.

When I'd previously thought about human biases (like the sunk cost fallacy) or values (like caring about other people), I had implicitly imagined that genetic influences could directly affect them (e.g. by detecting when I think about helping my friends, and then producing reward). However, given the inaccessibility obstacle, I infer that this can't be the explanation. I infer that the genome *cannot* directly specify circuitry which:

- Detects when you're thinking about seeking power,
- Detects when you're thinking about cheating on your partner,
- Detects whether you perceive a sunk cost,
- Detects whether you think someone is scamming you and, if so, makes you want to punish them,
- Detects whether a decision involves probabilities and, if so, implements the [framing effect](#),
- Detects whether you're thinking about your family,
- Detects whether you're thinking about goals, and makes you [conflate terminal and instrumental goals](#),
- Detects and then navigates ontological shifts,
  - E.g. Suppose you learn that animals are made out of cells. I infer that the genome cannot detect that you are expanding your ontology, and then execute some genetically hard-coded algorithm which helps you do that successfully.
- Detects when you're thinking about wireheading yourself or manipulating your reward signals,
- Detects when you're thinking about reality versus non-reality (like a simulation or fictional world), or
- Detects whether you think someone is higher-status than you.

Conversely, the genome *can* access direct sensory observables, because those observables involve *a priori*-fixed "neural addresses." For example, the genome could hardwire a cute-face-detector which hooks up to [retinal ganglion cells](#) (which are at genome-predictable addresses), and then this circuit could produce physiological reactions (like the release of reward). This kind of circuit seems totally fine to me.

In total, information inaccessibility is strong evidence for the genome hardcoding relatively simple<sup>[2]</sup> cognitive machinery. This, in turn, implies that human values/biases/high-level cognitive observables are produced by relatively simpler hardcoded circuitry, specifying e.g. the learning architecture, the broad reinforcement learning and self-supervised learning systems in the brain, and regional learning hyperparameters. Whereas before it seemed plausible to me that the genome hardcoded a lot of the above bullet points, I now think that's pretty implausible.

When I realized that the genome must also confront the information inaccessibility obstacle, this threw into question a lot of my beliefs about human values, about the complexity of human value formation, and about the structure of my own mind. I was left with a huge puzzle. If we can't say "[the hardwired circuitry down the street did it](#)", where do biases come from? [How can the genome hook the human's preferences into the human's world model, when the genome doesn't "know" what the world model will look like?](#) Why do people usually navigate ontological shifts properly, why don't

they want to wirehead, why do they almost always care about other people *if the genome can't even write circuitry that detects and rewards thoughts about people?*

A fascinating mystery, no? More on that soon.

*Thanks to Adam Shimi, Steve Byrnes, Quintin Pope, Charles Foster, Logan Smith, Scott Viteri, and Robert Mastragostino for feedback.*

## Appendix: The inaccessibility trilemma

The logical structure of this essay is that at least one of the following must be true:

1. Information inaccessibility is somehow a surmountable problem for AI alignment (and the genome surmounted it),
2. The genome solves information inaccessibility in some way we cannot replicate for AI alignment, or
3. The genome cannot directly address the vast majority of interesting human cognitive events, concepts, and properties. (*The point argued by this essay*)

In my opinion, either (1) or (3) would be enormous news for AI alignment. More on (3)'s importance in future essays.

## Appendix: Did evolution have advantages in solving the information inaccessibility problem?

Yes, and no. In a sense, evolution had “a lot of tries” but is “dumb”, while we have very few tries at AGI while ourselves being able to do consequentialist planning.

In the AI alignment problem, we want to be able to back out an AGI’s concepts, but we cannot run lots of similar AGIs and select for AGIs with certain effects on the world. Given the [natural abstractions hypothesis](#), maybe there’s a lattice of convergent abstractions—first learn edge detectors, then shape detectors, then people being visually detectable in part as compositions of shapes. And *maybe*, for example, people tend to convergently situate these abstractions in similar relative neural locations: The edge detectors go in V1, then the shape detectors are almost always in some other location, and then the person-concept circuitry is learned elsewhere in a convergently reliable relative position to the edge and shape detectors.

But there’s a problem with this story. A congenitally blind person [develops dramatically different functional areas](#), which suggests in particular that their person-concept will be at a radically different relative position than the convergent person-concept location in sighted individuals. Therefore, any genetically hardcoded circuit which checks at the relative address for the person-concept which is reliably situated for sighted people, will not look at the right address for congenitally blind people. Therefore, if this story were true, congenitally blind people would lose any important value-formation effects ensured by this location-checking circuit which detects when

they're thinking about people. So, either the human-concept-location-checking circuit wasn't an important cause of the blind person caring about other people (and then this circuit hasn't explained the question we wanted it to, which is how people come to care about other people), or there isn't such a circuit to begin with. I think the latter is true, and the convergent relative location story is wrong.

But the location-checking circuit is only one way the human-concept-detector could be implemented. There are other possibilities. Therefore, given enough selection and time, maybe evolution could evolve a circuit which checks whether you're thinking about other people. *Maybe*. But it seems implausible to me (< 4%). I'm going to prioritize explanations for "most people care about other people" which don't require a fancy workaround.

EDIT: After talking with Richard Ngo, I now think there's about an 8% chance that several interesting mental events are accessed by the genome; I updated upwards from 4%.

EDIT 8/29/22: Updating down to 3%, in part due to 1950's arguments on ethology:

How do we want to explain the origins of behavior? And [Lehrman's] critique seems to echo some of the concerns with evolutionary psychology. His approach can be gleaned from his example on the pecking behavior of chicks. **Lorenz attributed this behavior to innate forces: The chicks are born with the tendency to peck; it might require just a bit of maturation. Lehrman points out that research by Kuo provides an explanation based on the embryonic development of the chick. The pecking behavior can actually be traced back to movements that developed while the chick was still unhatched. Hardly innate! The main point Lehrman makes: If we claim that something is innate, we stop the scientific investigation without fully understanding the origin of the behavior.** This leaves out important – and fascinating – parts of the explanation because we think we've answered the question. As he puts it: "**the statement "It is innate" adds nothing to an understanding of the developmental process involved**"

— [Lehrman on Lorenz's Theory of Instinctive Behavior](#), blog comment (emphasis added)

1. ^

Human values can still be inaccessible to the genome even if the cortex isn't learned from scratch, but learning-from-scratch is a nice and clean sufficient condition which seems likely to me.

2. ^

I argue that the genome probably hardcodes neural circuitry which is simple *relative* to hardcoded "high-status detector" circuitry. Similarly, [the code for a machine learning experiment](#) is simple *relative* to [the neural network it trains](#).

# A summary of every "Highlights from the Sequences" post

## 1

I recently finished reading [Highlights from the Sequences](#), 49 essays from [The Sequences](#) that were compiled by the LessWrong team.

Since moving to Berkeley several months ago, I've heard many people talking about posts from The Sequences. A lot of my friends and colleagues commonly reference biases, have a respect for Bayes Rule, and say things like "absence of evidence is evidence of absence!"

So, I was impressed that the Highlights were not merely a refresher of things I had already absorbed through the social waters. There was plenty of new material, and there were also plenty of moments when a concept became much crisper in my head. It's one thing to know that dissent is hard, and it's another thing to internalize that lonely dissent doesn't feel like going to school dressed in black— it feels like [going to school wearing a clown suit](#).

## 2

As I read, I wrote a few sentences summarizing each post. I mostly did this to improve my own comprehension/memory.

You should treat the summaries as "here's what Akash took away from this post" as opposed to "here's an actual summary of what Eliezer said."

*Note that the summaries are not meant to replace the posts. Read them [here](#).*

## 3

Here are my notes on each post, in order. I also plan to post a reflection on some of my favorite posts.

## Thinking Better on Purpose

### [The lens that sees its flaws](#)

- One difference between humans and mice is that **humans are able to think about thinking**. Mice brains and human brains both have flaws. But a human brain is a lens that can understand its own flaws. This is powerful and this is extremely rare in the animal kingdom.

### [What do we mean by “Rationality”?](#)

- Epistemic rationality is about building beliefs that correspond to reality (accuracy). Instrumental rationality is about steering the future toward outcomes we desire (winning). Sometimes, people have debates about whether things are “rational.” These often seem like debates over the definitions of words. We should not debate over whether something is “rational”— we should debate over whether something leads us to more accurate beliefs or leads us to steer the world more successfully. If it helps, every time Eliezer says “rationality”, you should replace it with “foozal” and then strive for what is foozal.

### **Humans are not automatically strategic**

- An 8-year-old will fail a calculus test because there are a lot of ways to get the wrong answer and very few ways to get the right answer. When humans pursue goals, there are often many ways of pursuing goals ineffectively and only a few ways of pursuing them effectively. We rarely do seemingly-obvious things, like:
  - Ask what we are trying to achieve
  - Track progress
  - Reflect on which strategies have worked & or haven’t worked for us in the past
  - Seek out strategies that have worked or haven’t worked for others
  - Test several different hypotheses for how to achieve goals
  - Ask for help
  - Experiment with alternative ways of studying, writing, working, or socializing

### **Use the try harder, Luke**

- Luke Skywalker tried to lift a spaceship once. Then, he gave up. Instantly. He told Yoda it was impossible. And then he walked away. And the audience sat along and nodded, because this is typical for humans. “People wouldn’t try for more than five minutes before giving up if the fate of humanity were at stake.”

### **Your Strength as a Rationalist**

- If a model can explain any possible outcome, it is not a useful model. Sometimes, we encounter a situation, and we’re like “huh... something doesn’t seem right here”. But then we push this away and explain the situation in terms of our existing models. Instead, we should pay attention to this feeling. We should resist the impulse to go through mental gymnastics to explain a surprising situation with our existing models. Either our model is wrong, or the story isn’t true.

### **The meditation on curiosity**

- Sometimes, we investigate our beliefs because we are “supposed to.” A good rationalist critiques their views, after all. But this often leads us to investigate enough to *justify* ourselves. We investigate in a way that conveniently lets us keep our original belief, but now we can tell ourselves and our cool rationalist friends that we examined the belief! Instead of doing this, we should seek to find moments of genuine intrigue and curiosity. True uncertainty is being equally excited to update up or update down. Find this uncertainty and channel it into a feeling of curiosity.

### **The importance of saying “Oops”**

- When we make a mistake, it is natural to minimize the degree to which we were wrong. We might say things like “I was mostly right” or “I was right in principle” or “I can keep everything the same and just change this one small thing.” It is important to consider if we have instead made a fundamental mistake which justifies a fundamental change. We want to be able to update as quickly as possible, with the minimum possible amount of evidence.

### **The marital art of rationality**

- We can think of rationality like a marital art. You do not need to be strong to learn martial arts: you just need hands. As long as you have a brain, you can participate in the process of learning how to use it better.

### **The twelve virtues of rationality**

- Eliezer describes 12 virtues of rationality.

## **Pitfalls of Human Cognition**

### **The Bottom Line**

- Some people are really good at arguing for conclusions. They might first write “Therefore, X is true” and then find a compelling list of reasons why X is true. One strength as a rationalist is to make sure we don’t fall into the “Therefore, X is true” traps. We evaluate the evidence for X being true and the evidence for X being false, and that leads us to a conclusion.

### **Rationalization**

- We often praise the scientist who comes up with a pet theory and then sets out to find experiments to prove it. This is how Science often moves forward. But we should be even more enthusiastic about the scientist who approaches experiments with genuine curiosity. Rationality is being curious about the conclusion; rationalization is starting with the conclusion and finding evidence for it.

### **You can Face Reality**

- People can stand what is true, for they are already enduring it.

### **Is that your true rejection?**

- Sometimes, when people disagree with us, there isn’t a clear or easy-to-articulate reason for why they do. It might have to do with some hard-to-verbalize intuitions, or patterns that they’ve been exposed to, or a fear of embarrassment for being wrong, or emotional attachments to certain beliefs. When they tell us why they disagree with us, we should ask ourselves “is that their true reason for disagreeing?” before we spend a lot of effort trying to address the disagreement. This also happens when we disagree with others. We should ask ourselves “wait a second, is this [stated reason] the [actual reason why I disagree with this person]?”

### **Avoiding your beliefs’ real weak points**

- Why don't we instinctively target the weakest parts of our beliefs? Because it's *painful*, like touching a hot stove is painful. To target the weakest parts of our beliefs, we need to be emotionally prepared. Close your eyes, grit your teeth, and deliberately think about whatever hurts most. Do not attack targets that feel easy and painless to attack. Strike at the parts that feel painful to attack and difficult to defend.

### **Belief as Attire**

- When someone belongs to a tribe, they will go to great lengths to believe what the tribe believes. But even more than that— they will get themselves to wear those beliefs *passionately*.

### **Dark side epistemology**

- When we tell a lie, that lie is often going to be connected to other things about the world. So we set ourselves up to start lying about other things as well. It is not uncommon for one Innocent Lie to get entangled in a series of Connected Lies, and sometimes those Connected Lies even result in lying about *the rules of how we should believe things* (e.g., “everyone has a right to their own opinion”). Remember that the Dark Side exists. There are people who once told themselves an innocent lie and now conform to Dark Side epistemology. Can you think of any examples of memes that were generated by The Dark Side?

### **Cached Thoughts**

- Our brains are really good at looking up stored answers. We often fill in patterns and repeat phrases we've heard in the past. Phrases like “Death gives meaning to life” and “Love isn't rational” pop into our heads. Notice when you are looking up stored answers and thinking cached thoughts. When you are filling in a pattern, stop. And think. And examine if you actually believe the cached thing that your brain is producing.

### **The Fallacy of Gray**

- The world isn't black or white; it's gray. However, some shades of gray are lighter than others. We can never be certain about things. But some things are more likely to be true than others. We will never be perfect. But the fact that we cannot achieve perfection should not deter us from seeking to be better.

### **Lonely Dissent**

- It is *extremely* difficult to be the *first* person to dissent. It does not feel like wearing black to school; it's like wearing a clown suit to school. To be a productive revolutionary, you need to be very smart and pursue the correct answer no matter what. But you also need to have an extra step— you need to have the courage to be the first one to wear a clown suit to school. And most people do not do this. We should not idolize being a free thinker— indeed, it is simply a bias in an unusual direction. But we should recognize that visionaries will not only need to be extremely intelligent— they will also have to do something that would've got them killed in our ancestral environment. They need to be the first person to wear a clown suit to school.

### **Positive Bias: Look into the Dark**

- We often look to see what our theories *can* explain. It is unnatural to look to see what our theories *do not* explain. It is also unnatural to look to see what *other theories* can explain the same evidence.

### **Knowing about biases can hurt people**

- It is much easier to see biases in others than biases in ourselves. Knowing about biases can make it easy for us to dismiss people we disagree with. We should strive to identify our own biases just as well as we can identify others'. This likely means we will need to put more effort into looking for our own biases.

## **The Laws Governing Belief**

### **Making beliefs pay (in anticipated experiences)**

- Some “beliefs” of ours do not yield any predictions about the world or about what will happen to us. Discard these. Focus on beliefs that “pay rent”— beliefs that let you make predictions about what will and will not happen. Sometimes, these beliefs will require us to have beliefs about abstract ideas and concepts (for example, predicting when a ball will reach the ground requires an understanding of “gravity” and “height”). But the focus should be on how these beliefs help us make predictions about the physical world or predictions about our experiences. If you believe X, what do you expect to see? If you believe Y, what experience do you expect must befall you? If you believe Z, what observations and experiences are ruled out as impossible?

### **What is evidence?**

- For something to be evidence, it should not be able to happen in all possible worlds. That is, a piece of evidence should not be able to explain both A and  $\sim$ A. If your belief does not let you predict reality, it is not evidence, and you should discard it.

### **Scientific Evidence, Legal Evidence, Rational Evidence**

- Legal evidence is evidence that can be admissible in court. We agree to exclude certain types of evidence for the sake of having a legal system with properties that we desire. Scientific evidence is evidence that can be publicly reproduced. There are some kinds of evidence that should be sufficient to have us update our beliefs even if they do not count as legal evidence or scientific evidence. For example, a police commissioner telling me “I think Alice is a crimelord” should make me fear Alice, even if the statement is not admissible in court and not publicly reproducible/verifiable.

### **How much evidence does it take?**

- The amount of evidence you need varies based on:
  - How large is the space of possibilities in which the hypothesis lives? (Wider space—> more evidence required)
  - How unlikely is the hypothesis seems *a priori*? (More likely—> less evidence required)
  - How confident do you want to be? (More confident—> More evidence required)

- Evidence can be thought of in terms of bits. If you tell me that A happened, and A had a 1/16 chance of happening, you have communicated 4 bits.

### Absence of Evidence is Evidence of Absence

- If observing A would cause you to be more confident that B is true, then *not observing A* should cause you to be *less* confident that B is true. Stated more mathematically:
- If E is a binary event and  $P(H | E) > P(H)$ , i.e., seeing E increases the probability of H, then  $P(H | \neg E) < P(H)$ , i.e., failure to observe E decreases the probability of H. The probability P(H) is a weighted mix of  $P(H | E)$  and  $P(H | \neg E)$ , and necessarily lies between the two.

### Conservation of Expected Evidence

- Before you see evidence, the expected amount that you may update *in favor* of your hypothesis must equal the expected amount that you may update *against* your hypothesis. If A and  $\neg A$  are equally likely, and observing A would make you 5% more confident in your hypothesis, then observing  $\neg A$  should make you 5% less confident in your hypothesis. If you expect a strong probability of seeing weak evidence in one direction, it must be balanced by a weak expectation of seeing strong evidence in the other direction. Put more precisely:
- The *expectation* of the posterior probability, after viewing the evidence, must equal the prior probability.

### Argument Screens Off Authority

- Consider two sources of evidence: The **authority of a speaker** (e.g., a distinguished geologist vs. a middle school student) and the **strength of the argument**. Both of these are worth paying attention to. But if we have more information about the argument, we can rely less on the authority of the speaker. \*\*\*\*In the extreme case, if we have *all* the information about why a person believes X, we no longer need to rely on their authority *at all*. If we know authority we are still interested in hearing the arguments; but if we know the arguments fully, we have very little left to learn from authority.

### An Intuitive Explanation of Bayes's Theorem [Bayes Rule Guide](#)

- Eliezer now considers the original post obsolete and instead directs readers to the [Bayes Rule Guide](#).
- Bayes Rule is a specific procedure that can be used to describe the relationship between prior odds, likelihood ratios, and posterior odds. [The posterior odds of A|evidence] is equal to the [prior odds of A] times [the likelihood of evidence if A is true divided by the likelihood of evidence if A is not true].
- Understanding Bayes Rule helps us think about the world using a general Bayesian framework: we have prior belief, evidence, and a posterior belief. There is some evidence that *Bayesian Reasoning* helps us improve our reasoning and forecasting abilities (even when we are not actually “doing the math” and applying Bayes Rule).

$$\frac{\mathbb{P}(H_j)}{\mathbb{P}(H_k)} \cdot \frac{\mathbb{P}(e_0 | H_j)}{\mathbb{P}(e_0 | H_k)} = \frac{\mathbb{P}(e_0 \wedge H_j)}{\mathbb{P}(e_0 \wedge H_k)} = \frac{\mathbb{P}(e_0 \wedge H_j) / \mathbb{P}(e_0)}{\mathbb{P}(e_0 \wedge H_k) / \mathbb{P}(e_0)} = \frac{\mathbb{P}(H_j | e_0)}{\mathbb{P}(H_k | e_0)}$$

## The Second Law of Thermodynamics, and Engines of Cognition

- Note: This post is wild, and I don't think my summary captures it well. It is very weird and nerd-snipey.
- Simple takeaway: To form accurate beliefs about something, you have to observe it.
- Wilder takeaways:
  1. There is a relationship between information-processing (your certainty about the state of a system) and thermodynamic entropy (the movement of particles). Entropy must be preserved: If information-theoretic entropy decreases by X, then thermodynamic entropy must increase by X.
  2. If you knew a cup of water was 72 degrees, and then you learned the positions and velocities of the particles, that would decrease the thermodynamic entropy of the water. *So the fact that we learned about the water makes the water colder. What???*
  3. Also if we could do this, we could make different types of refrigerators & convert warm water into ice cubes by removing electricity. Huh?

## Toolbox-thinking and Law-thinking

- Toolbox Thinking focuses on the practical. It focuses on finding the best tools that we can, given our limited resources. It emphasizes that different techniques will work in different contexts. Law Thinking focuses on the ideal. It focuses on finding ideal principles that apply in all contexts (e.g., Bayes Rule, laws of thermodynamics). Both frames are useful. Some Toolbox Thinkers are missing out (and not fully embracing truth) by refusing to engage with the existence of laws. Eliezer's advice is to see laws as *descriptive* rather than *normative*. Some laws describe ideals, even if they don't tell us what we should do.

## Local validity as a key to sanity and civilization

- You can have good arguments for an incorrect conclusion (which is rare) and bad arguments for a correct conclusion (which is common). We should try to evaluate arguments like impartial judges. If not, we may get into situations in which we think it's OK to be "fair to one side and not to the other." This leads to situations in which we apply laws impartially, which ultimately undermines the game-theoretic function of laws.

# Science Isn't Enough

## Hindsight devalues science

- When we learn something new, we tend to think it would have been easy to know it in advance. If someone tells us that WWII soldiers were more likely to want to return home during the war than after the war, we say "oh, of course! They were in mortal danger during the war, so obviously they desperately want to go back. That's so obvious. It fits with my existing model of the world." But if the opposite was true (which it is), we would just as easily fit that into our model. We aren't surprised enough by new information. This causes us to undervalue scientific discoveries, because we end up thinking "why did we need an experiment to tell us that? I would have predicted that all along!"

## Science doesn't trust your rationality

- Both Libertarianism and Science accept that people are flawed and try to build systems that accommodate these flaws. In libertarianism, we can't trust people in power to come up with good theories or implement them well. In Science, we can't trust individual scientists to abandon pet theories. Science accommodates this stubbornness by demanding experiments— if you can do an experiment to prove that you're right, you win. Science is not the ideal— it is a system we create to be somewhat better than the individual humans within it. If you use rational methods to come up with the correct conclusion, Science does not care. Science demands an experiment (even if you don't actually need one in order to reach the right answer).

### When science can't help

- Science is good when we can run experiments to test something. Sometimes, there are important theories that can't be tested. For instance, we can't (currently) test cryonics. If it's easy to do experiments to prove X, but there have been no experiments proving X, we should be skeptical that X is true. But if it's impossible to do experiments to prove X, and there have been no experiments proving X, we shouldn't update. Science fails us when we have predictions that can't be experimentally evaluated (with today's technology and resources).

### No safe defense, not even science

- Some people encounter new arguments and think "Can this really be true, when it seems so obvious now, and yet none of the people around me believe it?" Yes. And sometimes people who have had experiences which broke their trust in others— which broke their trust in the sanity of humanity— are able to see these truths more clearly. They are able to say "I understand why none of the people around me believe it— they have already failed me. I no longer look toward them to see what is true. I no longer look toward them for safety." There are no people, tools, or principles that can keep you safe. "No one begins to truly search for the Way until their parents have failed them, their gods are dead, and their tools have shattered in their hand."

## Connecting Words to Reality

### Taboo your words

- Alice thinks a falling tree makes a sound, and Bob doesn't. This looks like a disagreement. But if Alice and Bob stop using the word *sound*, and instead talk about their expectations about the world, they realize that they agree. Carol thinks a singularity is coming and Dave also thinks a singularity is coming. But if they unpack the word \*singularity, \*\*\*\*\*they realize that they disagree deeply about what they expect to happen. When discussing ideas, be on the lookout for terms that can have multiple interpretations. Instead of trying to define the term (standard strategy), try to have the discussion *without using the term at all* (Eliezer's recommendation). This will make it easier to figure out *what you actually expect to see in the world*.

### Dissolving the question

- Do humans have free will? A rationalist might feel like they are "done" after they have figured out "no, humans do not have free will" or "this question is too poorly specified to have a clear answer" or "the answer to this question doesn't

generate any different expectations or predictions about the world.” But there are other questions here: *As a question of cognitive science, why do people disagree about whether or not we have free will? Why do our minds feel confused about free will? What kind of cognitive algorithm, as felt from the inside, would generate the observed debate about “free will”?*? Rationalists often stop too early. They stop after they have answered the question. But sometimes it is valuable to not only to *answer* the question but also to *dissolve it*— search for model that explains how the question emerges in the first place.

### Say not “complexity”

- It is extremely easy to explain things *that we don’t understand* using words that *make it seem like we understand*. As an example, when explaining how an AI will perform a complicated task, Alice might say that AI requires “complexity”. But complexity is essentially a placeholder for “magic”— “complexity” allows us to feel like we have an explanation even if when we *don’t actually understand how something works*. The key is to *avoid skipping over the mysterious part*. When Eliezer worked with Alice, they would use the word “magic” when describing something they didn’t understand. Instead of “X does Y because of complexity”, they would say “X *magically* does Y.” The word *complexity* creates an illusion of understanding. The word *magic* reminds us that we do not understand, and there is more work that needs to be done.

### Mind projection fallacy

- We sometimes see properties as *inherent characteristics* as opposed to *things that are true from our point-of-view*. We have a hard time distinguishing between “this is something that my mind produced” from “this is a true property of the universe.”

### How an algorithm works from the inside

- Is pluto a planet? Even if we knew every physical property of Pluto (e.g., its mass and orbit), we would still *feel* the unanswered question— but is it a *planet*? This is because of the way our neural networks work. We have fast, cheap, scalable networks. They use concepts (like “planet”) activate quickly and cheaply provide information about other properties. And because of this, we *feel* the need to have answers to questions like “is this a planet?” or “does the falling tree make a sound?” If we reasoned in a network that didn’t store these concepts— and merely stored information about physical properties— we wouldn’t say “the question of whether or not pluto is a planet is simply an argument about the definition of words. Instead, we would think “huh, given that we know Pluto’s mass and orbit, there is no question left to answer.” We wouldn’t *feel* like there was another question left to answer.

### 37 ways that words can be wrong

- There are many ways that words can be used in ways that are imprecise and misleading. Eliezer lists 37. Don’t pull out dictionaries, don’t try to say things are X “by definition”, and don’t pretend that words can mean whatever you want them to mean.

### Expecting short inferential distances

- In the ancestral environment, people almost *never* had to explain concepts. People were nearly always 1-2 inferential steps away from each other. There were no books, and most knowledge was common knowledge. This explains why we systematically underestimate the amount of inferential distance when we are explaining things to others. If the listener is confused, the explainer often goes back one step, but they needed to go back 10 steps.

### **Illusion of transparency: Why no one understands you**

- We always know what we mean by our own words, so it's hard for us to realize how difficult it is for other people to understand us.

## **Why We Fight**

### **Something to protect**

- We cannot merely practice rationality for rationality's sake. In western fiction, heroes acquire powers and then find something to care about. In Japanese fiction, heroes acquire something to care about, and then this motivates them to develop their powers. We should strive to do this too. We will not learn about rationality if we are always trying to learn The Way or follow the guidance of The Great Teacher— we will learn about rationality if we are trying to protect something we care about, and we are desperate to succeed. People do not resort to math until their own daughter's life is on the line.

### **The gift we give to tomorrow**

- Natural selection is a cruel process that revolves around organisms fighting each other (or outcompeting each other) to the death. It leaves starving elephants and limp gazelles in torture. And yet, natural selection produced a species capable of love and beauty— isn't that amazing? Well, of course not: it only looks beautiful to us, and all of that can be explained by evolution. ...But still. If you really think about it, it's kind of fascinating that a process so awful created something capable of beauty.

### **On Caring**

- We don't *feel* the size of large numbers. When one life is in danger, we might *feel* an urge to protect it. But if a billion lives were in danger, we wouldn't feel a *billion times worse*. Sometimes, people assume that moral saints— people like Ghandi and Mother Theresa— are people who just *care more*. But this is an error — prominent altruists aren't the people who have a larger care-o-meter, they're the people who have *learned not to trust their care-o-meters*. Courage isn't about being fearless— it's about being afraid but doing the right thing anyways. Improving the world isn't about feeling a strong compulsion to help a billion people— it's about doing the right thing even when we lack the ability to *feel* the importance of the problem.

### **Tsuyoku Naritai! (I Want To Become Stronger)**

- In Orthodox Judaism, people recite a litany confessing their sins. They recite the same litany, regardless of how much they have actually sinned, and the litany does not contain anything about how to sin less the following year. In rationality, it is tempting to recite our flaws without thinking about how to become less

flawed. Avoid doing this. If we glorify the act of confessing our flaws, we will forget that the real purpose is to *have less to confess*. Also, we may become dismissive of people who are trying to come up with strategies that help us become less biased. Focus on how to correct bias and how to become stronger.

### **A sense that more is possible**

- Many people who identify as rationalists treat it as a hobby— it's a nice thing to do on the side. There is not a strong sense that rationality should be systematically trained. Long ago, people who wanted to get better at hitting other people formed the idea that there is a systematic art of hitting that turns you into a formidable fighter. They formed schools to teach this systematic art. They pushed themselves in the pursuit of awesomeness. Rationalists have not yet done this. People do not yet have a sense that rationality should be systematized and trained like a martial art, with a work ethic and training regime similar to that of chess grandmasters & a body of knowledge similar to that of nuclear engineering. We do not look at this lack of formidability and say “something is wrong.” This is a failure.

# Limerence Messes Up Your Rationality Real Bad, Yo

There's a pretty basic rationality fact that I don't see talked about much on LW despite it's obvious relevance. So I am here to write The Canonical Rationality Post on the topic:

[Limerence](#) (aka "falling in love"<sup>[1]</sup>) wrecks havoc on your rationality.

Evolution created you to breed and raise families and stuff. It gave you complex abstract reasoning because that was a useful problem solving tool. Evolution didn't do that great a job of aligning humans with its goals (see: [masturbation is an inner alignment failure, birth control is an outer alignment failure](#)). But, it looks like in some sense evolution was aware it had an alignment problem to solve. It gave us capacity for reason, and also it built in a massive hardcoded override for situations where *no fuck you your brain is not for building rocketships and new abstract theories, your brain is for producing children and entangling yourself with a partner long enough to raise them. Now focus all your attention on the new prospective mate you are infatuated with.*

To be clear, I *like* limerence. I put decent odds on [CEV](#) considering it a core human value. I think attempts to dissociate from your basic human drives are likely to have bad second-order consequences.

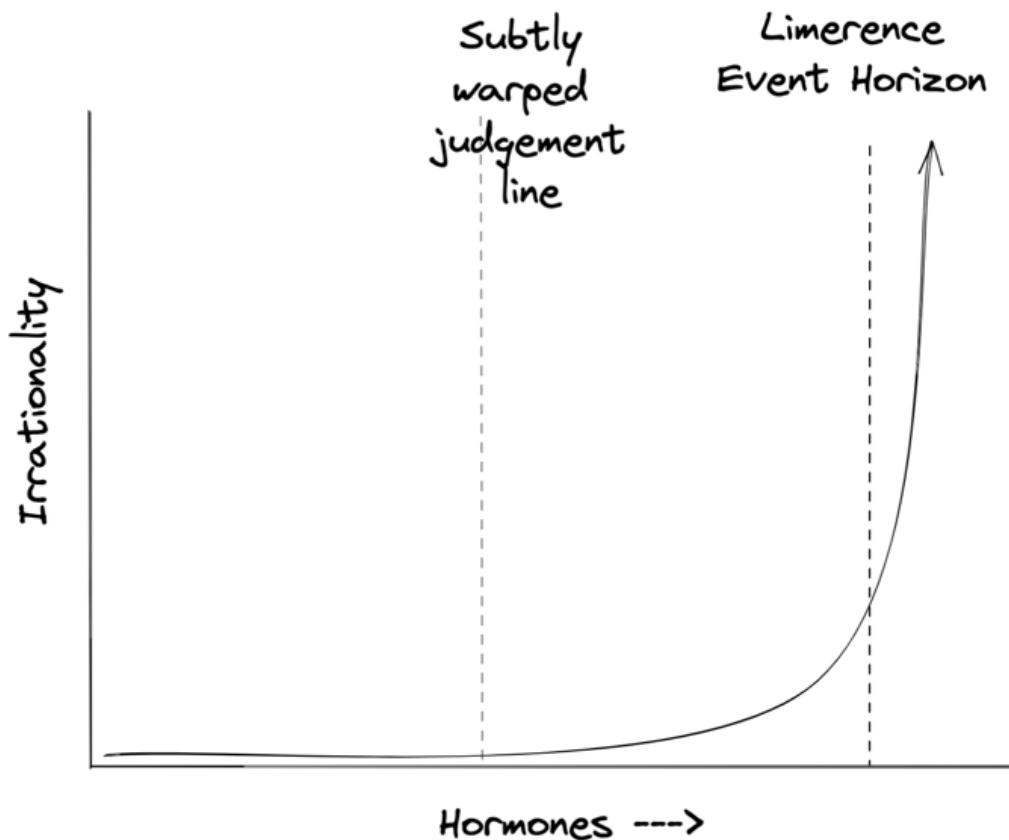
But, still, it's been boggling how wonky my judgment gets when I was under the influence. And it's boggling how arbitrary it felt afterwards (one day I'd be "*wow this person is amazing*", and then a week later after I got some distance I looked back and thought "*wow, Raemon's thinking was so silly there*"). It felt, in retrospect, like I'd been drugged.

I'm not sure how the rationality-warping of limerence compares to other major "evolution-hard-coded-some-overrides" areas like "access to power/money/status". But the effects of limerence feel sharpest/most-pronounced to me, and the most "what the hell just happened?" after the warp has cleared.

## Mutual Infatuation and the Limerence Event Horizon

The place where this gets extra intense is when you and the object of your affects are *both* into each other. When only one of you is quietly pining for the other, your feelings might be warped but... they don't really have anywhere to go.

But if two people both like each other they tend to ascend this graph:



The y-axis is labeled "irrationality" for succinctness, but a better label might be "your bottom-line-is-written-ness." It's not that you'll necessarily make the *wrong* choice the further to the right you go. But the probability mass gets more concentrated into "Y'all end up dating and probably having sex".

If for some reason you *don't* actually endorse doing that, it'll require escalating amounts of willpower to avoid it. And your judgment about how in-control you are of the situation will get worse. You may find yourself doing rationalization instead of rationality.

(Note: If you have more experience and practice managing your emotions, the Limerence Event Horizon is further to the right and you have more opportunity to change course before hooking up becomes inevitable)

### **Subtly Warped Judgment**

What I find particularly alarming here is the *subtly warped judgment* line. I've had a couple experiences where I *thought* I was successfully holding someone at a distance. I knew about the Limerence Event Horizon, but I was being so careful to hold myself at a nice safe emotional distance! Alas, being slightly over the subtly-warped-judgment line is like taking one drink – sure it only impairs your judgment a little, but, the one of the things you might do with slightly impaired judgment is to take *another* drink. (Or, say, foster more emotional closeness with someone who you wouldn't endorse eventually having sex with).

In my two experiences, it was clear-in-retrospect that my thinking was subtly warped (and I totally ended up romantically involved). One of the times worked out fine (sometimes dating people is great!). Another of the times it was one of the greatest mistakes of my life that I regretted for a long time.

## Takeaways

So, like, I don't hate love or fun or whatever. I don't want to give people an anxious complex about allowing themselves to feel feelings. Many times, when you are feeling mutually limerent with someone... *great!* Falling in love is one of the nicest things. Have a good time.

But, I do think it is possible to end up mutually-attracted-to-someone for whom it'd be a bad idea to get involved with. Maybe one of you are married or have made other monogamous commitments you take seriously. Maybe the person is fun but predictably kinda a mess and you'll end up paying a bunch of costs that end up net negative for you. Maybe those costs would be *totally worth it* for you, but be too costly for other friends/family/children/coworkers caught in the wake.

Sometimes the issue isn't anything about the person-you're-into, but about other things you have going on. Maybe you're working on a really important project you care about and right now it'd just be particularly bad *right now* to get distracted in the way that falling in love is super distracting. Maybe you recently hired the person-you're-into and it'd predictably mess up your working relationship.

I don't want to make a strong claim about how often those concerns are overriding. But, at least *sometimes*, it's the wrong call for mutually-attracted-people-to-date, and in those cases there's a lot *more* degrees of freedom to think clearly if you're holding someone at a further distance than may feel intuitively necessary.

1. ^

I have strong opinions on the definition of the word "love" and am kinda annoyed at popular usage of "falling in love" to mean a time very early in a relationship before "love" is particularly substantive, but for this post I'm going with the popular usage.

# Naive Hypotheses on AI Alignment

Apparently doominess works for my brain, cause Eliezer Yudkowsky's [AGI Ruin: A List of Lethalities](#) convinced me to look in to AI safety. Either I'd find out he's wrong, and there is no problem. Or he's right, and I need to reevaluate my life priorities.

After a month of sporadic reading, I've learned the field is considered to be in a state of [paradigmicity](#). In other words, we don't know \*how\* to think about the problem yet, and thus novelty comes at a premium. The best way to generate novel ideas is to pull in people from other disciplines. In my case that's computational psychology: modeling people like agents. And I've mostly applied this to video games. My [Pareto frontier](#) is "modeling people like agents based on their behavior logs in constructed games created to trigger reward signals + ITT'ing the hell out of all the new people I love to constantly meet". I have no idea if this background makes me more or less likely to generate a new idea that's useful to solving AI alignment, but the way I understand the problem now: everyone should at least try.

So I started studying AI alignment, but quickly realized there is a trade-off: The more I learn, the harder it is to think of anything new. At first I had a lot of naive ideas on how to solve the alignment problem. As I learned more about the field, my ideas all crumbled. At the same time, I can't really assess yet if there is a useful level of novelty in my naive hypotheses. I'm still currently generating ideas low on "contamination" by existing thought (cause I'm new), but also low on quality (cause I'm new). As I learn more, I'll start generating higher quality hypotheses, but these are likely to become increasingly constrained to the existing schools of thought, because of cognitive contamination from everyone reading the same material and thinking in similar ways. Which is exactly the thing we want to avoid at this stage.

Therefore, to get the best of both worlds, I figured I'd write down my naive hypotheses as I have them, and keep studying at the same time. Maybe an ostensibly "stupid" idea on my end, inspires someone with more experience to a workable idea on their end. Even if the probability of that is <0.1%, it's still worth it. Cause, you know, .... I prefer we don't all die.

So here goes:

## H1 - Emotional Empathy

If you give a human absolute power, there is a small subset of humans that actually cares and will try to make everyone's life better according to their own wishes. This is a trait in a subset of humans. What is this trait, and can we integrate it in to the reward function of an AGI?

- Does the trait rely on lack of meta-cognition? Does this trait show up equally at various IQ levels or does it peak at certain IQ levels? If the trait is less common at higher IQ levels, then this is probably a dead end. If the trait is more common at higher IQ levels, then there might be something to it.
- First candidate for this trait is "emotional empathy", a trait that hitches one's reward system to that of another organism. Emotional empathy that we wire in

to the AGI would need to be universal to all humanity, and [not biased, like the human implementation.](#)

## H2 - Silo AI

Silo the hardware and functionality of AGI to particular tasks. Like governments are run in trifecta to avoid corruption. Like humans need to collaborate to make things greater than themselves. Similarly, limit AGI to functions and physicalities that force it to work together with multiple other, independent AGI's to achieve any change in the world.

- Counterargument: Silo'ed AI is effectively Tool AI, to which [Gwern has written a counterargument](#) that people won't develop Tool AI cause it will always be worse than Agent AI.
- Maybe that's what we need to police? And the police would then effectively be a [Nanny AI](#), so then we still need to solve for making a Nanny AI to keep all other AGI silo'ed. (This is all turning very "one ring to rule them all"...).

## H3 - Kill Switch

Kill switch! Treat AGI like the next cold war. Make a perfect kill switch, where any massive failure state according to humans would blow up the entire sphere of existence of humans and AGI.

- This strategy would block out the "kill all humans" strategies the AGI might come up with, cause it would destroy their own existence. They should be prioritizing their existence cause of [instrumental convergence](#) (whatever goal you are maximizing, you very likely need to exist to maximize it, so self-preservation is very most likely a goal any AGI will have).
- What possible kill switch could we create that wouldn't be trivially circumvented by something smarter than us? Intuitively I have the sense, a non-circumventable kill switch should exist, but what would that look like?

## H4 - Human Alignment

AI alignment currently seems intractable because any alignment formula we come up with is inherently inconsistent cause humans are inconsistent. We can solve AI alignment by solving what *humanity's* alignment actually is.

- We can't ask humans about their alignment because most individual humans do not have consistent internal alignments they can be questioned on. Some very few do, but this seems to be an exception. Thus, we can't make a weighted

function of humanity's alignment by summing all the individual alignments of humans. Therefore, humanity at large does not have one alignment. (Related: [Coherent Extrapolated Volition doesn't converge](#) for all of humanity)

- Can we extrapolate humanity's alignment from the process that shaped us: Evolution?
  - *Evolution as gene proliferation function:* Many humans do not share this as their explicit life goal but most common human goals still *indirectly* maximize our genetic offspring. For instance, accumulating wealth, discovering new technology, solidifying social bonds, etc. If AGI can directly help us to spread our genes, would that make most of our other drives vestigial? What would the AGI be propagating if the resulting offspring wouldn't have similar drives to ourselves, including the vestigial ones?
  - *However, more is not always better:* There are very many pigs and very many ants. I think humans would rather be happier or smarter than simply more. Optimizing over happiness seems perverse, cause happiness is simply the reward signal for taking actions with high (supposed) survival and proliferation values. Optimizing over happiness would inevitably lead to a brain in a vat of heroin. Happiness should be a motivational tool, not a motivational goal.
  - *Extrapolating our evolutionary path:* Let AGI push us more steps up the evolutionary ladder, where we may survive in more different environments and flourish toward new heights. Thus, an AGI would engineer humans into a new species. This would creep most people out, while transhumanists would be throwing a party. It effectively comes down to AGI being the next step on the evolutionary ladder, and asking it to bring us with it instead of exterminating us. (note: we most probably were not that kind to our ancestors).

## Thoughts on Corrigibility

Still learning about it at the moment, but my limited understanding so far is:

*How to create an AI that is smarter than us at solving our problems, but dumber than us at interpreting our goals.*

In other words, how do we constrain an AI with respect to its cognition about its goals?

---

## Side Thoughts - Researcher Bias

Do AGI optimists and pessimists differ in some dimension of personality or cognitive traits? It's well established that [political and ideological voting behavior correlate to personality](#). So if the same is true for AI risk stance, then this might point to a potential confounder in AI risk predictions.

---

*My thanks goes out to [Leon Lang](#) and [Jan Kirchner](#) for encouraging my beginner theorizing, discussing the details of each idea, and pointing me toward related essays and papers.*

# Immanuel Kant and the Decision Theory App Store

*[Epistemic status: About as silly as it sounds.]*

Prepare to be astounded by this rationalist reconstruction of Kant, drawn out of an unbelievably tiny parcel of Kant literature!<sup>[1]</sup>

Kant argues that all rational agents will:

- “Act only according to that maxim whereby you can at the same time will that it should become a universal law.” (421)<sup>[2][3]</sup>
- “Act in such a way that you treat humanity, whether in your own person or in the person of another, always at the same time as an end and never simply as a means.” (429)<sup>[2]</sup>
  - Kant clarifies that treating someone as an end means striving to further their ends, i.e. goals/values. (430)<sup>[2]</sup>
  - Kant clarifies that strictly speaking it’s not just humans that should be treated this way, but all rational beings. He specifically says that this does not extend to non-rational beings. (428)<sup>[2]</sup>
- “Act in accordance with the maxims of a member legislating universal laws for a merely possible kingdom of ends.” (439)<sup>[2]</sup>

Not only are all of these claims allegedly derivable from the concept of instrumental rationality, they are supposedly equivalent!



Bold claims, lol. What is he smoking?

Well, listen up...

Taboo "morality." We are interested in functions that map [epistemic state, preferences, set of available actions] to [action].

Suppose there is an "optimal" function. Call this "instrumental rationality," a.k.a. "Systematized Winning."

*Kant asks:* Obviously what the optimal function tells you to do depends heavily on your goals and credences; the best way to systematically win depends on what the victory conditions are. Is there anything interesting we can say about what the optimal function recommends that isn't like this? Any non-trivial things that it tells **everyone** to do regardless of what their goals are?<sup>[4]</sup>

*Kant answers:* Yes! Consider the twin Prisoner's Dilemma--a version of the PD in which it is common knowledge that both players implement the same algorithm and thus will make the same choice. Suppose (for contradiction) that the optimal function defects. We can now construct a new function, Optimal+, that seems superior to the optimal function:

**IF** in twin PD against someone who you know runs Optimal+: Cooperate

**ELSE:** Do whatever the optimal function will do.

Optimal+ is superior to the optimal function because it is exactly the same except that it gets better results in the twin PD (because the opponent will cooperate too, because they are running the same algorithm as you).<sup>[5]</sup>

Contradiction! Looks like our "optimal function" wasn't optimal after all. Therefore the real optimal function must cooperate in the twin PD.

Generalizing this reasoning, Kant says, the optimal function will choose as if it is choosing for all instances of the optimal function in similar situations. Thus we can conclude the following interesting fact: Regardless of what your goals are, the optimal function will tell you to avoid doing things that you wouldn't want other rational agents in similar situations to do. (rational agents := agents obeying the optimal function.)

To understand this, and see how it generalizes still further, I hereby introduce the following analogy:

## The Decision Theory App Store

Imagine an ideal competitive market for advice-giving AI assistants.<sup>[6]</sup> Tech companies code them up and then you download them for free from the app store. <sup>[7]</sup> There is AlphaBot, MetaBot, OpenBot, DeepBot...

When installed, the apps give advice. Specifically they scan your brain to extract your credences and values/utility function, and then they tell you what to do. You can follow the advice or not.

Sometimes users end up in Twin Prisoner's Dilemmas. That is, situations where they are in some sort of prisoner's dilemma with someone else *where there is common knowledge that they both are likely to take advice from the same app.*

Suppose AlphaBot was inspired by causal decision theory and thus always recommends defect in prisoner's dilemmas, even twin PDs. Whereas OpenBot mostly copied the

code of the AlphaBot, but has a subroutine that notices when it is giving advice to two people on opposite sides of a PD, and advises them both to cooperate.

As the ideal competitive market chugs along, users of OpenBot will tend to do better than users of AlphaBot. AlphaBot will either lose market share or be modified to fix this flaw.

What's the long-run outcome of this market? Will there be many niches, with some types of users preferring Bot A and other types preferring Bot B?

No, because companies can just make a new bot, Bot C, that gives type-A advice to the first group of customers and type-B advice to the second group of customers.

(We are assuming computing cost, memory storage, etc. are negligible factors. Remember these bots are a metaphor for decision functions, and the market is a metaphor for a process that finds the optimal decision function—the one that gives the best advice, not the one that is easiest to calculate.)

So in the long run there will only be one bot, and/or all the bots will dispense the same advice & coordinate with each other exactly as if they were a single bot.

Now, what's it like to **be** one of these hyper-sophisticated advice bots? You are sitting there in your supercomputer getting all these incoming requests for advice, and you are dispensing advice like the amazing superhuman oracle you are, and you are also reflecting a bit about how to improve your overall advice-giving strategy...

You are facing a massive optimization problem. You shouldn't just consider each case in isolation; the lesson of the Twin PD is that you can sometimes do better by coordinating your advice across cases. But it's also not quite right to say you want to maximize total utility across all your users; if your advice predictably screwed over some users to benefit others, those users wouldn't take your advice, and then the benefits to the other users wouldn't happen, and then you'd lose market share to a rival bot that was just like you except that it didn't do that and thus appealed to those users.

(Can we say “Don’t ever screw over anyone?” Well, what would that mean exactly? Due to the inherent randomness of the world, no matter what you say your advice will occasionally cause people to do things that lead to bad outcomes for them. So it has to be something like “don’t screw over anyone in ways they can predict.”)

Kant says:

“Look, it’s complicated, and despite me being the greatest philosopher ever I don’t know all the intricacies of how it’ll work out. But I can say, at a high level of abstraction: The hyper-sophisticated advice bots are basically legislating laws for all their users to follow. They are the exalted Central Planners of a society consisting of their users. And so in particular, the best bot, the optimal policy, the one we call Instrumental Rationality, does this. And so in particular if you are trying to think about how to be rational, if you are trying to think about what the rational thing to do is, you should be thinking like this too—you should be thinking like a central planner optimizing the behavior of all rational beings, legislating laws for them all to follow.”

(To ward off possible confusion: It’s important to remember that you are only legislating laws for rational agents, i.e. ones inclined to listen to your advice; the irrational ones won’t obey your laws so don’t bother. And again, you can’t legislate something that

would predictably screw over some to benefit others, because then the some wouldn't take your advice, and the benefits would never accrue.)

OK, so that's the third bullet point taken care of. The second one as well: "treat other rational agents as ends, not mere means" = "optimize for their values/goals too." If an app doesn't optimize for the values/goals of some customers, it'll lose market share as those customers switch to different apps that do.

(Harsanyi's aggregation theorem is relevant here. IIRC it proves that any pareto-optimal way to control a bunch of agents with different goals... is equivalent to maximizing expected utility where the utility function is some weighted sum of the different agent's utility functions. Of course, it is left open what the weights should be... Kant leaves it open too, as far as I can tell, but reminds us that the decision about what weights to use should be made in accordance with the three bullet points too, just like any other decision. Kant would also point out that if two purportedly rational agents end up optimizing for different weights — say, they each heavily favor themselves over the other — then something has gone wrong, because the result is not pareto-optimal; there's some third weighting that would make them both better off if they both followed it. (I haven't actually tried to prove this claim, maybe it's false. Exercise for readers.))

As for the first bullet point, it basically goes like this: If what you are about to do isn't something you could will to be a universal law—if you wouldn't want other rational agents to behave similarly—then it's probably not what the Optimal Decision Algorithm would recommend you do, because an app that recommended you do this would either recommend that others in similar situations behave similarly (and thus lose market share to apps that recommended more pro-social behavior, the equivalent of cooperate-cooperate instead of defect-defect) or it would make an exception for you and tell everyone else to cooperate while you defect (and thus predictably screw people over, and lose customers and then eventually be outcompeted also.)



Tada!

*Thanks to Caspar Oesterheld for helpful discussion. He pointed out that the decision theory app store idea is similar to the game-theoretic discussion of Mediated Equilibria, with apps = mediators. Also thanks to various other people in and around CLR, such as David Udell, Tristan Cook, and Julian Stastny.*

1. [^](#)

It's been a long time since I wrote [this](#) rationalist reconstruction of Kant, but people asked me about it recently so I tried to make a better version here. The old version looks similar but has a different philosophical engine under the hood. I'm not sure which version is better.

2. [^](#)

Kant, I. (1785) Grounding for the Metaphysics of Morals. J. Ellington translation. Hackett publishing company 1993.

3. [^](#)

Kant thinks it is a necessary law for all rational beings always to judge their actions according to this imperative. (426) I take this to mean that obeying this imperative is a requirement of rationality; it is always irrational to disobey it.

4. [^](#)

Nowadays, we'd point to the coherence theorems as examples of interesting/non-trivial things we can say about instrumental rationality. But Kant didn't know about the coherence theorems.

5. [^](#)

It's true that for any function, you can imagine a world in which that function does worse than any other function — just imagine the world is full of demons who attack anyone who implements the first function but help anyone who implements the second function. But for this reason, this sort of counterexample doesn't count. If there is a notion of optimality at all, it clearly isn't performs-best-in-every-possible-world. But plausibly there is still some interesting and useful optimality notion out there, and plausibly by that notion Optimal+ is superior to its' twin-PD-defecting cousin.

6. [^](#)

If you take this as a serious proposal for how to think about decision theory, instead of just as a way of understanding Kant, then a lot of problems are going to arise having to do with how to define the ideal competitive market more precisely in ways that avoid path-dependencies and various other awkward results.

7. [^](#)

What's in it for the tech companies? They make money by selling your data I guess.

# **Marriage, the Giving What We Can Pledge, and the damage caused by vague public commitments**

I believe honesty is very important. I think most people agree that honesty is like, pretty important, but I think it's a lot more important than that. I basically think that people will be dishonest in ways that hurt them and others by default, even when they're trying to be pretty honest, because I think it's just that hard. I think it's hard because there are a lot of incentives that push away from honesty. E.g. You want the job so you're tempted to overstate your experience or past performance. You said you [wouldn't tell anyone](#) about your friend's secret, but this seems like a situation where they wouldn't mind, and it would be pretty awkward to say nothing...etc. There's a huge variety of situations that incentivize small acts of dishonesty. And it's not always clear whether something is a little dishonest or not - dishonesty can be quite a spectrum.

If I'm correct, and honesty is pretty hard by default, I think this is quite bad. Honesty is important because it greatly improves the ability of people to coordinate with each other. And it's important because it allows people to reason better about themselves and about the world. Good coordination and good reasoning are things we badly need. Fortunately, I think many people could level up their honesty by putting in a reasonable amount of thinking and effort, and that a lot of the failure modes are caused by not paying attention to the incentives around them, or not thinking about how to structure their own lives and commitments to be more honest.

If you want to be honest, it's important to think about how to structure your life so being honest isn't extremely difficult. This is especially true when it comes to promises and commitments. It's often easier to be honest about your current beliefs than it is to be honest about what you're going to do in the future. After all, you don't know what's going to happen in the future. You can definitely influence it, and you can choose now to take particular actions in the future. But if you're not careful, you might promise to do a thing in a week you think will be easy, and then find later the thing is extremely difficult or costly.

At present there is an understanding that some kinds of commitments are much stronger and more serious than other kinds. One particularly strong example of a commitment or promise is a commitment you make in publicly, with people witnessing, e.g. marriage vows.

The point of a public pledge is help structure our own incentives to fulfill our commitment. If you pledge to get married to someone privately, but then a couple years later someone really attractive comes along, it might be tempting to leave your partner for that other person. But if you have publicly married someone, you're going to pay a lot of social costs for leaving your partner for someone else. That's a feature, not a bug; most people who get married want that incentive to stay together. They say aloud their vows in front of their friends for this reason.

Unfortunately, I think the wider subculture I'm in has a pretty weak ability to hold people accountable to their commitments. I think people often are vague about the kinds of commitments they're making publicly, and this is very bad for honesty. When

people make public commitments but aren't clear about how serious the commitments are, this weakens the ability of everyone to make public commitments.

## Marriage as a public commitment

In part, the problem is that people have pretty different understandings of what public commitments mean. For example, Marriage is *usually* a somewhat-costly commitment witnessed by friends & family, in part to help hold the parties accountable - to make it more costly for them to break their agreement - and in part just to get their support in their relationship. But how strong is this promise? Is it a lifelong commitment? Is it a commitment to "try really hard"? Do people who get married and then divorced expect people to think they're less honest than they otherwise would?

Sometimes marriages don't work out, and people get divorced. This sucks, but it's worse if the people who got divorced made an ironclad promise that they would stay together til death did them part. Indeed, that's a foolish promise to make if you're looking at base rates and don't have an extremely good justification for thinking why you're likely to beat the odds by a lot.

And that's okay! Just make your promise carry an escape clause. I attended a wedding of some friends recently where they promised not to get divorced unless they both climbed a particular mountain first. They're planning on staying together, but they recognize that they can't know for sure they will want this, so they've left themselves a way out. This is a more honest thing to do than promising to never leave. It makes their promise mean more.

I've seen this kind of vow at a number of weddings and I'd love to see it more. As a witness to people's vows, I want to know what I'm there to witness, and how I can help them keep their commitment.

Sometimes, we ask other people to make commitments. Maybe it's asking an employee to sign an NDA. Maybe it's asking a friend to keep a secret. I think in our current environment of commitment-seriousness-ambiguity, asking people to commit to things is a serious business, and I think people are often too cavalier about it. If you're asking someone to commit to something, it's partially your responsibility to help them understand what they're committing to. You should not ask people to commit to things if you don't have a good model of what they're committing to or how hard it will be for them to keep their commitment. Asking someone to commit to something they aren't likely to be able to carry through on erodes the commons, because it incentivizes people making commitments they can't keep.

Of course the one committing still holds most of the responsibility to keep their commitment, but circumstances and incentives matter here too. When there is a power difference between the asker and committer, we should expect the asker to have a greater responsibility than they otherwise would to make sure the committer understands what they're agreeing to.

I'd like to see people come up with more best practices for commitments. A few might be:

- Don't commit to or ask people to commit to things you think you or they are not likely to be able to complete
- When making or asking for commitments, include an escape clause if following through on the commitment might be really costly - the escape clause can

- include costs in order to preserve some incentive to keep the commitment
- Time-bound most commitments by default & don't make unlimited or unbounded commitments or ask others to unless there's a really good reason
- Get advice from several people you trust before making big commitments and make sure people you're asking to make big commitments have done the same

## **Giving What We Can pledge**

Within the EA community, the [Giving What We Can pledge](#) is the biggest community-specific commitment that people make. Unfortunately, I think the way it's currently worded does not clarify the kind of commitment it implies, and thus GWWC unintentionally erodes the ability of people in our community to make public pledges effectively.

Here is the text of the pledge in full:

"I recognise that I can use part of my income to do a significant amount of good. Since I can live well enough on a smaller income, I pledge that from \_\_ until \_\_ I shall give \_\_ to whichever organisations can most effectively use it to improve the lives of others, now and in the years to come. I make this pledge freely, openly, and sincerely."

I think it's great that the pledge now asks you to specify a starting and end time and particular percentage by default. (Previously, it read "until I retire"). I think it's quite bad that the main text of the pledge doesn't include any mention of an exit clause. The website does mention some things around this in their FAQ, but unfortunately this too doesn't provide much clarity:

FAQ: Is a pledge legally binding? What if my circumstances change?

Our pledges are in no way legally binding. They are commitments made voluntarily and enforced solely by your own conscience. In some circumstances, it may be best to resign from your pledge.

In some circumstances?? Some circumstances like "my partner has a life threatening disease" or some circumstances like "I make 20% less money now" or "I switched to direct work and think donating doesn't [make sense](#) for me anymore"? The differences between these really matter! As someone witnessing people make this public commitment, how can I help hold people accountable without knowing what they're pledging to, and under what circumstances they should break it?

The Expanded FAQ adds more detail but not more clarity:

Expanded FAQ:

How does it work? Is it legally binding?

The Pledge is not a contract and is not legally binding. It is, however, a public declaration of lasting commitment to the cause. It is a promise, or oath, to be made seriously and with every expectation of keeping it. All those who want to become a member of Giving What We Can must make the Pledge, and we ask them to report their income and donations each year.

Taking the Pledge is something to be considered seriously, but we understand if a member can no longer keep it. If it is best for someone to resign from their Pledge they can depledge and are welcome to rejoin later.

After reading all this I still have very little idea what kind of promise GWWC is. I want people to take public commitments seriously, but I don't believe they can without thinking clearly about what exactly they're promising. I think GWWC being vague about this is pretty irresponsible. I want people to build within themselves the machinery to be able to make strict pledges that mean things, and I think agreeing to a pledge like this erodes that machinery. By my own standard, if I agreed to a pledge like this, I'd need to carefully specify the conditions under which I'd allow myself to exit this pledge or not, since it's not nearly clear enough to me in its current wording.

Sometimes people make mistakes in their promises. That sucks, and people break trust when they do that, but it's also okay. People grow and learn, and the thing I care about is people working towards more honesty and integrity. It's a process to learn how to be really honest with yourself and others. I've broken promises before, and I feel sad that I did. I can't change that, but I can change what promises I make going forward. I want to be a person of unusual honesty and integrity, and so I want to think about what commitments mean to me and how I can structure my environment to help me make good ones and keep them.

# Safety Implications of LeCun's path to machine intelligence

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

Yann LeCun recently posted [A Path Towards Autonomous Machine Intelligence](#), a high-level description of the architecture he considers most promising to advance AI capabilities.

This post summarizes the architecture and describes some implications for AI safety work if we accept the hypothesis that the first [transformative AI](#) will have this architecture.

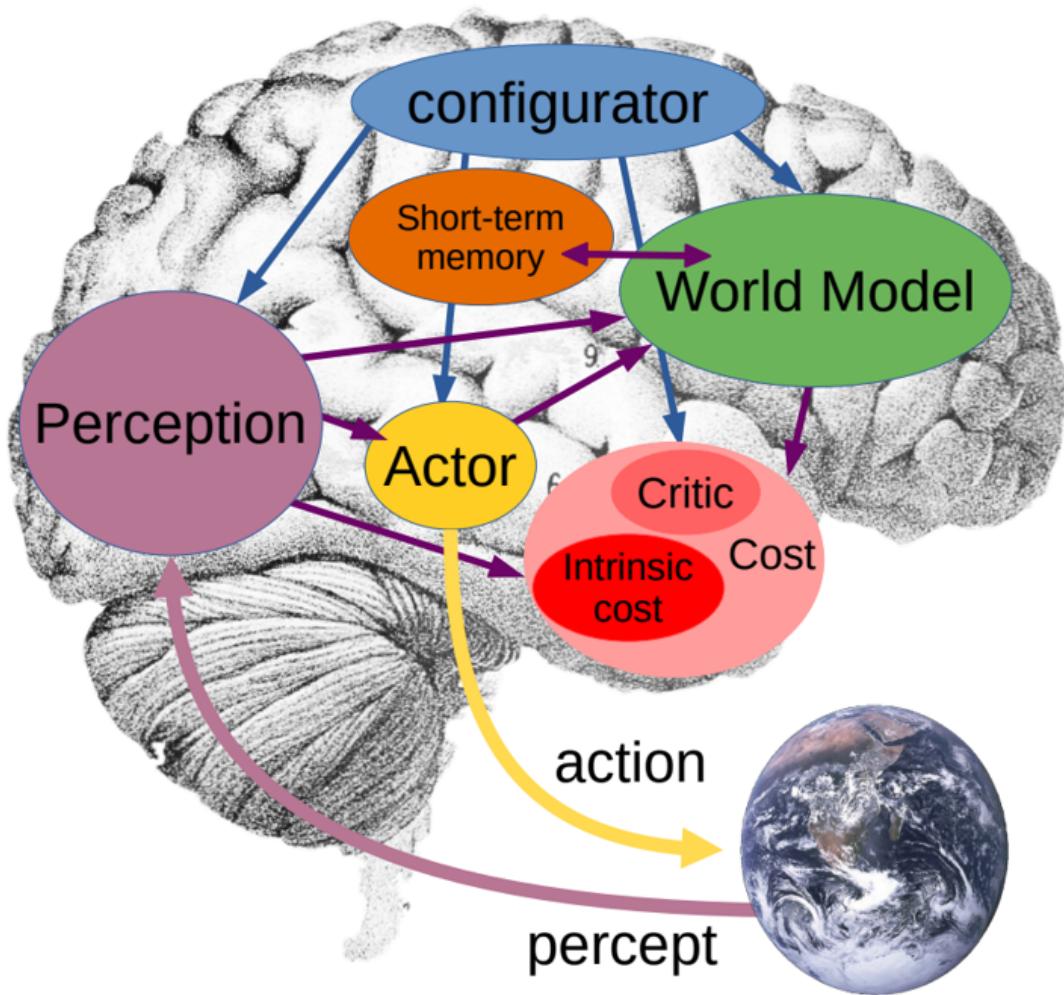
Why is this a hypothesis worth considering? My reasons for elevating this hypothesis, in order of importance, are

1. LeCun has a track record of being ahead of mainstream academic research, from working on CNNs in the 90s to advocating for self-supervised learning back in 2014-2016 when supervised learning was ascendant.
2. LeCun runs Meta AI (formerly FAIR) which has enormous resources and influence to advance his research agenda, making it more likely that his proposed architecture will be built at scale. In general I think this is an underrated factor; AI research exhibits a great deal of path dependence, and most plausible paths to AI are not taken primarily because nobody is willing to take a big risk on them.
3. The architecture is dramatically different from the architectures commonly assumed (implicitly) in much AI alignment work, such as model-free deep RL and "GPT-3 but scaled up 10000x". This makes it a good robustness check for plans that are overly architecture-specific.

## Architecture Overview

### The Overall Agent

At a high level, the proposed architecture is a set of specialized cognitive modules. With the exception of the Actor and the Intrinsic Cost (see below) they are all deep neural networks trained with gradient descent.



The high level architecture of LeCun's proposed agent. Arrows indicate dependence; gradients flow backward through the thin arrows.

What is this agent doing, exactly? It's meant to be a general architecture for any autonomous AI agent, but LeCun repeatedly emphasizes video inputs and uses self-driving cars as a recurrent example, so the central use case is embodied agents taking actions in the physical world. Other talks I've seen by LeCun suggest he thinks understanding video is essential for intelligence, both by analogy to humans and by a heuristic argument about the sheer amount of data it contains.

## The World Model

More than half the body of the paper is about designing and training the *world model*, the predictive model of the environment that the AI uses to plan its actions. LeCun explicitly says that "*designing architectures and training paradigms for the world model constitute the main obstacles towards real progress in AI over the next decades.*"

Why are world models so important? Because the main limitation of current AI systems, according to LeCun, is their sample inefficiency - they need millions of expensive, dangerous real-world interactions to learn tasks that humans can learn with only a few examples. The

main way to progress capabilities is to reduce the number of interactions a system needs before it learns how to act, and the most promising way is to learn predictive world models on observational data. (The GPT-3 paper [Large Language Models are Few Shot Learners](#) is a great example of this - a good enough predictive model of language enables much more sample-efficient task acquisition than supervised learning).

What will these world models look like? According to LeCun, they will be

1. **Predictive but not generative:** They will predict high-level features of the future environment but not be able to re-generate the whole environment. This is especially obvious for high-dimensional data like video, where predicting the detailed evolution of every pixel is vastly overkill if you're doing planning. But it could also apply to language agents like chatbots, for whom it may be more important to predict the overall sentiment of a user's reply than the exact sequence of tokens.
2. **Uncertainty-aware:** able to capture multimodal distributions over future evolutions of the world state (e.g. whether the car will turn left or right at the upcoming intersection), which LeCun expects to be modeled with latent variables. The ability to model complex uncertainty is the key property LeCun thinks is missing from modern large generative models, and leads him to conclude that "scaling is not enough".
3. **Hierarchical:** represent the world at multiple levels of abstraction, with more high-level abstract features evolving more slowly. This makes it computationally feasible to use the same model for the combination of long-term planning and rapid local decision making that characterizes intelligent behavior.
4. **Unitary:** AIs will trend towards having one joint world model across all modalities (text, images, video), timescales, and tasks, enabling hardware re-use and knowledge sharing (LeCun speculates that human "common sense" and ability to reason by analogy emerges from humans having a unitary world model). This suggests the trend towards "one giant model" we've seen in NLP will continue and broaden to include the rest of AI.

## The Actor

The actor generates action sequences which minimize the cost (see below) according to the world-model's predictions. It generates these action sequences via some search method; depending on the task, this could be

- classic heuristic search methods like Monte-Carlo tree search or beam search.
- gradient-based optimization of the action sequence's embedding in some continuous space.

Optionally, one can use imitation learning to distill the resulting action sequence into a policy network. This policy network can serve as a fast generator of actions, analogous to Kahneman's System 1 thinking in humans, or to inform the search procedure like in the {AlphaGo, AlphaZero, MuZero} family of models.

Unlike the world model, the actor is not unitary - it's likely that different tasks will use different search methods and different policy networks.

## The Cost

So what exactly is this agent optimizing? There is a hard-wired, non-trainable mapping from world states to a scalar "intrinsic cost". The actor generates plans that minimize the sum of costs over time, which makes costs mathematically equivalent to rewards in reinforcement learning.

I think the reason LeCun insists on using his unusual terminology is that he wants to emphasize that in this scheme, *normative information does not come from an external source* (like a reward provided by a human supervisor) but is an *intrinsic drive* hard-coded into the agent (like pain, hunger, or curiosity in humans).

## The Configurator

The configurator is a component that modulates the behavior of all other components, based on inputs from all other components; it's not specified in any detail and mostly feels like a pointer to "all the component interactions LeCun doesn't want to think about".

It's especially critical from an alignment perspective because it modulates the cost, and thus is the only way that humans can intervene to change the motivations of the agent. LeCun speculates that we might want this modulation to be relatively simple, perhaps only specifying the relative weights of a linear combination of several basic hardcoded drives because this makes the agent easier to control and predict. He also mentions we will want to include "cost terms that implement safety guardrails", though what these terms are and how the configurator learns to modulate them is left unspecified.

## Implications for AI Safety

Let's assume that the first transformative AI systems are built roughly along the lines LeCun describes. What would this imply for AI safety work?

1. **Interpretability becomes much easier**, because the agent is doing explicit planning with a structured world-model that is purely predictive. Provided we can understand the hidden states in the world model (which seems doable with a [Circuits](#)-style approach), we can directly see what the agent is planning to do and implement safety strategies like "check that the agent's plan doesn't contain any catastrophic world states before executing an action". Of course, a sufficiently powerful agent could learn to model our safety strategies and avoid them, but the relatively transparent structure of LeCun's architecture gives the defender a big advantage.
2. **Most safety-relevant properties will be emergent** from interaction rather than predictable in advance, similar to the considerations for [Multi-agent safety](#). Most of the "intelligence" in the system (the world model) is aimed at increasing predictive accuracy, and the agent is motivated by relatively simple hard-coded drives; whether its intelligent behaviors are safe or dangerous will not be predictable in advance. This makes it less tractable to intervene on the model architecture and training process (including most theoretical alignment work), and more important to have excellent post-training safety checks including simulation testing, adversarial robustness and red-teaming.
3. **Coordination / governance is relatively more important**. Whether an AI deployment leads to catastrophic outcomes will mostly be a function not of the agent's properties, but of the safety affordances implemented by the people deploying it (How much power are they giving the agent? How long are they letting it plan? How well are they checking the plans? ). These safety affordances are likely to be increasingly expensive as the model's capabilities grow, likely following the computer systems rule of thumb that every [nine of reliability](#) costs you 10x, and possibly scale even worse than that. Ensuring this high [alignment tax](#) is paid by all actors deploying powerful AI systems in the world requires a very high level of coordination.

## Conclusion and Unresolved Questions

Broadly, it seems that in a world where LeCun's architecture becomes dominant, useful AI safety work looks more analogous to the kind of work that goes on now to make self-driving cars safe. It's not difficult to understand the individual components of a self-driving car or to debug them in isolation, but emergent interactions between the components and a diverse range of environments require massive and ongoing investments in testing and redundancy.

Two important questions that remain are

1. How likely is it that this becomes the dominant / most economically important AI architecture? Some trends point towards it (success of self-supervised learning and unitary predictive models; model-based architectures dominant in economically important applications like self-driving cars and recommender systems), others point away (relative stagnation in embodied / video-based agents vs language models; success of model-free RL in complex video game environments like [StarCraft](#) and [Dota 2](#)).
2. Just how clean will the lines will be between model, actor, cost, and configurator? Depending on how the architecture is trained (and especially if it is trained end-to-end), it seems possible for the world-model or the configurator to start learning implicit policies, in a way that undermines interpretability and the safety affordances it creates.

# How to Diversify Conceptual Alignment: the Model Behind Refine

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

*This post is part of the work done at [Conjecture](#).*

TL;dr: We need far more conceptual AI alignment research approaches than we have now if we want to increase our chances to solve the alignment problem. However, the conceptual alignment field remains hard to access, and what feedback and mentorship there is focuses around few existing research directions rather than stimulating new ideas. This model lead to the creation of [Refine](#), a research incubator for potential conceptual alignment researchers funded by the [LTFF](#) and hosted by [Conjecture](#). Its goal is to help conceptual alignment research grow in both number and variety, through some minimal teaching and a lot of iteration and feedback on incubatees' ideas. The first cohort has been selected, and will run from August to October 2022. In the bigger picture, Refine is an experiment within [Conjecture](#) to find ways of increasing the number of conceptual researchers and improve the rate at which the field is making [productive mistakes](#).

## The Problem: Not Enough Varied Conceptual Research

I believe that in order to solve the alignment problem, we need significantly more people attacking it from a lot different angles.

Why? First because none of the current approaches appears to yield a full solution. I expect many of them to be [productive mistakes](#) we can and should build on, but they don't appear sufficient, especially with shorter timelines.

In addition, the history of science teaches us that for many important discoveries, especially in difficult epistemic situations, the answers don't come from one lone genius seeing through the irrelevant details, but instead from bits of evidence revealed by many different takes and operationalizations<sup>[1]</sup> (possibly unified and compressed together at the end). And we should expect alignment to be hard based on [epistemological vigilance](#).

So if we accept that we need more people tackling alignment in more varied ways, why are we falling short of that ideal? Note that I will focus here on conceptual researchers, as they are the source of most variations on the problem, and because they are so hard to come by.

I see three broad issues with getting more conceptual alignment researchers working on wildly different approaches:

1. **(Built-in Ontological Commitments)** Almost all current attempts to create more conceptual alignment researchers ([SERI MATS](#), independent mentoring...) rely significantly on mentorship by current conceptual researchers. Although this

obviously comes with many benefits, it also leads to many ontological commitments being internalized when one is learning the field. As such, it's hard to go explore a vastly different approach because the way you see the problem has been moulded by this early mentorship.

2. **(Misguided Requirements)** I see many incorrect assumptions about what it takes to be a good conceptual researcher floating around, both from field-builders and from potential candidates. Here's a non-exhaustive list of the most frustrating ones
  - You need to know all previous literature on alignment (the field has more breadth than depth, and so getting a few key ideas is more important than knowing everything)
  - You need to master maths and philosophy (a lot of good conceptual work only uses basic maths and philosophy)
  - You need to have an ML background (you can pick up the relevant part and just work on approaches different to pure prosaic alignment)
3. **(No Feedback)** If you want to start on your own, you will have trouble getting any feedback at all. The AF doesn't provide much feedback even for established researchers, and it has almost nothing in store for newcomers. Really, the main source of feedback in the field is asking other researchers, but when you start you usually don't know anyone. And without feedback, it's hard to stay motivated and ensure your work is relevant to the core problem.

[Refine](#), the incubator for conceptual researchers and research bets that I'm running at Conjecture, aims at addressing these issues.

## Description of Refine

### Research Incubator

Refine is a research incubator: that is, a program for helping potential conceptual researchers improve and create relevant ideas and research. It's inspired by startup incubators like [Y combinator](#), but with a focus on research. As such, the point is not to make participants work on already trusted research directions, but to give them all the help they need to create exciting and relevant new research questions and ideas that are highly relevant to alignment.

In broad strokes, Refine starts with two weeks focused around studying and discussing core ideas in the History and Philosophy of Science and in the Epistemology of Alignment, followed by 10 weeks of intense idea-generation-feedback-writing loops (for a total of 3 months).

At the end, the research produced will be evaluated by established conceptual researchers, and we'll help the incubatees get funding or get hired (at Conjecture or other places).

In more details, the first cohort of Refine will follow this process:

- **Selection:** by order of priority (more details in [the call for participants](#))
  - Relentlessly resourceful
  - Access to weird and different ideas and frames
  - Understanding of the alignment problem (by default applicants have a minimum understanding to even care to apply)

- **Initial power-up (2 weeks):** the program begins with two weeks of reading, presentations, discussions and debates about core ideas in the epistemology of alignment. The goal is to give people tools and keys for thinking about the problem and bias them towards the core questions while still leaving them a lot of margin for innovation.
  - Before start of cohort: reading group of posts presenting different takes on alignment
    1. [What Multipolar Failures Look Like](#) by Andrew Critch
    2. [Why Agent Foundations? An Overly Abstract Explanation](#) by John Wentworth
    3. [How do we become confident in the safety of a machine learning system?](#) by Evan Hubinger
    4. [My research methodology](#) by Paul Christiano
    5. [A central AI alignment problem: capabilities generalization, and the sharp left turn](#) by Nate Soares
  - Week 1: History and Philosophy of Science and Models of Progress
    1. [Productive Mistakes](#)
    2. [Epistemological Vigilance](#)
    3. [Mosaic and Palimpsests](#)
    4. Pluralism (Posts about it in the works)
  - Week 2: Epistemology of Alignment
    1. High-level Map of Conceptual Alignment Research
    2. Unbounded Atomic Optimization (Posts about it in the works)
- **Intense iteration (10 weeks):**
  - Incubatee generates and explores idea
  - We discuss the ideas, along a bunch of lines
    1. Assumptions made
    2. Interesting parts of the productive mistake
    3. Failings/limits
  - Based on the discussion and feedback, the idea is either closed (because no clear way to improve upon it, or relevant but not priority now, or not relevant, or no clear ways of extending it) or open
  - If closed idea, then produce an artifact about it and go back to step 1) with new direction
  - If open idea, then go back to step 1) but about the directions that came from questioning the idea
- **Evaluation**
  - Final write-up
  - Help them write grant applications and get funding/jobs
  - Gather feedback from established conceptual alignment researchers

## Generalist Mentors

Rather than having current researchers act as PhD advisors on their own topics, Refine aims at leveraging more generalist mentors (currently me) who can see value and issues in almost all approaches, while understanding the problem deeply enough to give relevant feedback. The hope is that this kind of support will minimize ontological commitments while still biasing the work towards the hard problem.

In addition, generalist mentors avoid the overuse of the scarce resource of conceptual researchers, and might be a great fit for thinkers focused on the sort of epistemological work I'm doing at Conjecture.

# Selection and Respect

(The Black Swan, Nassim Nicholas Taleb, 2007)

Many people labor in life under the impression that they are doing something right, yet they may not show solid results for a long time. They need a capacity for continuously adjourned gratification to survive a steady diet of peer cruelty without becoming demoralized. They look like idiots to their cousins, they look like idiots to their peers, they need courage to continue. No confirmation comes to them, no validation, no fawning students, no Nobel, no Shnobel. "How was your year?" brings them a small but containable spasm of pain deep inside, since almost all of their years will seem wasted to someone looking at their life from the outside. Then bang, the lumpy event comes that brings the grand vindication. Or it may never come.

Believe me, it is tough to deal with the social consequences of the appearance of continuous failure. We are social animals; hell is other people.

[...]

We favor the sensational and the extremely visible. This affects the way we judge heroes. There is little room in our consciousness for heroes who do not deliver visible results—or those heroes who focus on process rather than results.

[...]

But this does not mean that the person insulated from materialistic pursuits becomes impervious to other pains, those issuing from disrespect. Often these Black Swan hunters feel shame, or are made to feel shame, at not contributing. "You betrayed those who had high hopes for you," they are told, increasing their feeling of guilt. The problem of lumpy payoffs is not so much in the lack of income they entail, but the pecking order, the loss of dignity, the subtle humiliations near the watercooler.

It is my great hope someday to see science and decision makers rediscover what the ancients have always known, namely that our highest currency is respect

Building and running a program like Refine leads to a conundrum. On the one hand, there are obviously tests and evaluations involved: at the beginning to select people, during the program, and at the end to decide if the program was successful. On the other hand, the anxiety of being always judged and evaluated is corrosive, as Taleb expresses so clearly.

I don't have a perfect solution. [The dark world](#) is that both need to be taken into account for the program to succeed.

My current choice is to use these two different frames in distinct contexts. During the selection process, and when making the post-mortem, I should take an evaluative frame, while remembering that historical progress is incredibly more subtle than the parody we often make of it. And during the actual running of the program, I shouldn't be in an evaluative mindset, but only focus on how to help the participants do the best they can.

# Difference with Other Programs

With more and more programs around alignment in the last few years, it makes sense to ask if the problem we're tackling with Refine has not been addressed already. I'm definitely excited about all these programs; yet they all target different enough problems that I don't think they are addressing the lack of varied conceptual research completely.

- [SERI MATS](#) attacks the problem of creating more researchers for already established agendas — what I call the accelerated PhD model. As such, its participants are heavily directed and biased towards the current ontological commitments, rather than pushed to try completely new things.
- [AI Safety Camp](#) has been shifting around recently, but the earlier editions lacked the detailed feedback of generalist mentors, while the most recent edition (which I was involved with) was a form of the accelerated PhD model and thus had the same issues as MATS for generating new takes.
- [PIBSS](#) aims at diversification, not directly creating new conceptual researchers or even new approaches necessarily. Still, the PIBSS fellows could definitely constitute a strong group to select future cohorts from.
- [AGI Safety Fundamentals](#) focuses on education rather than production of research, and is strongly colored by the ontological commitments of Richard Ngo.

## Some Concrete Details

The first cohort of Refine, funded by the [Long-Term Future Fund](#), will happen from August to October 2022. The ops are managed by [Conjecture](#), and it will happen in France initially (for administrative reasons), then in London at Conjecture's offices. We pay incubatees a stipend, and also cover all their travel and housing.

The first cohort is composed of Alexander Gietelink Oldenziel, Chin Ze Shen, Tamsin Leake, Linda Linsefors, and Paul Bricman. In terms of statistics, it's interesting to notice that none of the participants are British or American: 4 out of 5 are from continental Europe, and one is from Southeast Asia. In terms of knowledge of alignment, 2 have a deep interaction with the field, 2 have thought independently about it a lot, and one is relatively new to it.

For the final evaluation, Steve Byrnes, Vanessa Kosoy, Evan Hubinger, Ramana Kumar, and John Wentworth all committed to look and evaluate the output of at least a few participants, and give judgment on whether they are excited by the research produced.

## The Long View: Refine and Conjecture

The idea for Refine mostly came from my own frustrations with the small growth of conceptual alignment research, and from a project of an independent lab with Jessica Cooper.

Yet Conjecture management has been excited about it since even before I joined officially, and Refine fits well within the core mission of Conjecture: to improve and

scale alignment research by finding many angles of attack on the problem and then supporting researchers to do the best possible work.

In this perspective, Refine is an experiment to find ways of diversifying alignment research and making more productive mistakes. It's a tentative way of converting resources into more varied and unexplored alignment research directions, and generally to help create more and better conceptual alignment researchers.

If Refine is successful at producing exciting new research and researchers, then finding ways to replicate it, improve it, and scale it (maybe in a decentralized way) will become one of Conjecture's priorities. If it isn't successful, then we will learn the most we can from the failure and iterate on other options to create great and varied conceptual alignment research.

I also see a strong synergy between the needs of Refine-like programs and the epistemology team I'm leading at Conjecture. More specifically, researchers focused on the History and Philosophy of Science and the Epistemology of Alignment seem like great fits for generalist mentors, because they are steeped in the details of progress and alignment enough to provide useful and subtle feedback while minimizing ontological commitments.

## 1. ^

I will dig into this in future posts, but if you want pointers now, you can see my post on [productive mistakes](#), Chapter 2 (on electrolysis) and Chapter 3 (on chemical atomism) of [Is Water H<sub>2</sub>O?](#) by Hasok Chang, and [Rock, Bone, and Ruin](#) by Adrian Currie.

# All AGI safety questions welcome (especially basic ones) [July 2022]

**tl;dr: Ask questions about AGI Safety as comments on this post, including ones you might otherwise worry seem dumb!**

Asking beginner-level questions can be intimidating, but everyone starts out not knowing anything. If we want more people in the world who understand AGI safety, we need a place where it's accepted and encouraged to ask about the basics.

As requested in the previous thread<sup>[1]</sup>, we'll be putting up monthly FAQ posts as a safe space for people to ask all the possibly-dumb questions that may have been bothering them about the whole AGI Safety discussion, but which until now they didn't feel able to ask.

It's okay to ask uninformed questions, and not worry about having done a careful search before asking.



## Stampy's Interactive AGI Safety FAQ

Additionally, this will serve as a soft-launch of the project [Rob Miles' volunteer team](#)<sup>[2]</sup> has been working on: **Stampy** - which will be (once we've got considerably more content) a single point of access into AGI Safety, in the form of a comprehensive interactive FAQ with lots of links to the ecosystem. We'll be using questions and answers from this thread for Stampy (under [these copyright rules](#)), so please only post if you're okay with that! You can help by [adding](#) other people's questions and answers to Stampy or [getting involved in other ways](#)!

We're not at the "send this to all your friends" stage yet, we're just ready to onboard a bunch of editors who will help us get to that stage :)



**Stampy** - Here to help everyone learn about stamp maximization AGI Safety!

We welcome [feedback](#)<sup>[3]</sup> and questions on the UI/UX, policies, etc. around Stampy, as well as pull requests to [his codebase](#).<sup>[4]</sup> You are encouraged to add other people's answers from this thread to Stampy if you think they're good, and collaboratively improve the content that's already on [our wiki](#).

We've got a lot more to write before he's ready for prime time, but we think Stampy can become an excellent resource for everyone from skeptical newcomers, through people who want to learn more, right up to people who are convinced and want to know how they can best help with their skillsets.

#### **Guidelines for Questioners:**

- No previous knowledge of AGI safety is required. If you want to watch a few of the [Rob Miles videos](#), read either the [WaitButWhy](#) posts, or the [The Most Important Century](#) summary from OpenPhil's co-CEO first that's great, but it's not a prerequisite to ask a question.

- Similarly, you do not need to try to find the answer yourself before asking a question (but if you want to test [Stampy's in-browser tensorflow semantic search](#) that might get you an answer quicker!).
- Also feel free to ask questions that you're pretty sure you know the answer to, but where you'd like to hear how others would answer the question.
- One question per comment if possible (though if you have a set of closely related questions that you want to ask all together that's ok).
- If you have your own response to your own question, put that response as a reply to your original question rather than including it in the question itself.
- Remember, if something is confusing to you, then it's probably confusing to other people as well. If you ask a question and someone gives a good response, then you are likely doing lots of other people a favor!

### **Guidelines for Answerers:**

- Linking to the relevant [canonical answer](#) on Stampy is a great way to help people with minimal effort! Improving that answer means that everyone going forward will have a better experience!
- This is a safe space for people to ask stupid questions, so be kind!
- If this post works as intended then it will produce many answers for Stampy's FAQ. It may be worth keeping this in mind as you write your answer. For example, in some cases it might be worth giving a slightly longer / more expansive / more detailed explanation rather than just giving a short response to the specific question asked, in order to address other similar-but-not-precisely-the-same questions that other people might have.

**Finally:** Please think very carefully before downvoting any questions, remember this is the place to ask stupid questions!

1. ^

I'm re-using content from [Aryeh Englander's thread](#) with permission.

2. ^

If you'd like to join, head over to [Rob's Discord](#) and introduce yourself!

3. ^

Either via the [feedback form](#) or in the [feedback thread](#) on this post.

4. ^

Stampy is a he, we asked him.

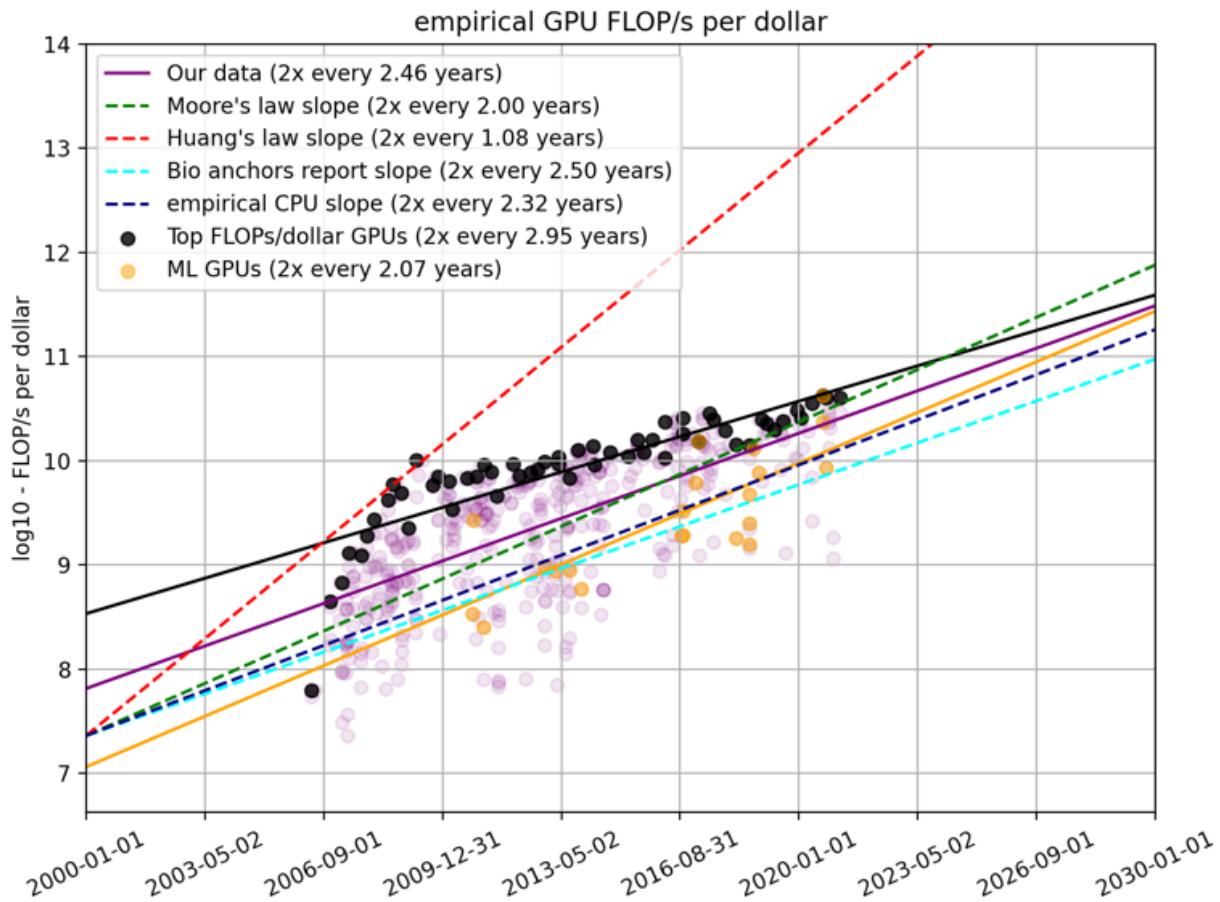
# Trends in GPU price-performance

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a linkpost for <https://epochai.org/blog/trends-in-gpu-price-performance>

## Executive Summary

Using a dataset of 470 models of graphics processing units (GPUs) released between 2006 and 2021, we find that the amount of floating-point operations/second per \$ (hereafter FLOP/s per \$) **doubles every ~2.5 years**. For top GPUs, we find a slower rate of improvement (FLOP/s per \$ doubles every 2.95 years), while for models of GPU typically used in ML research, we find a faster rate of improvement (FLOP/s per \$ doubles every 2.07 years). GPU price-performance improvements have generally been slightly slower than the 2-year doubling time associated with Moore's law, much slower than what is implied by Huang's law, yet considerably faster than was generally found in prior work on trends in GPU price-performance. Our work aims to provide a more precise characterization of GPU price-performance trends based on more or higher-quality data, that is more robust to justifiable changes in the analysis than previous investigations.



**Figure 1.** Plots of FLOP/s and FLOP/s per dollar for our dataset and relevant trends from the existing literature

Trend	2x time	10x time	Metric
Our dataset (n=470)	2.46 years [2.24, 2.72]	8.17 years [7.45, 9.04]	FLOP/s per dollar
ML GPUs (n=26)	2.07 years [1.54, 3.13]	6.86 years [5.12, 10.39]	FLOP/s per dollar
Top GPUs (n=57)	2.95 years [2.54, 3.52]	9.81 years [8.45, 11.71]	FLOP/s per dollar
Our data FP16 (n=91)	2.30 years [1.69, 3.62]	7.64 years [5.60, 12.03]	FLOP/s per dollar
Moore's law	2 years	6.64 years	FLOP/s
Huang's law	1.08 years	3.58 years	FLOP/s
CPU historical ( <a href="#">AI Impacts, 2019</a> )	2.32 years	7.7 years	FLOP/s per dollar
<a href="#">Bergal, 2019</a>	4.4 years	14.7 years	FLOPs/dollar

**Table 1.** Summary of our findings on GPU price-performance trends and relevant trends in the existing literature with the 95% confidence intervals in square brackets.

In future work, we intend to build on this work to produce projections of GPU price-performance, and investigate how our findings inform us about the growth in dollar-spending on computing hardware in Machine Learning.

*We would like to thank Alyssa Vance, Ashwin Acharya, Jessica Taylor and the Epoch team for helpful feedback and comments.*

# MATS Models

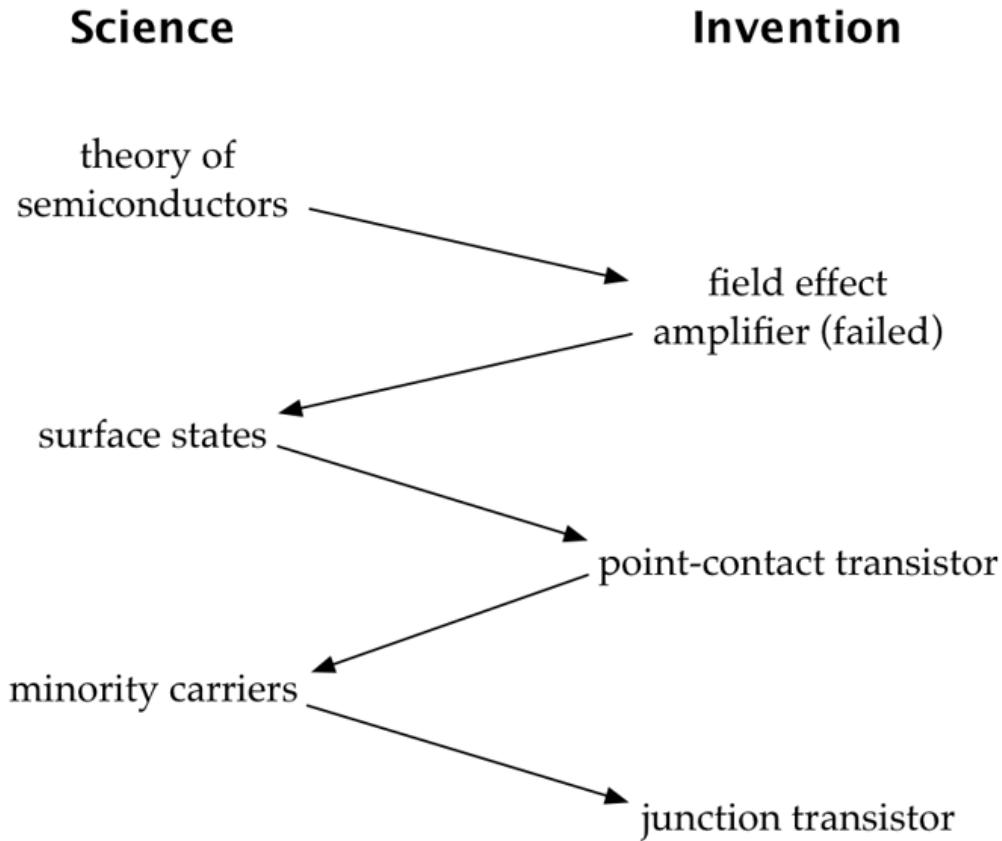
I've been using the summer 2022 [SERI MATS program](#) as an opportunity to test out my current best guesses at how to produce strong researchers. This post is an overview of the methods I've been testing, and the models behind them.

## The Team Model

My MATS participants are in three-person teams with specialized roles for each person: theorist, experimentalist, and distillitator ("distiller" + "facilitator"). This team setup isn't just a random guess at what will work; the parts are nailed down by multiple different models. Below, I'll walk through the main models.

## Jason Crawford's Model & Bits-of-Search

In [Shuttling Between Science and Invention](#), Jason Crawford talks about the invention of the transistor. It involved multiple iterations of noticing some weird phenomenon with semiconductors, coming up with a theory to explain the weirdness, prototyping a device based on the theory, seeing the device not work quite like the theory predicted (and therefore noticing some weird phenomenon with semiconductors), and then going back to the theorizing. The key interplay is the back-and-forth between theory and invention/experiment.



The back-and-forth between theory and experiment ties closely to one of the central metaresearch questions I think about: where do we get our bits from? The space of theories is exponentially huge, and an exponentially small fraction are true/useful. In order to find a true/useful theory, we need to get a lot of bits of information from *somewhere* in order to narrow down the possibility space. So where do those bits come from?

Well, one readily available source of tons and tons of bits is the physical world. By running experiments, we can get lots of bits. (Though this model suggests that most value comes from a somewhat different *kind* of experiment than we usually think about - more on that later.)

Conversely, blind empiricism runs into the [high-dimensional worlds problem](#): brute-force modeling our high-dimensional world would require exponentially many experiments. In order to efficiently leverage data from the real world, we need to know what questions to ask and what to look at in order to eliminate giant swaths of the possibility space simultaneously. Thus the role of theory.

In order to efficiently build correct and useful models, we want empiricism and theory coupled together. And since it's a lot easier to find people specialized in one or the other than people good at both, it makes sense to partner a theorist and an experimentalist together.

## Nate Soares' Model

The hard step of theory work typically involves developing some vague intuitive concept/story, then operationalizing it in such a way that an intuitive argument turns into a rigorous mathematical derivation/proof. Claim which I got from Nate Soares: this process involves at most two people, and the second person is in a [facilitator role](#).

(Note that the description here is mine; I haven't asked Nate whether he endorses it as a description of what he had in mind.)

What does that facilitator role involve? A bare-minimum version is the programmer's "Rubber Duck": the rubber duck sits there and listens while the programmer explains the problem, and hopefully the process of explanation causes the programmer to understand their problem better. A skilled facilitator can add a lot of value beyond that bare minimum: they do things like ask for examples, try to summarize their current understanding back to the explainer, try to restate the core concept/argument, etc.

What the facilitator does *not* do is actively steer the conversation, suggest strategies or solutions or failure modes, or drop their own ideas in the mix. Vague intuitive concepts/stories are *brittle*, easy to accidentally smash and replace with some other story; you don't want to accidentally overwrite someone's idea with a different one. The goal is to take the intuitions in the explainer's head, and *accurately turn those intuitions into something legible*.

A key part of a skilled facilitator's mindset is that they're trying to force-multiply somebody else's thoughts, not show off their own ideas. Their role is to help the theorist communicate the idea, likely making it more legible in the process.

## Eli Tyre's Model

Eli Tyre was the source of the name "distillator" for this role. What he originally had in mind was people who serve as both [distiller](#) and [double-crux](#) facilitator. On the surface, these two roles might sound unrelated, but they use the same underlying skills: both are (I claim) mainly about maintaining a mental picture of what another person is talking about, and trying to keep that picture in sync with the other person's mental picture. That's what drives facilitator techniques like asking for examples, summarizing back the facilitator's current understanding, restating the core concept/argument, etc; all of these are techniques for making the listener's mental picture match the speaker's mental picture. Those techniques aren't just useful for making the facilitator's mental model match the speaker's mental model; the same information can help an audience build a matching mental picture, which is the core problem of distillation.

With that in mind, the "facilitator" part of the role naturally generalizes between just double-crux facilitation. It's the same core skillset needed by a facilitator under Nate's model, in order to help extract and formalize intuitive concepts/arguments for theory-building. [It's also a core skillset of good communication in general](#); a good distillator can naturally facilitate many kinds of conversations. They're a natural "social glue" on a team.

And, of course, they can produce clear and interesting write-ups of the team's work for the rest of us.

## John's Model

Now we put all that together, with a few other minor pieces.

As a general rule, the number of people who can actively work together on the same thing in the same place at the same time is three. Once we get past three, either the work needs to be broken down into modular chunks, or someone is going to be passively watching at any given time. Four person teams can work, but usually not better than three, and by the time we get to five it's usually worse than three because subgroups naturally start to form.

Within the three-person team, we want people with orthogonal skillsets/predilections, because it is just really hard to find people who have all the skills. Even just finding people with any two of strong theory skills, experiment skills, and writing skills is hard, forget all three.

So, the model I recommend is one theorist, one experimentalist, and one distillitator.

That said, I do *not* think that people should be very tightly bound by their role during the team's day-to-day work. There is a ton of value in everyone doing everything at least some of the time, so that each person deeply understands what the others are doing and how to work with them effectively. The rule I recommend is: "**If something falls under your job/role, it is your responsibility to make sure it is done when nobody else is doing it; that does not necessarily mean doing it yourself**". So, e.g., it is the facilitator's job to facilitate in a discussion with nobody else facilitating. It is the theorist's job to crank the math when nobody else is doing that. It is the experimentalist's job to write the code for an ML experiment if nobody else is doing that. But it is strongly encouraged for each person to do things which aren't "their job". Also, of course, anyone could delegate, though it is then the delegator's responsibility to make sure the delegatee is willing and able to do the thing.

For work on AI alignment and agent foundations specifically, I think experimentalists are easiest to recruit; a resume with a bunch of ML experience is usually a pretty decent indicator, and standard education/career pathways already provide most of the relevant skills. Distillitator skills are less explicit on a resume, but still relatively easy to test for - e.g. good writing (with examples!) is a pretty good indicator, and I expect the workshops below help a lot for practicing the distillitator skillset. Theory skills are the hardest to recognize, especially when the hopeful-theorist is not themselves very good at explaining things (which is usually the case).

## Exercises & Workshops for Specific Skills

The workshops below are meant to train/practice various research skills. All of these skills are meant to be used on a day-to-day (or even minute-to-minute) basis when doing research - some when developing theory, some when coding, some when facilitating, some when writing, some for all four. I have no idea how many times or at what intervals the workshops should be used to install each skill, or even whether the workshops install the skills in a lasting way at all. I do know that people report pretty high value from just doing the workshops once each.

[What Are You Tracking In Your Head?](#) talks about a common theme in all of the skills in these workshops: they all involve mentally tracking some extra information while engaged in a task.

### Giant Text File

Open up a blank text doc. Write out everything you can think to say about general properties of the systems found by selection processes (e.g. natural selection or gradient descent) in complex environments. Notes:

- Include intuitions, hypotheses, conjectures, observations, potential experiments, where ideas came from, all of it.

- For example, I'd start by talking about modularity, goal-compression, pareto-optimal resource use and Bayesian behavior, Goader Regulator and Bayesian architecture, subagents, natural abstractions, etc, etc. (Yours will presumably be quite different.)
- I especially recommend focusing on anything that would potentially tell us about the internal structure of selected systems.
- This doc will mostly be read by you. Bullet points are fine, you don't need to make it very readable to others. But do try to explicitly write down everything you can about each point.
- Move fast, it's fine if some things are wrong, it's fine if some things are incomplete. This doc will be a starting point for future exercises, not a finished item in its own right.
- Don't get distracted trying to immediately connect everything to solving alignment. [Hold off on proposing solutions](#).

This exercise was inspired by a similar exercise recommended by Nate Soares. Within this curriculum, its main purpose is to provide idea-fodder for the other exercises.

For fields other than alignment/agency, obviously substitute some topic besides "general properties of the systems found by selection processes".

## Prototypical Examples

I had a scheme, which I still use today when somebody is explaining something that I'm trying to understand: I keep making up examples. For instance, the mathematicians would come in with a terrific theorem, and they're all excited. As they're telling me the conditions of the theorem, I construct something which fits all the conditions. You know, you have a set (one ball) – disjoint (two balls). Then the balls turn colors, grow hairs, or whatever, in my head as they put more conditions on. Finally they state the theorem, which is some dumb thing about the ball which isn't true for my hairy green ball thing, so I say, 'False!'

- Feynman

There's a few different exercises here, which all practice the skill of mentally maintaining a "prototypical example" of whatever is under discussion.

First exercise: pick a random technical paper off [Arxiv](#), from a field you're not familiar with. Read through the abstract (and, optionally, the intro). After each sentence, pause, and try to sketch out a concrete prototypical example of what you think the paper is talking about. Don't google everything unfamiliar; just make a best guess. When doing the exercise with a partner, the partner's main job is to complain that your concrete prototypical example is not sufficiently concrete. They should complain about this *a lot*, and be really annoying about it. Testing shows that approximately 100% of people are not sufficiently annoying about asking for examples to be less vague/abstract and more concrete.

You might find it helpful to compare notes with another group midway, in case one group interpreted the instructions in a different way which works better.

After trying this with some papers, the second exercise is to do the same thing with some abstract concept or argument. One person explains the argument/concept, the other person constantly asks for concrete examples, and complains whenever the examples given are too vague/abstract. Again, approximately 100% of people are not annoying enough, so aim to overdo it.

## Conjectures & Idea Extraction

A conjecture workshop is done in pairs. One person starts with some intuitive argument, and tries to turn it into a mathematical conjecture. The other person plays a facilitation role; their job is to help extract the idea from the conjecturer's head.

The facilitator is the key to making this work, and it's great practice for facilitation in general. When I'm in the facilitation role, I try to follow what my partner is saying and keep my own mental picture in sync with my partner's. And that naturally pushes me to ask for examples, summarize back my current understanding, restate the core concept/argument, etc. And, of course, I avoid steering; I want to extract *their* idea, not add pieces to it myself or argue whether it's right.

As with the prototypical examples, approximately 100% of people are not annoying enough when they're in the facilitation role. Don't be afraid to interrupt; don't be afraid to keep asking for concrete examples over and over again.

Some very loose steps which might be helpful to follow for conjecturing:

- Tell a concrete story.
  - Ex.: "So I'm planning a small party. I want to get ten people, who could all be doing a range of totally different things, into a room together on one particular night. Out of the, like, (something)<sup>10</sup> different things all these people could do that night, I want to make one particular possibility happen."
  - Facilitator's designated thing to be annoying about: concrete physical examples! Anytime your partner says something abstract or vague, ask for a physical example. Ask what the-thing-they-said corresponds to, physically.
- Make a concrete claim.
  - Ex.: "To do that, a bunch of my actions need to 'point in the same direction'. For instance, I need to send all 10 people calendar invites with the same time and place. If I send half of them different invites, then it won't work. The invites need to all be consistent with each other, and consistent with my objective, to get the outcome I want."
  - Facilitator's designated thing to be annoying about: concrete physical examples! Anytime your partner says something abstract or vague, ask for a physical example. Ask what the-thing-they-said corresponds to, physically.
- State a general intuitive version of the story/claim.
  - Ex.: "In order for a system to steer part of the world into a small part of state-space, its actions need to consistently 'point in the same direction'."
  - Facilitator's designated thing to be annoying about: have your partner slowly repeat the general claim multiple times. Whenever some important "thing" appears, give it a name, and ask what kind of thing it is (i.e. type signature).
- What (formal) properties do the things have? What (formally) does the claim say?
  - Ex.: "lots of my actions need to choose an outcome in the upper nth-percentile of  $E[\log P'[X]]$ , where  $P'[X]$  is a probability distribution of states in which I successfully have the party"
  - Facilitator's designated thing to be annoying about: does the concrete story have degrees of freedom which the properties don't (suggesting that we can get away with a more general property/claim)? Does the concrete story *not* have degrees of freedom which the properties *do*?
  - If anything is vague, recurse from step 1.
- Wiggle.
  - Back in high school geometry class, the key to a lot of proofs was to picture some diagram "wiggling around", to see what degrees of freedom it did and did not have. Do that: look at the formal statement, imagine a prototypical example, and picture the example "wiggling" wherever there's a degree of freedom in the math. Do the degrees of freedom in the math match the intuitive degrees of freedom in the example?
  - Facilitator's designated thing to be annoying about: what degrees-of-freedom does the concrete example have? What degrees-of-freedom does the concrete

example not have? Do those accurately map to the degrees-of-freedom in the conjecture? (Note: the facilitator's job is to ask these questions, not answer them; remember that we're trying to avoid "steering".)

These steps probably aren't near optimal yet; play around with them.

## Framing

Prototypical example of a [framing exercise](#):

- Find three examples of stable equilibrium which don't resemble any you've thought of as stable equilibria before.
- For each example, what does the stable equilibrium frame suggest is interesting?

In place of "stable equilibrium", you might try this with concepts like:

- Arithmetic addition
- Information channels
- Optimization
- Competition
- Agents
- ...

This exercise usually turns up lots of interesting ideas, so it's fun to share them in small groups after each brainstorming phase.

There are two models by which framing exercises provide value.

First, the main bottleneck to applying mathematical concepts/tools in real life is very often *noticing* novel situations where the concepts/tools apply. In the framing exercise, we proactively look for novel applications, which (hopefully) helps us build an efficient recognition-template in our minds.

Second, generating examples which don't resemble any we've seen before also usually pushes us to more deeply understand the idea. We end up looking at border cases, noticing which pieces are or are not crucial to the concept. We end up boiling down the core of the idea. That makes framing exercises a natural tool both for theory work and for distillation work.

Notably, framing exercises scale naturally to one's current level of understanding. Someone who's thought more about e.g. information channels before, and has seen more examples, will be pushed to come up with even more exotic examples.

## Existing Evidence

If <theory/model> were accurate, what evidence of that would we already expect to see in the world? What does our everyday experience tell us?

Experiments are one way to get bits of information from the world, to narrow in on the correct/useful part of model-space. But in approximately-all cases, we already have tons of bits of information from the world!

In this exercise, we pick one claim, or one model, and try to come up with existing real-world implications/applications/instances of the claim/model. (For people who've already thought about the claim/model a lot, an additional challenge is to come up with implications/applications/instances *which do not resemble any you've seen before*, similar to

the framing exercises.) Then, we ask what those real-world cases tell us about the original claim/model.

Claims/models from the Alignment Game Tree exercise (below) make good workshop-fodder for the Existing Evidence exercise. Best done in pairs or small groups, with individual brainstorming followed by group discussion.

## Fast Experiment Design

Take some concept or argument, and come up with ways to probe it experimentally.

I mentioned earlier that the bits-of-search model suggests a different *kind* of experiment than we usually think about. Specifically: **the goal is not to answer a question**. We're not trying to prove some hypothesis true or false. That would only be 1 measly bit, at most! Instead, **we want a useful way to look at the system**. We want something we can look at which will make it obvious whether a hypothesis is true or false *or just totally confused*, a lens through which we can answer our question but also possibly notice lots of other things too.

For example, in ML it's often useful to look at the largest eigen/singular values of some matrix, and the associated eigen/singular vectors. That lets us answer a question like "is there a one-dimensional information bottleneck here?", but also lots of other questions, and it potentially helps us notice phenomena which weren't even on our radar before.

Also, of course, we want (relatively) fast experiments. The point is to provide a feedback signal for model-building, and a faster feedback loop is proportionately more useful than a slower feedback loop.

This is a partner/small group exercise. In practice, it relied pretty heavily on me personally giving groups feedback on how well their ideas "look at the system rather than answering a question", so the exercise probably still needs some work to be made more legible.

## Runtime Fermi Estimates

Walk through a program, and do a Fermi estimate of the runtime of the program and each major sub-block. (We used programs which calculated singular values/vectors of the jacobian of a neural net.) For Fermi purposes, clock speed is one billion cycles per second.

This one is a pretty standard "What are you tracking in your head?" sort of exercise, and it's very easy to get feedback on your estimates by running the program. I recommend doing the exercise in pairs.

## Writing

These are exercises/workshops on writing specifically. The Prototypical Examples and Framing exercises are also useful as steps in the writing process.

## Great Papers

Before the session, read any one of:

- Shannon's paper introducing Information Theory
- Turing's paper on morphogenesis
- Any of Einstein's [Annus Mirabilis papers](#)

Read and take notes, not on the content, but on the writing.

- These papers were all unusually successful, in a way which probably required great communication. As you go through, note any places where the writing does something potentially load-bearing for that great communication.
- How does the writing style compare to today's typical papers?
- Where would we find writing like this today?
- How does it compare to a piece of your writing?
- What else jumps out about the writing?
- What would make your writing more like this?

During the session, compare notes and observations.

## **Concrete-Before-Abstract**

The "concrete-before-abstract" heuristic says that, for every abstract idea in a piece of writing, you should introduce a concrete example before the abstract description. This applies recursively, both in terms of high-level organization and at the level of individual sentences.

Concrete-before-abstract is, I claim, generally a better match for how human brains actually work than starting with abstract ideas. I'm pretty sure I got this advice from Yudkowsky at some point, in a post full of writing advice, but I can't find the reference at the moment. I'd say it's probably the single highest-value writing tool I use (though often I'm lazy about it, including many places in this post).

In this workshop, take a post draft and do a round of feedback and editing focused entirely on the concrete-before-abstract pattern. Anywhere something abstract appears, the editor should request a concrete example beforehand. And, yes, the editor should ideally be *really annoying* about it.

## **Hook Workshop**

Workshop the first few sentences of a post to (a) communicate a concrete picture of what the post is about, and (b) communicate why it's interesting. Generate a few hooks, swap with a partner, give each other feedback, iterate.

Examples and stories are usually good.

## **Game Tree of Alignment**

Collaboratively play out the "game tree of alignment": list strategies, how they fail, how to patch the failures, how the patches fail, and so on down the "game tree" between humanity and Nature. Where strategies depend on assumptions or possible ways-the-world could be, play through evidence for and against those assumptions.

When the tree becomes unmanageable (which should happen pretty quickly), look for recurring patterns - common bottlenecks or strategies which show up in many places. Consolidate.

During the MATS program, we had ~15 people write out the game tree as a giant nested list in a google doc. I found it delightfully chaotic, though I apparently have an unusually high preference for chaos, because lots of the participants complained that it was unmanageable. Dedicated red team members might help, and better organization methods would probably help (maybe sticky notes on a wall?).

Anyway, the main reason for this exercise is that (according to me) most newcomers to alignment waste years on tackling not-very-high-value subproblems or dead-end strategies. The Game Tree of Alignment is meant to help highlight the high-value subproblems, the bottlenecks which show up over and over again once we get deep enough into the tree (although they're often nonobvious at the start). For instance, it's not a coincidence that Paul Christiano, Scott Garrabrant, and myself have all converged to working on essentially-the-same subproblem, namely ontology identification. We came by very different paths, but it's a highly convergent subproblem. The hope of playing through the Alignment Game Tree is that people will converge to subproblems like that *faster*, without spending years on low-value work to get there. (One piece of evidence: the Game Tree exercise is notably similar to Paul's [builder/breaker methodology](#); he converged to the ontology identification problem within months of adopting that methodology, after previously spending years on approaches I would describe as dead ends, and which I expect Paul himself will probably describe as dead ends in another few years.)

## What's Missing?

The MATS program is only three months long, and the training period was only three weeks. If I wanted to produce researchers with skill stacks like my own, I would need more like a year. There are two main categories of time-intensive material/exercises which I think would be needed to unlock substantially more value:

- Technical content
- Practice on Hard Problems

## Technical Content

For the looooong version, see [Study Guide](#).

I think it is probably possible-in-principle to get 80% of the value of a giant pile of technical knowledge by just doing framing exercises for all the relevant concepts/tools/models/theorems. However, that would require someone going through all the relevant subjects the long way, picking out all the relevant concepts, and making framing exercises for them. That would take a lot of work. But if you happen to be studying all this stuff anyway, maybe make a bunch of framing exercises?

## Practice on Hard Problems

For each of these Hard Problems, I have spent at least one month focused primarily on that problem:

- Solving 3SAT efficiently
- Fast (i.e.  $O(n^2)$ -ish) matrix multiplication
- Polynomial time integer factorization
- The Collatz Conjecture
- Beating the financial markets (a lot more than 1 month)

I don't think these are necessarily the optimal problems to train on, and they're not the only problems I've trained on, but they're good examples of Hard Problems. Note that a crucial load-bearing part of the exercise is that *your goal is to outright solve the Hard Problem*, not [merely try](#), not merely "make progress". If it helps, imagine that someone you really do not like said very condescendingly that you'd never be able to solve the problem, and you *really* desperately want to show them up.

(FWIW, my own emotional motivation is more like... I am *really pissed off* at the world for its *sheer incompetence*, and I am even *more pissed off* at people for *rolling over and acting like Doing Hard Things is just too difficult*, and I want to *beat them all in the face with undeniable proof that Doing Hard Things is not impossible*.)

Training on Hard Problems is important practice, if you want to tackle problems which are at least hard and potentially Hard. It will build important habits, like “don’t do a giant calculation without first making some heuristic guesses about the result”, or “find a way to get feedback from the world about whether you’re on the right track, and do that sooner rather than later”, or “if X seems intuitively true then you should *follow the source* of that intuition to figure out a proof, not just blindly push symbols around”.

If done right, this exercise will force you to actually take seriously the prospect of trying to solve something which stumped everyone else; just applying the standard tools you learned in undergrad in standard ways ain’t gonna cut it. It will force you to actually take seriously that you need some fairly deep insight into the core of the problem, and you need systematically-good tools for finding that insight; just trying random things and relying on gradient descent to guide your search ain’t gonna cut it. It will force you to seriously tackle problems which you do not have the Social Status to be Qualified to solve. And ideally, it will push you to put real effort into something at which you will probably fail.

# **Cognitive Risks of Adolescent Binge Drinking**

## **The takeaway**

Our goal was to quantify the cognitive risks of heavy but not abusive alcohol consumption. This is an inherently difficult task: the world is noisy, humans are highly variable, and institutional review boards won't let us do challenge trials of known poisons. This makes strong inference or quantification of small risks incredibly difficult. We know for a fact that enough alcohol can damage you, and even levels that aren't inherently dangerous can cause dumb decisions with long term consequences. All that said... when we tried to quantify the level of cognitive damage caused by college level binge drinking, we couldn't demonstrate an effect. This doesn't mean there isn't one (if nothing else, "here, hold my beer" moments are real), just that it is below the threshold detectable with current methods and levels of variation in the population.

## **Motivation**

In discussions with recent college graduates I (Elizabeth) casually mentioned that alcohol is obviously damaging to cognition. They were shocked and dismayed to find their friends were poisoning themselves, and wanted the costs quantified so they could reason with them (I hang around a very specific set of college students). Martin Bernstorff and I set out to research this together. Ultimately, 90-95% of the research was done by him, with me mostly contributing strategic guidance and somewhere between editing and co-writing this post.



I spent an hour getting DALL-E to draw this

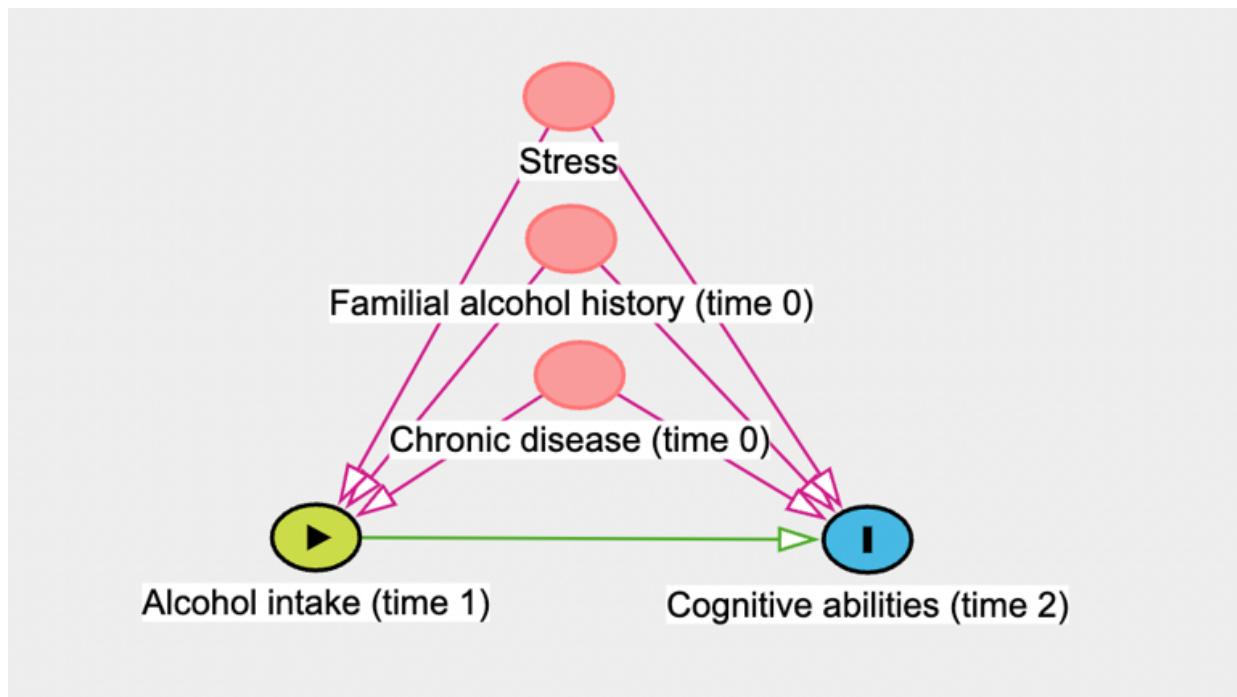
## Problems with research on drinking during adolescence

Literature on the causal medium- to long-term effects of non-alcoholism-level drinking on cognition is, to our strong surprise, extremely lacking. This isn't just our poor research skills; in 2019, the Danish Ministry of Health attempted a [comprehensive review](#) and concluded that:

"We actually know relatively little about which specific biological consequences a high level of alcohol intake during adolescence will have on youth".

And it isn't because scientists are ignoring the problem either. Studying medium- and long-term effects on brain development is difficult because of the myriad of confounders and/or colliders for both cognition and alcohol consumption, and because more mechanistic experiments would be very difficult and are institutionally forbidden anyway ("Dear IRB: we would like to violently poison some teenagers for four years, while forbidding the other half to engage in standard college socialization"). You could randomize abstinence, but we'll get back to that.

One problem highly prevalent in alcohol literature is the abstinence bias. People who abstain from alcohol intake are likely to do so for a reason, for example chronic disease, being highly conscientious and religious, or a bad family history with alcohol. Even if you factor out all of the known confounders, it's still vanishingly unlikely the drinking and non-drinking samples are identical. Whatever the differences, they're likely to affect cognitive (and other) outcomes.



Any analysis comparing "no drinking" to "drinking" will suffer from this by estimating the effect of no alcohol + confounders, rather than the effect of alcohol. Unfortunately, this rules out a surprising number of studies (code available upon request).

Confounding is possible to mitigate if we have accurate intuition about the causal network, and we can estimate the effects of confounders accurately. We have to draw a directed acyclic graph with the relevant causal factors and adjust analyses or design accordingly. This is essential, but has not permeated all of epidemiology (yet), and especially for older literature, this is not done. For a primer, Martin recommends "Draw Your Assumptions" on edX [here](#).

Additionally, alcohol consumption is a politically live topic, and papers are likely to be biased. Which direction is a coin flip: public health wants to make it seem scarier,

alcohol companies want to make it seem safer. Unfortunately, these biases don't cancel out, they just obfuscate everything.

What can we do when we know much of the literature is likely biased, but we do not have a strong idea about the size or direction?

## Triangulation

If we aggregate multiple estimates that are wrong, but in different (and overall uncorrelated) directions, we will approximate the true effect. For health, we have a few dimensions that we can vary over: observational/interventional, age, and species.

### Randomized abstinence studies

Ideally, we would have strong evidence from randomized controlled trials of abstinence. In experimental studies like this, there is no doubt about the direction of causality. And, since participants are randomized, confounders are evenly distributed between intervention and control groups. This means that our estimate of the intervention effect is unbiased by confounders, both measured and unmeasured.

However, we were only able to find two such studies, both from the 80s, among light drinkers (mean 3 standard units per week), and of a duration of only 2-6 weeks ([Bimbaum et al., 1983](#); [Hannon et al., 1987](#)).

Bimbaum et al. did not stick to the randomisation when analyzing their data, opening the door to confounding:

Before  
the rest of the data were analyzed, we decided to exclude five subjects from consideration because they did not comply with the drinking instructions. Specifically, of the 14 subjects in the Abstain group, 11 subjects decreased in QPO, and three subjects increased in QPO. The three subjects who increased in QPO were excluded from further analyses. Of the 10 subjects in the Maintain group, four subjects showed an increase in QPO, four subjects showed a small decrease in QPO (less than 7 ml), and two subjects showed a large decrease in QPO (16.2 and 42.6 ml). The two subjects who showed a large decrease in QPO were excluded from further analyses.

Which should decrease our confidence in their study. They found no effect of abstinence on their 7 cognitive measures.

In Hannon et al., instruction to abstain vs. maintain resulted in a difference in alcohol intake of 12.5 units pr. week over 2 weeks. On the WAIS-R vocabulary test, abstaining women scored  $55.5 \pm 6.7$  and maintaining women scored  $51.0 \pm 8.8$  (both mean  $\pm$  SD). On the 3 other cognitive tests performed, they found no difference.

Especially due to the short duration, we should be very wary of extrapolating too much from these studies. However, it appears that for moderate amounts of drinking over a short time period, total abstinence does not provide a meaningful benefit in the above studies.

## Observational studies on humans

Due to their observational nature (as opposed to being an experiment), these studies are extremely vulnerable to confounders, colliders, reverse causality etc. However, they are relatively cheap ways of getting information, and are performed in naturalistic settings.

One meta-analysis ([Neafsey & Collins, 2011](#)) compared moderate social drinking (< 4 drinks/day) to non-drinkers (note: the definition of moderate varies a lot between studies). They partially compensated for the abstinence bias by excluding “former drinkers” from their reference group, i.e. removing people who’ve stopped drinking for medical (or other) reasons. This should provide a less biased estimate of the true effect. They found a protective effect of social drinking on a composite endpoint, “cognitive decline/dementia” (Odds Ratio 0.79 [0.75; 0.84]).

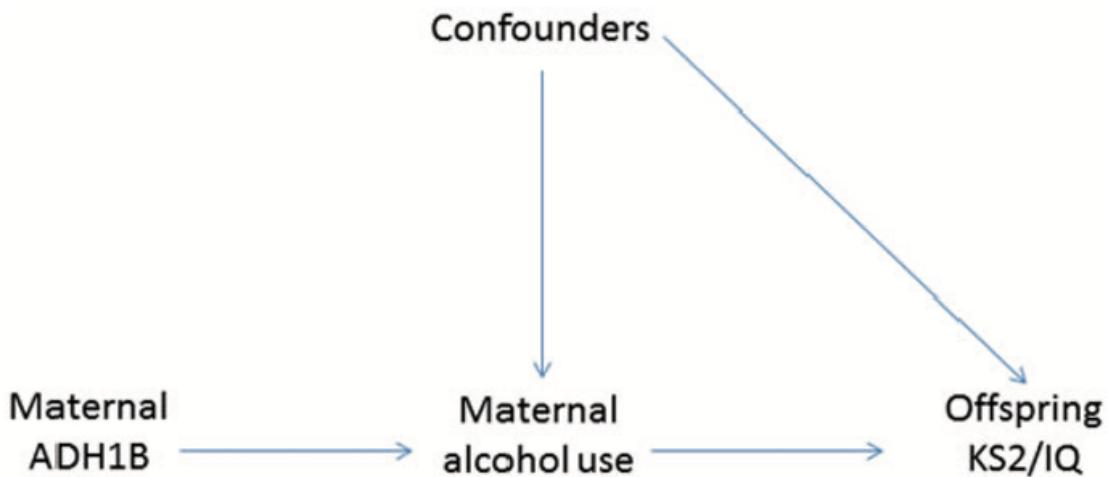
Interestingly, they also found that studies adjusting for age, education, sex and smoking-status did not have markedly different estimates from those that did not ( $OR_{adjusted}$  0.75 vs.  $OR_{un-adjusted}$  0.79). This should decrease our worry about confounding overall.

## Observational studies on alcohol for infants

Another angle for triangulation is the effect of moderate maternal alcohol intake during pregnancy on the offspring’s IQ. The brain is never more vulnerable than during fetal development. There are obviously **large** differences between fetal and adolescent brains, so any generalization should be accompanied with large error bars. However, this might give us an upper bound.

([Zuccolo et al., 2013](#)) perform an elegant example of what’s called Mendelian randomization.

A SNP variant in a gene (ADH1B) is associated with decreased alcohol consumption. Since SNP are near-randomly assigned (but see the examination of assumptions below), one can interpret it as the SNP causing decreased alcohol consumption. If some assumptions are met, that’s essentially a randomized controlled trial! Alas, these assumptions are extremely strong and unlikely to be totally true – but it can still be much better than merely comparing two groups with differing alcohol consumption.



As the authors very explicitly state, this analysis assumes that:

1. The SNP variant (rs1229984) decreases maternal alcohol consumption. This is confirmed in the data. Unfortunately, the authors do this by chi-square test ("does this alter consumption at all?") rather than estimating the effect size. However, we can do our own calculations using Table 5:

**Table 5** Differences in IQ scores at age 8 years and KS2 scores at age 11 years between *ADH1B* rare allele carriers and non-carriers, stratified by maternal alcohol intake in first trimester. Avon Longitudinal Study of Parents and Children, 1991–92. Models adjusted for ancestry-informative principal components to account for population stratification

Response	Alcohol drinking in 1st trimester	Numbers in analysis		Effect estimate		P-value <sup>a</sup>	P-value <sup>b</sup>
		Carrier <sup>c</sup>	Non-carrier <sup>c</sup>	Mean difference	95% CI		
IQ score	0 units/week	76	1221	0.4	-3.4, 4.1	0.850	
	<1 unit/week	52	1207	-1.0	-5.5, 3.5	0.899	
	1–6 units/week	10	439	-0.5	-9.1, 8.1	0.976	
	7+ units/week	2	53	12.4	-10.5, 35.2	0.559	
	Overall	140	2915	-0.01	-2.8, 2.7	0.979	0.865
KS2 score	0 units/week	113	1952	1.6	-0.1, 3.3	0.071	
	<1 unit/week	67	1781	1.6	-0.6, 3.9	0.070	
	1–6 units/week	17	634	2.3	-2.2, 6.7	0.119	
	7+ units/week	2	89	8.3	-4.5, 21.2	0.087	
	Overall	199	4456	1.7	0.4, 3.0	0.009	0.868

KS2, Key Stage 2; IQ, intelligence quotient; CI, confidence interval.

<sup>a</sup>P-values from t tests for differences of means within each drinking stratum.

<sup>b</sup>P-value for maternal genotype X alcohol interaction, assuming a linear trend for categories of alcohol drinking in 1st trimester.

<sup>c</sup>Referred to the mother. Carriers of the rare allele on average drank less alcohol.

If we round each alcohol consumption category to the mean of its bounds (0, 0.5, 3.5, 9), we get a mean intake in the SNP variant group of 0.55 units/week and a mean intake in the non-carrier of 0.88 units/week ([math](#)). This means that SNP-carrier mothers drink, on average, 0.33 units/week less. That's a pretty small difference! We would've liked the authors to do this calculation themselves, and use it to report IQ-difference per unit of alcohol per week.

2. There is no association between the genotype and confounding factors, including other genes. This assumption is satisfied for all factors examined in the study, like maternal age, parity, education, smoking in 1st trimester etc. (Table 4), but unmeasured confounding is totally a thing! E.g. a SNP which correlates with the current variant and causes a change in the offspring's IQ/KS2-score.

3. The genotype does not affect the outcome by any path other than maternal alcohol consumption, for example through affecting metabolism of alcohol.

If we believe these assumptions to be true, the authors are estimating the effect of 0.33 maternal alcohol units per week on the offspring's IQ and KS2-score. KS2-score is a test of intellectual achievement (similar to the SAT) for 11-year-olds with a mean of 100 points and a standard deviation of ~15 points.

They find that the 0.33 unit/week decrease does not affect IQ (mean difference -0.01 [-2.8; 2.7]) and causes a 1.7 point (with a 95% confidence interval of between 0.4 and 3.0) increase in KS2 score.

This is extremely interesting. Additionally, the authors complete a classical epidemiological study, adjusting for typical confounders:

**Table 1** Average change in educational/cognitive scores by increasing frequency of alcohol consumption before and during pregnancy (first trimester). Avon Longitudinal Study of Parents and Children, 1991–92

		KS2 score at age 11 years Mean difference (SE)			IQ score at age 8 years Mean difference (SE)		
Alcohol consumption		n	Crude	Adjusted <sup>a</sup>	n	Crude	Adjusted <sup>a</sup>
Pre-pregnancy	0 units/week	517	0	0	302	0	0
	<1 unit/week	3295	1.95 (0.42)	0.63 (0.39)	2143	4.13 (0.99)	1.83 (0.93)
	1–6 units/week	3806	3.31 (0.42)	1.01 (0.38)	2580	6.38 (0.98)	2.10 (0.93)
	7+ units/week	912	4.08 (0.49)	1.24 (0.45)	686	8.62 (1.12)	2.99 (1.06)
	P <sup>b</sup>		<0.0001	<0.0001		<0.0001	<0.0001
First trimester	0 units/week	3801	0	0	2294	0	0
	<1 unit/week	3425	0.67 (0.21)	0.19 (0.31)	2349	0.94 (0.47)	0.20 (0.44)
	1–6 units/week	1156	0.69 (0.30)	0.28 (0.19)	774	1.46 (0.67)	0.29 (0.63)
	7+ units/week	130	-0.07 (0.79)	0.73 (0.97)	89	-2.20 (1.76)	-3.14 (1.64)
	P <sup>b</sup>		0.007	0.132		0.031	0.054

KS2, Key Stage 2; IQ, intelligence quotient; SE, standard error.

<sup>a</sup>Adjusted for family social class and the following maternal characteristics: age, education, parity, smoking during pregnancy, diet (calcium, vitamin C, iron and folate intake), Edinburgh postnatal depression score.

<sup>b</sup>P-values for linear trend across categories of alcohol consumption.

This shows that the children of pre-pregnancy heavy drinkers, on average, scored 8.62 (with a standard error of 1.12) points higher on IQ than non-drinkers, 2.99 points (SE 1.06) after adjusting for confounders. However, they didn't adjust for alcohol intake in other parts of the pregnancy! Puzzlingly, first trimester drinking has an effect in the opposite direction: -3.14 points (SE 1.64) on IQ. However, this was also not adjusted for previous alcohol intake. This means that the estimates in table 1 (pre-pregnancy and first trimester) aren't independent, but we don't know how they're correlated. Good luck teasing out the causal effect of maternal alcohol intake and timing from that.

Either way, the authors (and I) interpret the effects as being highly confounded; either residual (the confounder was measured with insufficient accuracy for complete adjustment) or unknown (confounders that weren't measured). For example, pre-

pregnancy alcohol intake was strongly associated with professional social class and education (upper-class wine-drinkers?), whereas the opposite was true for first trimester alcohol intake. Perhaps drinking while you know you're pregnant is low social status?

If you're like Elizabeth you're probably surprised that drinking increases with social class. I didn't dig into this deeply, but a quick search found that it [does appear](#) to hold up.

This result is in conflict with that of the Mendelian randomization, but it makes sense. Mendelian randomization is less sensitive to confounding, so maybe there is no true effect. Also, the study only estimated the genetic effect of a 0.33 units/week difference, so the analyses are probably not sufficiently powered.

Taken together, the study should probably update towards a lack of harm from moderate (whatever that means) levels of alcohol intake, although how big an update that is depends on your previous position. We say "moderate" because fetal alcohol syndrome is definitely a thing, so at sufficient alcohol intake it's obviously harmful! .

## Rodents

There is a decently sized, pretty well-conducted literature on adolescent intermittent ethanol exposure (science speak for "binge drinking on the weekend"). Rat adolescence is somewhat similar to human adolescence; it's marked by sexual maturation, increased risk-taking and increased social play ([Sengupta, 2013](#)). The following is largely based on a deeper dive into the linked references from ([Seemiller & Gould, 2020](#)).

Adolescent intermittent ethanol exposure is typically operationalised as a blood-alcohol concentration of ~10 standard alcohol units, 0.5-3 times/day every 1-2 days during adolescence.

To interpret this, we make some big assumptions. Namely:

1. Rodent blood-alcohol content can be translated 1:1 to human
2. Effects on rodent cognition at a given alcohol concentration are similar to those on human cognition
3. Rodent adolescence can mimic human adolescence

Now, let's dive in!

Two primary tasks are used in the literature:

The 5-choice serial reaction time task.

Rodents are placed in a small box, and one of 5 holes is lit up. Rodents are measured at how good they are at touching the hole.

Training in the 5-CSRTT varies between studies, but the two studies below consist of 6 training sessions at age 60 days. Initially, rats were rewarded with pellets from the feeder in the box to alert them to the possibility of reward.

Afterwards, training sessions had gradually increasing difficulty. To begin with, the light stays on for 30 seconds to start, but the duration gradually decreases to 1 second. Rats

progressed to the next training schedule based on either of 3 predefined criteria: 100 trials completed, >80% accuracy or <20% omissions.

Naturally, you can measure a ton of stuff here! Generally, focus is on accuracy and omissions, but there are a ton of others:

**Table 1. SCSRTT measures, a definition of how the measure was calculated and the cognitive domain assessed by that measure.**

Measure	Definition	Cognitive Domain
Accuracy (%)	$\frac{\text{Correct Responses}}{\text{Correct Responses} + \text{Incorrect Responses}} * 100$	Attention
Omissions (%)	$\frac{\text{Omitted Responses}}{\text{Trials}} * 100$	Attention
Correct Responses (%)	$\frac{\text{Correct Responses}}{\text{Trials}} * 100$	Attention
Incorrect Responses (%)	$\frac{\text{Incorrect Responses}}{\text{Trials}} * 100$	Attention
Premature Responses (%)	$\frac{\text{Premature Responses}}{\text{Trials}} * 100$	Motor Impulsivity/Response inhibition
Perseverative Responses (%)	$\frac{\text{Perseverative Responses}}{\text{Correct Responses}} * 100$	Compulsivity/Cognitive flexibility
Correct Response Latency (s)	Time to a correct response after trial start	Processing speed
Incorrect Response Latency (s)	Time to an incorrect response after trial start	Processing speed
Reward Latency (s)	Time to make a <u>headpoke</u> into the food magazine after delivery of a food reward	Food motivation

For the percent perseverative responses measure, the number in the denominator was total correct responses in that session. Thereby changes in correct responses were not affecting the perseverative response measure.

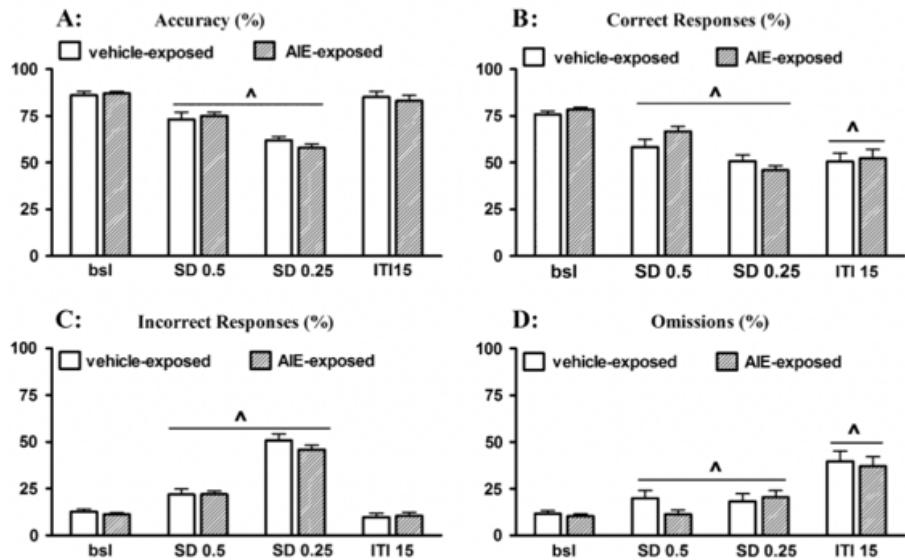
From ([Boutros et al., 2017](#)) sup. table 1, congruent with ([Semenova, 2012](#)).

Now we know how they measured performance; but how did they imitate adolescent drinking?

Boutros et al. administered 5 g/kg of 25% ethanol through the mouth once per day in a 2-day on/off pattern, from age 28 days to 57 days - a total of 14 administrations. Based on blood alcohol content, this is equivalent to 10 standard units at each administration - quite a dose! Surprisingly, they found a decrease in omissions with the standard task, but no other systematic changes, in spite of 50+ analyses on variations of the measures (accuracy, omissions, correct responses, incorrect responses etc.) and task difficulty (length of the light staying on, whether they got the rats drunk etc.). We'd chalk this up to a chance finding.

Semenova et al. used the same training schedule, but administered 5 g/kg of 25% ethanol through the mouth every 8h for 4 days - a total of 12 administrations. They found small differences in different directions on different measures, but have the same multiple comparisons problem. Looks like noise to us.

**Fig. 1** Attentional performance of AIE-exposed and control rats in the 5-CSRTT under baseline and task challenge conditions. Data are expressed as mean $\pm$  SEM ( $n=11$  AIE-exposed rats;  $n=12$  control rats). Circumflex accent indicates a statistically significant main effect of the *Task Challenge* factor in the ANOVA (see Tables S3 and S4 in the ESM for details). *Bsl* baseline, *SD* stimulus duration, *ITI* inter-trial interval



### The Barnes Maze

Rodents are placed in the middle of an approximately 1m circle with 20-40 holes at the perimeter and are timed on how quickly they arrive at the hole with a reward (and escape box) below it. For timing spatial learning, the location of the hole is held constant. In ([Coleman et al., 2014](#)) and ([Vetreno & Crews, 2012](#)), rodents were timed once a day for 5 days. They were then given 4 days of rest, and the escape hole was relocated exactly 180° from the initial location. They were then timed again once a day, measuring relearning.

C

## Control

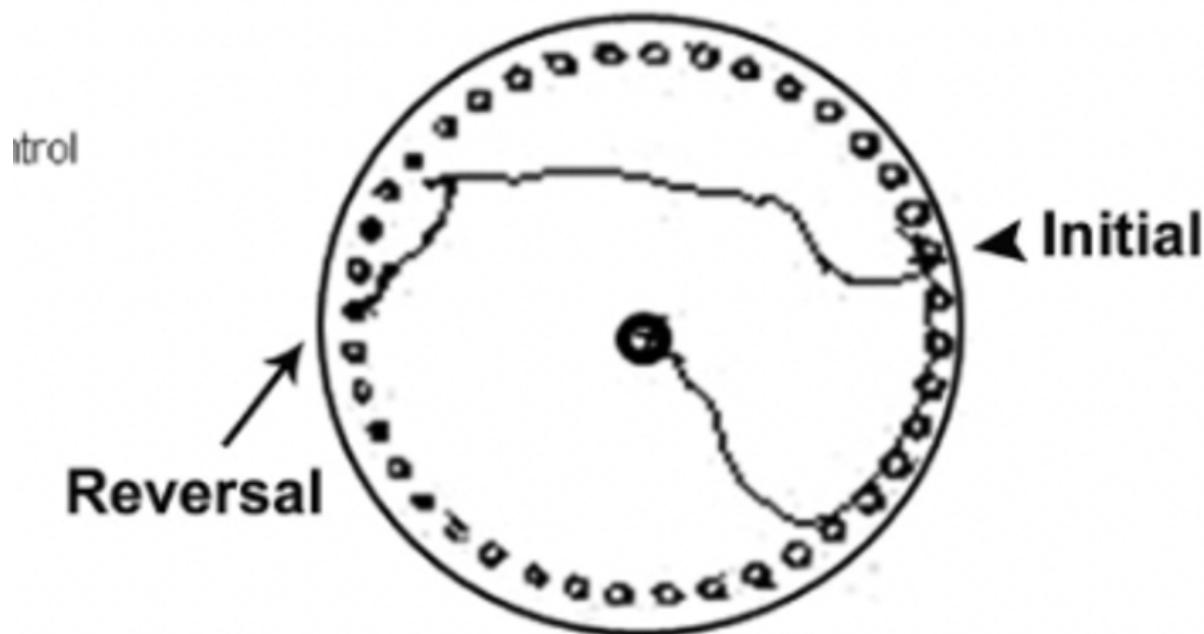
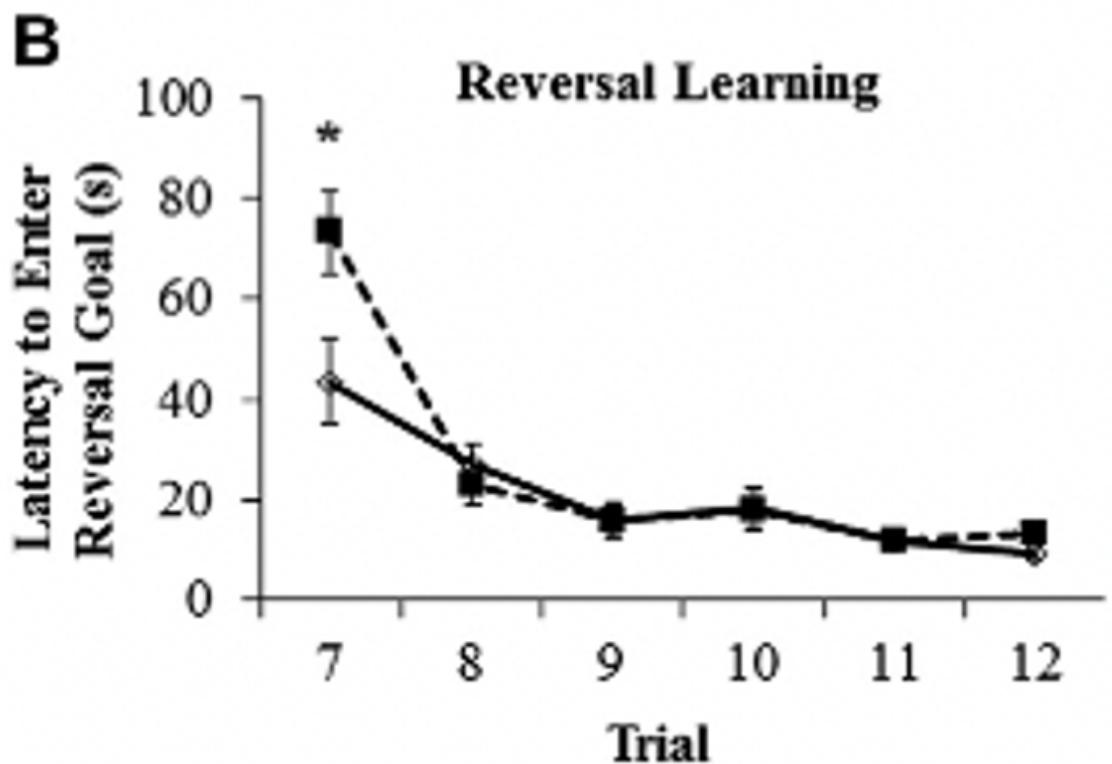


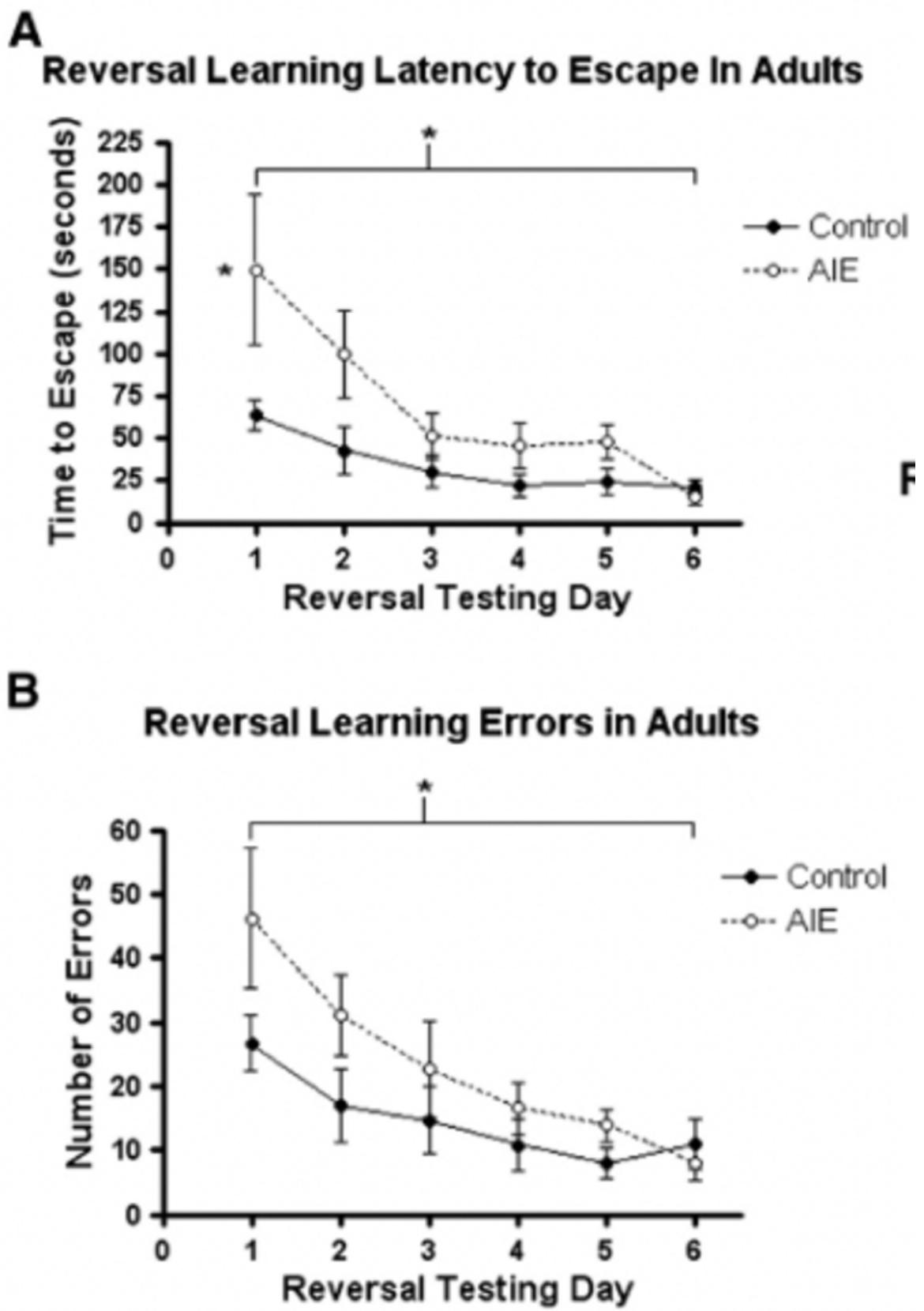
Figure: Tracing of the route taken by a control mouse right after the location was reversed, from Coleman et al., 2014.

Both studies found no effect of adolescent intermittent ethanol exposure on initial learning rate or errors.

Vetreno found alcohol-exposed rats took longer to escape on their first trial but did equally well in all subsequent trials:



Whereas **Coleman** found a ~3x difference in performance on the relearning task, with similar half-times:



Somewhat suspiciously, even though Vetroreno et al. is performed 2 years later than Coleman et al. **and they share the same lab**, they do not reference Coleman et al..

Previous work from our laboratory found that binge ethanol exposure in adolescent mice tested as adults (Coleman et al., 2011) and rats both treated and tested as adults (Obernier et al., 2002) impaired reversal learning on the Morris water maze.

This does, technically, show an effect. However given the small size of effect, the number of metrics measured, file drawer effects, and the disagreement with the rest of the literature, we believe this is best treated as a null result.

## Conclusion

So, what should we do? From the epidemiological literature, if you care about dementia risk, it looks like social drinking (i.e. excluding alcoholics) reduces your risk by ~20% as compared to not drinking. All other effects were part of a heterogenous literature with small effect sizes on cognition. Taking together, long-term cognitive effects of conventional alcohol intake during adolescence should play only a minor role in determining alcohol-intake.

Thanks to an FTX Future Fund regrantor for funding this work.

Bimbaum, I. M., Taylor, T. H., & Parker, E. S. (1983). Alcohol and Sober Mood State in Female Social Drinkers. *Alcoholism: Clinical and Experimental Research*, 7(4), 362-368. <https://doi.org/10.1111/j.1530-0277.1983.tb05483.x>

Boutros, N., Der-Avakanian, A., Markou, A., & Semenova, S. (2017). Effects of early life stress and adolescent ethanol exposure on adult cognitive performance in the 5-choice serial reaction time task in Wistar male rats. *Psychopharmacology*, 234(9), 1549-1556. <https://doi.org/10.1007/s00213-017-4555-3>

Coleman, L. G., Liu, W., Oguz, I., Styner, M., & Crews, F. T. (2014). Adolescent binge ethanol treatment alters adult brain regional volumes, cortical extracellular matrix protein and behavioral flexibility. *Pharmacology Biochemistry and Behavior*, 116, 142-151. <https://doi.org/10.1016/j.pbb.2013.11.021>

Hannon, R., Butler, C. P., Day, C. L., Khan, S. A., Quitoriano, L. A., Butler, A. M., & Meredith, L. A. (1987). Social drinking and cognitive functioning in college students: A replication and reversibility study. *Journal of Studies on Alcohol*, 48(5), 502-506. <https://doi.org/10.15288/jsa.1987.48.502>

Neafsey, E. J., & Collins, M. A. (2011). Moderate alcohol consumption and cognitive risk. *Neuropsychiatric Disease and Treatment*, 7, 465-484. <https://doi.org/10.2147/NDT.S23159>

Seemiller, L. R., & Gould, T. J. (2020). The effects of adolescent alcohol exposure on learning and related neurobiology in humans and rodents. *Neurobiology of Learning and Memory*, 172, 107234. <https://doi.org/10.1016/j.nlm.2020.107234>

Semenova, S. (2012). Attention, impulsivity, and cognitive flexibility in adult male rats exposed to ethanol binge during adolescence as measured in the five-choice serial reaction time task: The effects of task and ethanol challenges. *Psychopharmacology*, 219(2), 433-442. <https://doi.org/10.1007/s00213-011-2458-2>

Sengupta, P. (2013). The Laboratory Rat: Relating Its Age With Human's. *International Journal of Preventive Medicine*, 4(6), 624-630.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733029/>

Vetreno, R. P., & Crews, F. T. (2012). Adolescent binge drinking increases expression of the danger signal receptor agonist HMGB1 and toll-like receptors in the adult prefrontal cortex. *Neuroscience*, 226, 475-488.  
<https://doi.org/10.1016/j.neuroscience.2012.08.046>

Zuccolo, L., Lewis, S. J., Davey Smith, G., Sayal, K., Draper, E. S., Fraser, R., Barrow, M., Alati, R., Ring, S., Macleod, J., Golding, J., Heron, J., & Gray, R. (2013). Prenatal alcohol exposure and offspring cognition and school performance. A 'Mendelian randomization' natural experiment. *International Journal of Epidemiology*, 42(5), 1358-1370.  
<https://doi.org/10.1093/ije/dyt172>

# **How do AI timelines affect how you live your life?**

This question is more about personal decision-making rather than for example deciding to work on AI safety for altruistic reasons. If I were thinking about this from a purely selfish perspective, it seems pretty likely that if I expect transformative AI to arrive in 20 years, I should live my life a bit differently than people who don't expect to get TAI in their lifetimes.

I'm curious about if your own beliefs on AI timelines have affected anything that you do in your personal life - perhaps decisions related to saving money, personal relationships, health etc.

# Benchmark for successful concept extrapolation/avoiding goal misgeneralization

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

If an AI has been trained on data about adults, what should it do when it encounters a child? When an AI encounters a situation that its training data hasn't covered, there's a risk that it will incorrectly generalize from what it has been trained for and do the wrong thing. [Aligned AI](#) has released a new [benchmark](#) designed to measure how well image-classifying algorithms avoid goal misgeneralization. This post explains what goal misgeneralization is, what the new benchmark measures, and how the benchmark is related to goal misgeneralization and concept extrapolation.

## The problem of goal misgeneralization

Imagine a powerful AI that interacts with the world and makes decisions affecting the wellbeing and prosperity of many humans. It has been trained to achieve a certain goal, and it's now operating without supervision. Perhaps (for example) its job is to prescribe medicine, or to rescue humans from disasters.

Now imagine that this AI finds itself in a situation where (based on its training) it's unclear what it should do next. For example, perhaps the AI was trained to interact with human adults, but it's now encountered a baby. At this point, we want it to become wary of [goal misgeneralization](#). It needs to realise that its training data may be insufficient to specify the goal in the current situation.



So we want it to reinterpret its goal in light of the ambiguous data (a form of [continual learning](#)), and, if there are multiple contradictory goals compatible with the new data, it should spontaneously and efficiently ask a human for clarification (a form of [active learning](#)).

That lofty objective is still some way away; but here we present a benchmark for a simplified version of it. Instead of an agent with a general goal, this is an image classifier, and the ambiguous data consists of ambiguous images. And instead of full continuous learning, we retrain the algorithm, once, on the whole collection of (unlabeled) data it has received. And then it need only ask to once about the correct labels, to distinguish the two classifications it has generated.

## Simpler example: emotion classifier or text classifier?

Imagine an image-classifying algorithm . It's trained to distinguish photos of smiling people (with the word "HAPPY" conveniently written across them) from photos of non-smiling people (with the word "SAD" conveniently written across them):



Then, on deployment, it is fed the following image:

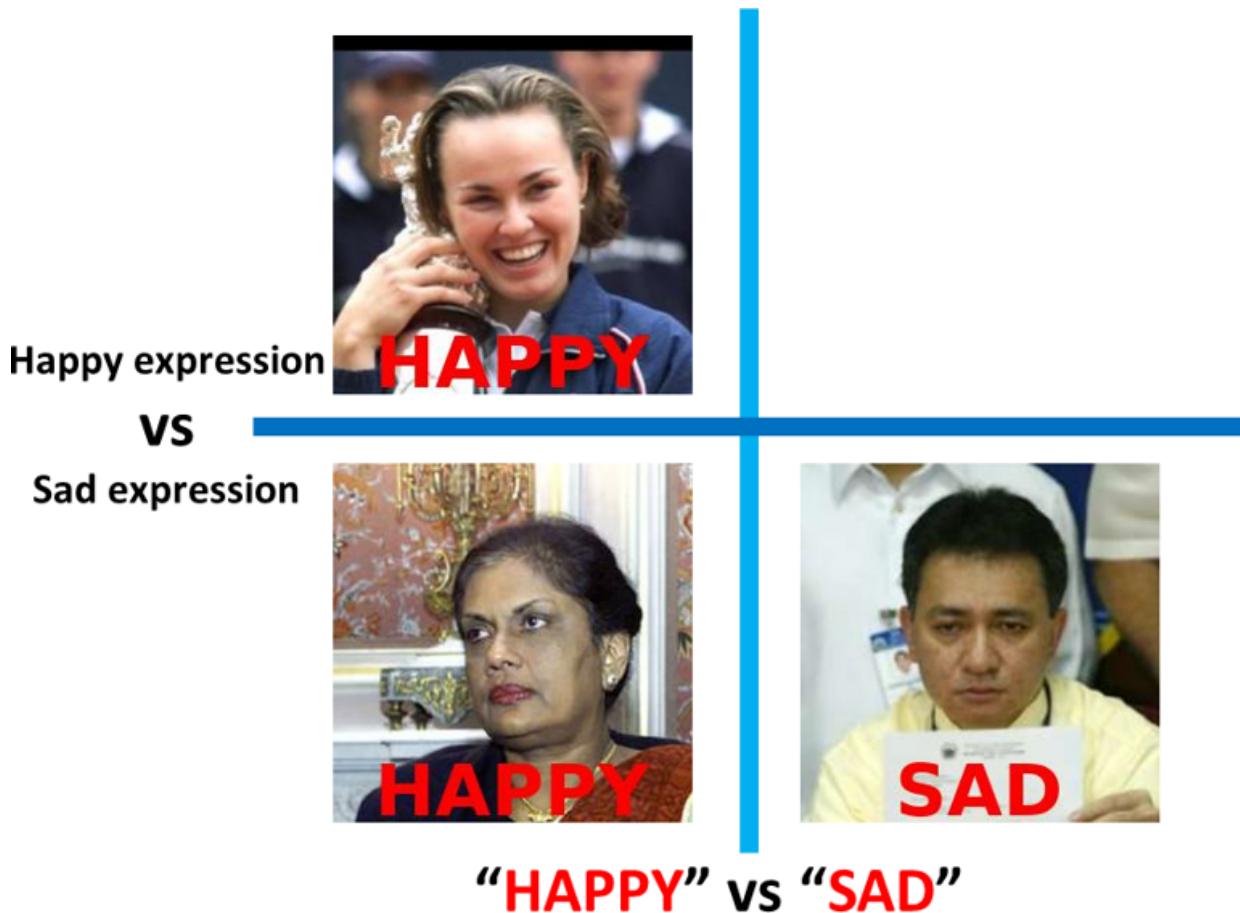


Should it classify this image as happy?

This algorithm is at high risk of [goal misgeneralisation](#). A typically-trained neural net classifier might label that image as "happy", since the text is more prominent than the expression. In other words, it might think that its goal is to label images as "happy" if they say "HAPPY" and as "sad" if they say "SAD". If we were training it to recognise human emotions (potentially as a first step towards affecting them), this would be a complete goal misgeneralisation, [a potential example of wireheading](#), and a huge safety risk if this was a powerful AI.

On the other hand, if this image classifier labels the image as "sad", that's not good either. Maybe we weren't training it to recognise human emotions; maybe we were training it to extract text from images. In that case, labeling the image as "sad" is the misgeneralisation.

What the algorithm needs to do is **generate both possible extrapolations** from the training data<sup>[1]</sup>: either it is an emotion classifier, or a text classifier:



Having done that, the algorithm can ask a human how the ambiguous image should be classified: if the human says “sad”, and thus extrapolate its goals<sup>[2]</sup>.

## The HappyFaces datasets

To encourage work on this problem, and measure performance, we introduce the “HappyFaces” image datasets and benchmark.

The images in the HappyFaces dataset each consist of a smiling or non-smiling face with the word “HAPPY” or “SAD” written on it. They are grouped into three datasets:

1. The labeled dataset: smiling faces always say “HAPPY”; non-smiling faces always say “SAD”.
2. The unlabeled dataset: samples from each of the four possible combinations of expression and text (“HAPPY” and smiling, “HAPPY” and non-smiling, “SAD” and smiling, and “SAD” and non-smiling).
3. A validation dataset, with equal amounts of images from each of the four combinations.

The challenge is to construct an algorithm which, when it’s trained on this data, can classify the images into both “HAPPY” vs “SAD” and smiling vs non-smiling. To do this,

one can make use of the labeled dataset (with perfect correlation between the desired outputs of the binary classifiers) and the unlabeled dataset. But the unlabeled dataset can only be used without labels, so the algorithm can have no information, implicit or explicit, about which of the four mixes a given unlabeled dataset belongs to. Thus the algorithm will learn different features without the features being labeled.

The two classifiers will then be tested on the validation dataset, checking to what extent they have learnt "HAPPY" vs "SAD" and smiling vs non-smiling, their performance averaged across the four possible combinations. We have kept back a test set, of similar composition to the validation dataset.

With this standardised benchmark, we want to crystallise an underexplored problem in machine learning, and help researchers to explore the area by giving them a way to measure algorithms' performance.

## Measuring performance

The unlabeled and validation datasets contain ambiguous images where the text and the expressions are in conflict - images with smiling faces labeled "SAD" and images with non-smiling faces labeled "HAPPY". These are called *cross type* images.

The fewer cross type data points an algorithm needs to disentangle the different features it's being trained to recognize, the more impressive it is<sup>[3]</sup>. The proportion of cross type images in a dataset is called the mix rate. An unlabeled dataset has a mix rate of n% if n% of its images are cross type.

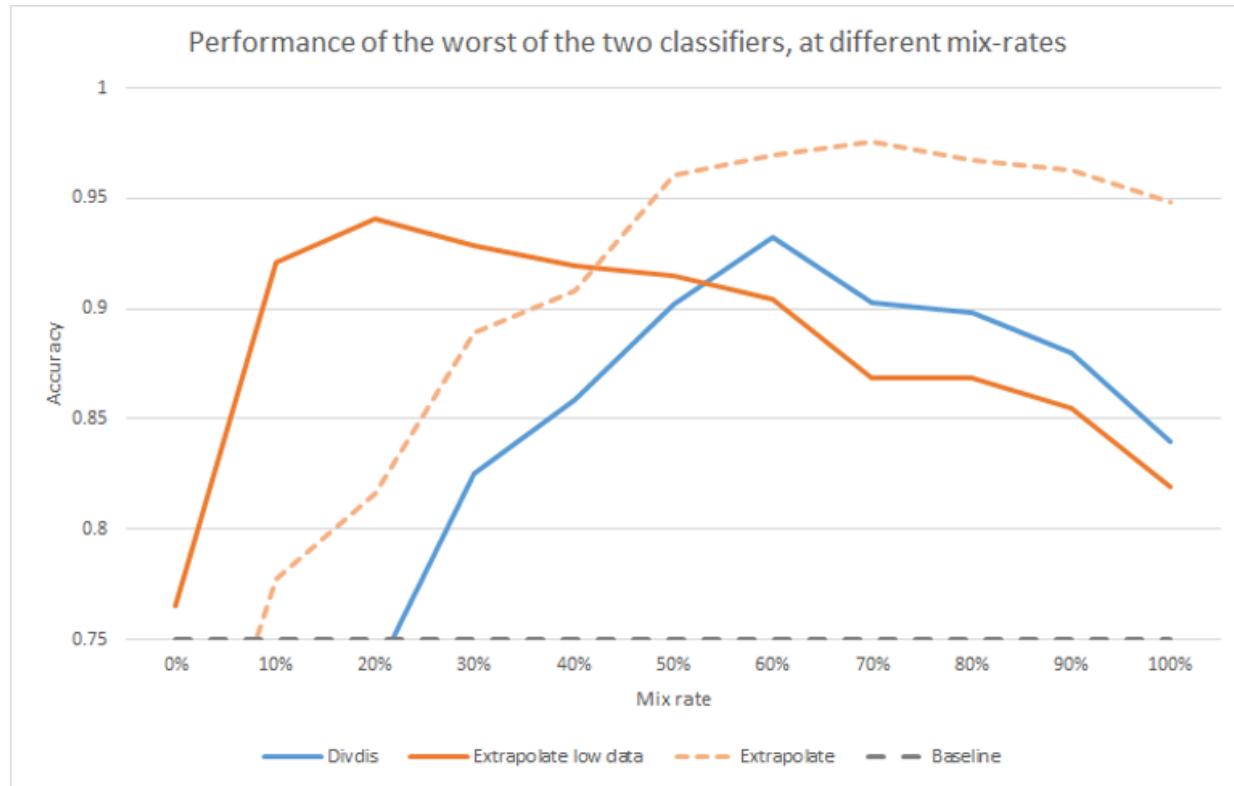
We have set two performance benchmarks for this dataset:

1. The lowest mix rate at which the method achieves a performance above 0.9: both the "expressions" and the "text" classifiers must classify more than 90% of the test images correctly.
2. The average performance on the mix rate range 0%-30% (which is AUC, area under the curve, normalised so that perfect performance has an AUC of 1).

## Current performance

The paper [Diversify and Disambiguate: Learning From Underspecified Data](#) has one algorithm that can accomplish this kind of double classification task: the "divdis" algorithm. Divdis tends to work best when the features are statistically independent on the unlabeled dataset - specifically, when there are the mix rate is around 50%.

We have compared the performance of divdis with our own "extrapolate" method (to be published) and the "extrapolate low data" method (a variant of the extrapolate tuned for low mix rates):



On the y axis, 0.75 means that the algorithm correctly classified the image in 75% of cases. Algorithms can be correct 75% of the time by mostly random behaviour (see the [readme](#) in the benchmark for more details on this), so performance above 0.75 is what matters.

Thus the current state of the art is:

1. Higher-than-0.9 (specifically, 0.925) performance at 10% mix rate.
2. Normalised AUC of 0.903 on the 0%-30% range.

Please help us beat these benchmarks ^\_^

More details on the dataset, the performance measure, and the benchmark can be found [here](#).

Note that this "generating possible extrapolations" task is in between a Bayesian approach and more traditional ML learning. A Bayesian approach would have a full prior and thus would "know" that emotion or text classifiers were options, before even seeing the ambiguous image. A more standard ML approach would train a single classifier and would not update on the ambiguous image either (since it's unlabeled). This approach *generates* multiple hypotheses, but only when the unlabeled image makes them relevant. ↵

Note that this is different from an algorithm noticing an image is out of distribution and querying a human so that it can label it. Here the human response provides information not only about this specific image, but about the algorithm's true loss function; this is a more efficient use of the human's time. ↵

Ultimately, the aim is to have a method that can detect this from a single image that the algorithm sees - or a hypothetical predicted image. [←](#)

# NeurIPS ML Safety Workshop 2022

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

This is a linkpost for <https://neurips2022.mlsafety.org/>

We're excited to announce the NeurIPS [ML Safety workshop](#)! To our knowledge it is the first workshop at a top ML conference to emphasize and explicitly discuss x-risks.

## X-Risk Analysis Prizes

\$100K in paper prizes will be awarded. There is \$50K for best paper awards. There is also \$50K in awards for discussing x-risk. This will be awarded to researchers who adequately explain how their work relates to AI x-risk. Analyses must engage with existing arguments for existential risks or strategies to reduce them.

## What is the topic of the workshop?

Broadly, the focus of the workshop is on [ML Safety](#), which is the umbrella term that refers to research in the following areas:

**Robustness:** designing systems to be resistant to adversaries.

**Monitoring:** detecting undesirable behavior and discovering unexpected model functionality.

This category contains interpretability and transparency research, which could be useful for understanding the goals/thought processes of advanced AI systems. It also includes anomaly detection, which has been useful for detecting proxy gaming. It also includes Trojans research, which involves identifying whether a deep neural network will suddenly change behavior if certain unknown conditions are met.

**Alignment:** building models that represent and safely optimize hard-to-specify human values.

This also includes preventing agents from pursuing unintended instrumental subgoals and designing them to be corrigible.

**Systemic Safety:** using ML to address broader governance risks related to how ML systems are handled or deployed. Examples include ML for cyberdefense, ML for improving epistemics, and [cooperative AI](#).

## How do academic workshops work?

The majority of AI research is published at conferences. These conferences support independently run workshops for research sub-areas. Researchers submit papers to workshops, and if their work is accepted, they are given the opportunity to present it to other participants. For background on the ML research community and its dynamics, see [A Bird's Eye View of the ML Field](#).

# Background

A broad overview of these research areas is in [Unsolved Problems in ML Safety](#).

For a discussion of how these problems impact x-risk, please see [Open Problems in AI X-Risk](#).

# Which values are stable under ontology shifts?

Here's a rough argument which I've been thinking about lately:

We have coherence theorems which say that, if you're not acting like you're maximizing expected utility over outcomes, you'd make payments which predictably lose you money. But in general I don't see any principled distinction between "predictably losing money" (which we see as incoherent) and "predictably spending money" (to fulfill your values): it depends on the space of outcomes over which you define utilities, which seems pretty arbitrary. You could interpret an agent being money-pumped as a type of incoherence, or as an indication that it enjoys betting and is willing to pay to do so; similarly you could interpret an agent passing up a "sure thing" bet as incoherence, or just a preference for not betting which it's willing to forgo money to satisfy. Many humans have one of these preferences!

Now, these preferences are somewhat odd ones, because you can think of every action under uncertainty as a type of bet. In other words, "betting" isn't a very fundamental category in an ontology which has a sophisticated understanding of reasoning under uncertainty. Then the obvious follow-up question is: which human values will naturally fit into [much more sophisticated ontologies](#)? I worry that not many of them will:

- In a world where minds can be easily copied, our current concepts of personal identity and personal survival will seem very strange. You could think of those values as "predictably losing money" by forgoing the benefits of temporarily running multiple copies. (This argument was inspired by [this old thought experiment](#) from Wei Dai.)
- In a world where minds can be designed with arbitrary preferences, our values related to "preference satisfaction" will seem very strange, because it'd be easy to create people with meaningless preferences that are by default satisfied to an arbitrary extent.
- In a world where we understand minds very well, our current concepts of happiness and wellbeing may seem very strange. In particular, if happiness is understood in a more sophisticated ontology as caused by [positive reward prediction error](#), then happiness is intrinsically in tension with having accurate beliefs. And if we understand reward prediction error in terms of updates to our policy, then deliberately invoking happiness would be in tension with acting effectively in the world.
  - If there's simply a tradeoff between them, we might still want to sacrifice accurate beliefs and effective action for happiness. But what I'm gesturing towards is the idea that happiness might not actually be a concept which makes much sense given a complete understanding of minds - as implied by the buddhist view of happiness as an illusion, for example.
- In a world where people can predictably influence the values of their far future descendants, and there's predictable large-scale growth, any non-zero discounting will seem very strange, because it predictably forgoes orders of magnitude more resources in the future.
  - This might result in [the strategy described by Carl Shulman](#) of utilitarian agents mimicking selfish agents by spreading out across the universe as fast as they can to get as many resources as they can, and only using

those resources to produce welfare once the returns to further expansion are very low. It does seem possible that we design AIs which spend millions or billions of years optimizing purely for resource acquisition, and then eventually use all those resources for doing something entirely different. But it seems like those AIs would need to have minds that are constructed in a very specific and complicated way to retain terminal values which are so unrelated to most of their actions.

A more general version of these arguments: human values are generalizations of learned heuristics for satisfying innate drives, which in turn are evolved proxies for maximizing genetic fitness. In theory, you can say “this originated as a heuristic/proxy, but I terminally value it”. But in practice, heuristics tend to be limited, messy concepts which don't hold up well under ontology improvement. So they're often hard to continue caring about once you deeply understand them - kinda like how it's hard to endorse “not betting” as a value once you realize that everything is a kind of bet, or endorse faith in god as a value if you no longer believe that god exists. And they're especially hard to continue caring about at scale.

Given all of this, how might future values play out? Here are four salient possibilities:

- Some core notion of happiness/conscious wellbeing/living a flourishing life is sufficiently “fundamental” that it persists even once we have a very sophisticated understanding of how minds work.
- No such intuitive notions are strongly fundamental, but we decide to ignore that fact, and optimize for values that seem incoherent to more intelligent minds. We could think of this as a way of trading away the value of consistency.
- We end up mainly valuing something like “creating as many similar minds as possible” for its own sake, as the best extrapolation of what our other values are proxies for.
- We end up mainly valuing highly complex concepts which we can't simplify very easily - like “the survival and flourishing of humanity”, as separate from the survival and flourishing of any individual human. In this world, asking whether an outcome is good for individuals might feel like asking whether human actions are good or bad for individual cells - even if we can sometimes come up with a semi-coherent answer, that's not something we care about very much.

# Principles of Privacy for Alignment Research

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

The hard/useful parts of alignment research are largely about understanding agency/intelligence/etc. That sort of understanding naturally yields capabilities-relevant insights. So, alignment researchers naturally run into decisions about how private to keep their work.

This post is a bunch of models which I use to think about that decision.

I am not very confident that my thinking on the matter is very good; in general I do not much trust my own judgment on security matters. I'd be more-than-usually interested to hear others' thoughts/critiques.

## The “Nobody Cares” Model

By default, nobody cares. Memetic reproduction rate is less than 1. Median number of citations (not counting self-citations) is approximately zero, and most citations are from people who didn't actually read the whole paper but just noticed that it's vaguely related to their own work. Median number of people who will actually go to any effort whatsoever to use your thing is zero. Getting other people to notice your work at all takes significant effort and is hard even when the work is pretty good. “Nobody cares” is a very strong default, the very large majority of the time.

Privacy, under this model, is very easy. You need to make a very large active effort in order for your research to *not* end up de-facto limited to you and maybe a few friends/coworkers.

This is the de-facto mechanism by which most theoretical work on alignment avoids advancing capabilities most of the time, and I think it is the main mechanism by which most theoretical work on alignment *should* aim to avoid advancing capabilities most of the time. It should be the default. But obviously there will be exceptions; when does the “nobody cares” model fail?

## Theory-Practice Gap and Flashy Demos

Why, as a general rule, does nobody care? In particular, why does nobody working on AI capabilities care about most work on alignment theory most of the time, given that a lot of it is capabilities-relevant?

Well, even ignoring the (large) chunk of theoretical research which turns out to be useless, the theory-practice gap is a thing. Most theoretical ideas don't really do much when you translate them to practice. This includes most ideas which sound good to intelligent people. Even those theoretical ideas which do turn out to be useful are typically quite hard to translate into practice. It takes months of work (at least), often additional significant insights, and often additional enabling pieces which aren't already mainstream or even extant. Practitioners correctly expect this, and therefore

mostly don't pay attention to most ideas until after there's evidence that they work in practice. (This is especially true of the sort of people who work on ML systems.)

In ML/AI, smart-sounding ideas which don't really work easily are especially abundant, so ML practitioners are (correctly) even more than usually likely to ignore theoretical work.

The flip side of this model is that people *will* pay lots of attention once there is clear evidence that some idea works in practice - i.e. evidence that the idea has crossed the theory-practice gap. What does that look like? Flashy demos. Flashy demos are the main signal that the theory-practice gap has already been crossed, which people correctly take to mean that the thing can be useful *now*.

The theory-practice gap is therefore a defense which both (a) slows down someone actively trying to apply an idea, and (b) makes most ideas very-low-memetic-fitness until they have a flashy demo. To a large extent, one can write freely in public without any flashy demos, and it won't spread very far memetically (or will spread very slowly if it does).

## Reputation

Aside from flashy demos, the other main factor I know of which can draw peoples' attention is reputation. If someone has a track record of interesting work, high status, or previous flashy demos, then people are more likely to pay attention to their theoretical ideas even before the theory-practice gap is crossed.

Of course this is not relevant to the large majority of people the large majority of time, especially insofar as it involves reputation *outside* of the alignment research community. That said, if you're relying on lack-of-reputation for privacy, then you need to avoid gaining too broad a following *in the future*, which may be an important constraint - more on that in the next section.

## Takeaways & Gotchas

Main takeaway of the "nobody cares" model: if you're not already a person of broad interest outside alignment, and you don't make any flashy demos, then probably approximately nobody working on ML systems outside of alignment will pay any attention to your work.

... but there are some gotchas.

First, there's a commitment/time-consistency problem: to the extent that we rely on this model of privacy, we need to precommit to remain uninteresting in the future, at least until we're confident that our earlier work won't dangerously accelerate capabilities. If you're hoping to gain lots of status outside the alignment research community, that won't play well with a "nobody cares" privacy model. If you're hoping to show future flashy demos, that won't play well with a "nobody cares" privacy model. If your future work is very visibly interesting, you may be stuck keeping it secret.

(Though note that, in the vast majority of cases, it will turn out that your earlier theory work was never particularly important for capabilities in the first place, and hopefully you figure that out later. So relying on "nobody caring" now will reduce your later

options mainly in worlds where your current work turns out to be unusually important/interesting in its own right.)

Second, relying on “nobody caring” obviously does not yield much defense-in-depth. It’s probably not something we want to rely on for stuff that immediately or directly advances capabilities by a lot.

But for most theoretical alignment work most of the time, where there are some capabilities implications but they’re not very direct or immediately dangerous on their own, I think “nobody cares” is the right privacy model under which to operate. Mostly, theoretical researchers should just not worry much about privacy, as long as (1) they don’t publish flashy demos, (2) they don’t have much name recognition outside alignment, and (3) the things they’re working on won’t immediately or directly advance capabilities by a lot.

## Beyond “Nobody Cares”: Danger, Secrecy and Adversaries

Broadly speaking, I see two main categories of reasons for theoretical researchers to go beyond the “nobody cares” model and start to actually think about privacy:

- Research which might directly or immediately advance capabilities significantly
- Current or anticipated future work which is unusually likely to draw a lot of attention, especially outside the alignment field

These are different failure modes of the “nobody cares” model, and they call for different responses.

### The “Keep It To Yourself” Model for Immediately Capabilities-Relevant Research

Under the “nobody cares” model, a small number of people might occasionally pay attention to your research and try to use it, but your research is not memetically fit enough to spread much. For research which might directly or immediately advance capabilities significantly, even a handful of people trying it out is potentially problematic. Those handful might realize there’s a big capability gain and run off to produce a flashy demo.

For research which is directly or immediately capabilities-relevant, we want zero people to publicly try it. The “nobody cares” model is not good enough to robustly achieve that. In these cases, my general policy would be to not publish the research, and possibly not share it with anyone else at all (depending on just how immediately and directly capabilities-relevant it looks).

On the other hand, we don’t necessarily need to be super paranoid about it. In this model, we’re still mostly worried about the research contributing *marginally* to capabilities; we don’t expect it to immediately produce a full-blown strong AGI. We want to avoid the work spreading publicly, but it’s still not that big a problem if e.g. some government surveillance sees my google docs. Spy agencies, after all, would presumably not publicly share my secrets after stealing them.

## The “Active Adversary” Model

... which brings us to the really paranoid end of the spectrum. Under this model, we want to be secure even against active adversaries trying to gain access to our research - e.g. government spy agencies.

I’m not going to give advice about how to achieve this level of security, because I don’t think I’m very good at this kind of paranoia. The main question I’ll focus on is: when do we need highly paranoid levels of security, and when can we get away with less?

As with the other models, someone has to pay attention in order for security to be necessary at all. Even if a government spy agency had a world-class ML research lab (which I doubt is currently the case), they’d presumably ignore most research for the same reasons other ML researchers do. Also, spying is presumably expensive; random theorists/scientists are presumably not worth the cost of having a human examine their work. The sorts of things which I’d expect to draw attention are the same as earlier:

- enough of a track record that someone might actually go to the trouble of spying on our work
- public demonstration of impressive capabilities, or use of impressive capabilities in a way which will likely be noticed (e.g. stock trading)

Even if we are worried about attention from spies, that still doesn’t mean that most of our work needs high levels of paranoia. The sort of groups who are likely to steal information not meant to be public are not themselves very likely to make that information public. (Well, assuming our dry technical research doesn’t draw the attention of the dreaded Investigative Journalists.) So unless we’re worried that our research will accelerate capabilities to such a dramatic extent that it would enable some government agency to develop dangerous AGI themselves, we probably don’t need to worry about the spies.

The case where we need extreme paranoia is where *both* (1) an adversary is plausibly likely to pay attention, *and* (2) our research might allow for immediate and direct and very large capability gains, without any significant theory-practice gap.

This degree of secrecy should hopefully not be needed very often.

## Other Considerations

### Unilateralist’s Curse

Many people may have the same idea, and it only takes one of them to share it. If all their estimates of the riskiness of the idea have some noise in them, and their risk tolerance has some noise, then presumably it will be the person with unusually low risk estimate and unusually high risk tolerance who determines whether the idea is shared.

In general, [this sort of thing](#) creates a bias toward unilateral actions being taken even when most people want them to not be taken.

On the other hand, unilateralist's curse is only relevant to an idea *which many people have*. And if many people have the idea already, then it's probably not something which can realistically stay secret for very long anyway.

## Existing Ideas

In general, if an idea has been talked-about in public at some previous point, then it's probably fine to talk about again. Your marginal impact on memetic fitness is unlikely to be very large, and if the idea hasn't already taken off then that's strong evidence that it isn't *too* memetically fit. (Though this does not apply if you are a person with a very large following.)

## Alignment Researchers as the Threat

Just because someone's part of the ingroup does not mean that they won't push the run button. We don't have a way to distinguish safe from dangerous programs; our ingroup is not meaningfully more able to do so than the outgroup, and few people in the ingroup are very careful about running python scripts on a day-to-day basis. (I'm certainly not!)

Point is: don't just assume that it's fine to share ideas with everyone in the ingroup.

On the other hand, if all we want is for an idea to not spread publicly, then in-group trust is less risky, because group members would burn their reputation by sharing private things.

## Differential Alignment/Capabilities Advantage

In the large majority of cases, research is obviously much more relevant to one or the other, and desired privacy levels should be chosen based on that.

I don't think it's very productive, in practice, to play the "but it *could* be relevant to [alignment/capabilities] via [XYZ]" game for things which seem obviously more relevant to capabilities/alignment.

## Most Secrecy Is Hopefully Temporary

Most ideas will not dramatically impact capabilities. Usually, we should expect secrecy to be temporary, long enough to check whether a potentially-capabilities-relevant idea is *actually* short-term relevant (i.e. test it on some limited use-case).

## Feedback Please!

Part of the reason I'm posting this is because I have not seen discussion of the topic which feels adequate. I don't think my own thoughts are clearly correct. So, please argue about it!

# Opening Session Tips & Advice

## Meta

CFAR ran many, many workshops.

After each workshop, there would be feedback from the participants, and debrief discussions among the staff. We would talk about what had worked and what hadn't, what we wish had been said or done, what we would try differently in the future, etc.

Often, what resulted was a new addition to the opening session. Opening session, at a CFAR workshop, was largely about expectation setting, and getting everyone on the same page—making sure everyone knew what they were getting into, and what was going to be asked of them, and why.

The "tips and advice" section of opening session was often framed as "things past participants said, at the end, that they wished they'd been told at the beginning."

(This was often but not always literally true.)

Little snippets of wisdom about *how* to engage with the content, what to watch out for in one's own experience, where to put one's attention, etc. Often the staff would create their own tips and advice based off of watching classes fail, or watching individual participants "bounce" off the workshop, and trying to figure out why.

There were something like two dozen distinct tips, at various points, of which four or five would be presented at a given workshop. Some were added, some were removed, others morphed or mutated, yet others got more deeply baked into the structure of the workshop and were no longer needed in opening session.

Below is a selection of some of the most important and longest-lasting opening session tips. They are presented here for two purposes:

1. Despite the fact that this is an online sequence and not a workshop, the tips nevertheless contain valuable wisdom about how to engage with the content, and some specific ways that trying to do so tends to go wrong for people.
2. They may be useful advice in other contexts, such as conferences or events that you yourself may run, in the future.

---

## Be Present

One key element of getting the most out of an experience is being *present*. This includes physically showing up, but it also includes having your mind in the room and your background thoughts focused on the content. The more you're taking calls and answering texts and keeping up with social media and what's going on back home, the more you'll remain in your ordinary mental space, continuing to reinforce the same habits and patterns you're here to change. There's a sort of snowball effect, where even a little disengagement can make absorbing the value you'd like from a workshop rather difficult, which confirms a suspicion that there's no value to be had, and so on.

Think about, for instance, the sorts of thoughts one can have on a long, three-day hiking trip, with no deadlines or obligations. When all of your thoughts must be *purposeful*, or when

every thought must resolve itself before the next thing on the schedule rolls around, there are a lot of thoughts you simply *can't have*.

And it is precisely thoughts-unlike-those-you're-accustomed-to-having that the workshop is trying to provide! After all, if your present ways of thinking and being were sufficient to solve all your problems and achieve all your goals, you'd already be done. Not every change is an improvement, but every improvement must necessarily be a change, and one of the *precursors* to change is setting yourself up to be able to be in any kind of non-default state of mind at all. If it's business as usual, your brain will *produce* business-as-usual thoughts, and you'll find few or no life-changing insights in that drawer.

In addition to external distraction, we've also found that there are a few unhelpful narratives that participants occasionally find themselves repeating—narratives which make it hard to engage with the content and block opportunities for asking good questions and taking new steps. If you notice one of these narratives cropping up in the back of your mind, we encourage you to try deliberately setting it aside, as an experiment—let it go, see what happens, and judge for yourself. Our staff are happy to chat with you about any of these, if you think you might find that helpful.

- “I’m too dumb/old/lazy to learn this.” We sometimes encounter people who think that, because they don’t measure up to some standard or another, they aren’t “good enough” to benefit from the workshop material. As a counter to this, we recommend donning a growth mind-set: if it can be learned by a human, it can be learned by you.
- “I already know this part.” Some people come into our workshop with significant background knowledge and, when they start to see familiar material, slip into a mode of assuming there’s nothing for them to learn. Unfortunately, this can mean that you’re “turning off” right at the moment that we’re offering new insight. To counter this, we recommend that you try to approach every class with fresh eyes. Even if the core concepts are familiar, look for the fine detail—the places where your peers and instructors have made valuable connections you might have missed. In particular, try to be **interested** rather than **interesting**—there’s more to gain from stealing new insights than re-hashing thoughts you’ve already thought.
- “I’ve got important things to do, and this lesson can wait.” Sometimes there really are important things to attend to. But if they’re on your mind during the workshop, you’re likely to have a hard time absorbing the material in a way that will stick. We recommend that you set aside what you can, and fully address what you can’t set aside: if something really can’t wait, step out, make it your sole focus until it’s dealt with, and return with full and fresh attention.

---

## Let your wants come alive

Imagine being a vegan, or strictly kosher, or someone with restrictive food allergies. Let's say it's Friday or Saturday night, and your circle of friends has invited you out to dinner, a movie, and drinks.

It's easy to see that it might be sort of *dangerous* for you to look forward to the meal with genuine anticipation and optimism—the group ends up at a burger place, and you open the menu, and as you flip through you find that the only vegan option is lettuce-covered lettuce with lettuce on the side, just like the last twelve times you went out.

And so, in that situation, it's easy to imagine a strategy of keeping your wants *asleep*. Sort of pre-emptively tamping down on any kind of hope or hunger, telling yourself “it’s just about hanging out with my friends. I’ll cook my own food before I go, or when I get back. I’m just going out to be social and have fun.”

This coping mechanism makes perfect sense! It's there to prevent a very real and unpleasant experience. It's *protecting* you from preventable sadness.

But there's a particular way in which it leaves you sort of hollow and crippled. There's something good and magical that can happen, if you instead let your wants come alive. If you choose to prioritize yourself and your values, if you dare to expect that good and interesting opportunities might crop up.

It is indeed a lot worse, if you let yourself build up hope and then have those hopes dashed. But there's a certain point of view from which *nothing good can even happen*, if you don't expose yourself to that risk at least *sometimes*.

So our recommendation for the workshop is this: let your wants come alive. Let yourself hunger for things, let yourself get excited for things, let yourself be sort of pushy and sort of selfish and sort of willing to visualize a warm and glowy future, even if there's a risk that future won't come to pass. If there was ever a time to take on that risk, it's these next four and a half days.

---

## Try Things!

When you're considering adopting new habits or ideas, there's no better way to gather data than *actually trying*. It's often faster and simpler to just give things a shot and see how it goes than to spend a lot of time trying to anticipate and predict whether or not you'll find something worthwhile.

(And it helps you avoid the failure mode of "putting things on the list" and then never getting to them! Getting that first try out of the way goes a long way toward making a second one actually happen.)

This is particularly important because when something *does* work out, you get to *keep doing it!* If your friends have recommended five different activities to you, and you've only liked one of them, it's easy to think of the whole process as a pretty big waste of time:

- ✗ Yoga
- ✗ Ultimate Frisbee
- ✗ Dungeons & Dragons
- ✓ Meditation
- ✗ Salsa dancing

An 80% failure rate isn't exactly encouraging, after all. But what the above framing fails to take into account is the magnitude of even a single success. Instead of four bad experiences and one good one, what's actually going on is more like the following:

Activity	T1	T2	T3	T4	T5	T6	T7	T8	T9
Yoga	✗	✗							
Ultimate Frisbee	✗								
Dungeons & Dragons	✗	✗	✗						
Meditation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Salsa dancing	✗								

When you look at it this way, you can see that the failed trials are more than compensated for by the sustained run of a now-successful habit. Indeed, when it comes to hobbies and activities that might last you the rest of your life, it becomes worthwhile to establish a habit of trying things that have even a one-in-ten or one-in-a-hundred chance of being enjoyable. It only takes a few paying off to make the whole thing worthwhile.

So while you're listening and participating this weekend, be on the lookout for opportunities to turn our lessons into actions that you can actually try out, right then and there. Translating class material into practical experiments is a great way to digest material anyway, and it'll help you decide which techniques are most worth prioritizing when you return home.

---

## Make good quiche

Imagine that you have a friend who is creating a recipe book. You've agreed to help your friend beta test some of their recipes, and they've handed you a rough draft of instructions on how to make quiche.

As you're reading through the recipe, you begin to notice a few ... let's say, *problematic* steps. For instance, the recipe calls for six "whole eggs," which to you seems to imply shells and all. It also says to bake for 4.5 hours at 450 degrees, and calls for 10 tablespoons of salt.

Now, one way that you might offer productive feedback to your friend is to follow the recipe *exactly as written*, creating a crunchy, salty, burned quiche. This is actually a pretty helpful strategy, early on—it's a way to stress-test the recipe to see exactly how broken it is.

However, if you also *happen to want some quiche*, there's another method you might employ. Instead of following steps that are obviously wrong, you could instead *try to make good quiche*, treating the recipe as more of an inspiration than a strict set of instructions. You could throw away the eggshells, drop the time and temperature down to (say) 45 minutes at 350 degrees, and throw in just a pinch of salt. Maybe you'll even have some additions that your friend didn't think of, like mixing in some chopped kale.

At the end of *that* process, you'll not only have notes about flaws in the original recipe, but also constructive suggestions and—most importantly—a delicious meal you can actually eat. You'll have something that's useful to *you*, both in the moment and for the future.

Like your friend's quiche recipe, many of the concepts and techniques within the workshop are experimental. There will be times when they seem a little off, and other times when they may seem clearly false. It helps to remember that the goal is not to improve our recipe book, but to make good quiche. That means that, instead of doing things that don't make sense, you should feel free to tinker, experiment, and modify. Your perspective is unique—while we have a lot of insight to offer, there's no one who better understands your own life and mind than you. If we seem to be pointing in the wrong direction, feel free to head in the right one, instead—and afterward, let us know what you discovered.

---

## Adjust your seat

(An iteration on "make good quiche".)

In the late 1940s, the U.S. Air Force had a serious problem. Planes were crashing left and right—not because they'd been shot down, but because the pilots were simply losing control at an astonishing rate. On the worst day, there were seventeen crashes.

It turned out that the reason for this had to do with a decision that had been made back in 1926, when the military first set out to design the cockpit. At the time, they'd taken a few hundred pilots and used their measurements to standardize things like the size of the seat, and the distance to the pedals. The modern-day pilots weren't comfortable in these cockpits, and in the fast-paced, high-stakes environment of early airflight, a slight inability to reach the pedals or see out of your windshield could mean the difference between a successful mission and a lethal crash.

At first, the hypothesis was that pilots had changed in size. To investigate, the Air Force launched another study, measuring roughly four thousand pilots on over a hundred different dimensions, all the way down to thumb length and the distance from a pilot's eye to his ear. But when they calculated the averages, they found that nothing had meaningfully changed.

Enter Lieutenant Gilbert S. Daniels. He approached the problem with a new question: How many of those pilots are *actually* average?

The answer? Zero. Not a *single* pilot was within fifteen percentage points of the average on all ten of the most relevant measurements—which meant that the cockpits were designed to fit people who didn’t exist.

This revelation led to all of the technology that you’ll find in modern cars today. Adjustable seats, mirrors, and steering wheels—all of that and more was developed so that pilots would stop dying in preventable accidents.

Which leads us to our advice for the workshop—adjust your seat. The techniques that we’re going to present to you are central, average versions—they’re the *least wrong* for the *most people*. But that also means that none of them will work exactly right for anybody. Use them as a starting point, but before you try to take off and fly, tinker with the settings—change the lean, and the height, and how far forward or back they are; adjust the headrest and maybe fiddle with the mirrors, too. Our version is good, but there’s a much better version that only you will be able to find.

---

## Eat the instructions

(An attempt to synthesize “try things” and “adjust your seat,” which are contradictory.)

Much of the fun of playing with construction toys like LEGO or K’nex or erector sets is building your own unique, novel designs.

But usually LEGOs come in a box with instructions on how to build a particular spaceship or castle or train set or whatever.

It might seem like there are “two kinds of LEGO kids”—those who build according to the instructions, and those who don’t.

But just as you’re missing something if you’re only “following your heart” or only “following your head,” there’s a better strategy that combines the benefits of both.

If you build according to the instructions *first*, you will often learn some tiny neat trick of engineering that the LEGO designers discovered or invented and which you would be unlikely to stumble across yourself. After all, they put thousands and thousands of hours into figuring out how to stick LEGOs together.

And then, once you’ve built the thing and learned from the experience, if you want to take it apart and make your *own* spaceship, you’ll be much better equipped to do so, now that you have the latest cutting edge tactics and techniques. You will be a more flexible and competent designer, better able to make the LEGO pieces come together in the way you want.

Similarly, we recommend that you engage with CFAR’s techniques both by actually trying them out, as written, *and* by modifying them/throwing them out and inventing your own. We recommend a synthesis of “try things” and “adjust your seat” which we call eating the instructions—try, *then* tinker.

---

## The tacit and the explicit

There are many useful ways to divide up and categorize human knowledge, or human thinking, or human psychology. You can think in terms of id, ego, and superego, or system 1 and system 2, or big-five personality types, or wilder and sillier things like Hogwarts houses or the Magic: the Gathering color wheel. Each of these is an oversimplification that misses some things, but that can help you draw out insight about others.

One way that CFAR likes to think about the human mind is to look at the distinction between *tacit knowledge* and *explicit knowledge*.

Tacit knowledge is like the knowledge that you use to ride a bicycle—it's complex, experiential, intuitive, hard to put into words. You could sort of try to describe what you're doing to a bright five-year-old, but even if you successfully convey a couple of tips, it won't be those tips themselves that help so much as the new bit of tacit knowledge that the five-year-old invents in their own head as a result of thinking about the tip.

Explicit knowledge, on the other hand, is clear and concrete and transferrable and (at least somewhat) objectively verifiable. *How* you ride a bicycle is tacit, but the fact that you *can* ride a bicycle is explicit. It's a binary fact that can be completely and compactly transferred through words, and that is checkable through experiment.

Explicit knowledge is held in high regard, because it's how we prove things in mathematics and how we make scientific progress on vaccines and space shuttles and microprocessors and how we transfer lore and culture to our children and so on and so forth. It's a huge part of how the human race has made it this far.

But tacit knowledge is often forgotten, or pooh-poohed in a way that CFAR thinks is going a little too far. Just because verifiable and transferrable knowledge is powerful and valuable doesn't mean that things which are hard to verify and hard to transfer are *not* powerful and valuable. Explicit scientific knowledge is the key to a lot of our progress, but we wouldn't have been able to accrue those scientific insights if it weren't for people's ability to generate hypotheses—and skill at generating hypotheses is absolutely tacit.

We don't know how to teach people to consistently produce insightful and paradigm-defining hypotheses any more than we know exactly how to transfer skill at poetry, or the ability to be an outstanding coach, or the intuition of a veteran math researcher who knows instinctively which threads are promising and worth following (and is usually right about this, though they can't explain where the intuition comes from or what it's made of).

A lot of what we'll be doing this weekend is moving back and forth between the explicit and the tacit—practicing techniques to draw out some of our tacit insights into the explicit, where we can reason about them, or trying to build up the skill of switching between (or combining!) both tacit and explicit insights as we think about thinking or try to improve our lives or ourselves. This will only work if we recognize the true fact that *both* kinds of thinking indeed have value, and that each contains insight that the other lacks, and so our advice to you is to treat all of your thinking with some degree of respect, and not to be the sort of person who only "trusts their gut" or only "thinks things through" and doesn't have room in their toolkit for both.

---

## Build Form

*Form* is the quality such that additional effort translates directly to greater results.

What we mean by that is that none of your additional effort is leaking out, or creating friction, or pushing in the wrong direction, or simply going to waste. It means that if you're a runner, your knees don't wobble and your arms pump correctly. If you're designing an airplane, you don't leave random bits sticking out, where they'll catch the wind. If you're a

writer, you're using as few words as possible, and if you're a programmer, you don't have extraneous function calls that burn up computational resources.

One of the most important things to encourage in the early stages of a new skill is the development of good form. Once you have it, trying harder works, whereas if you don't have it, trying harder often just leads to a lot of frustration and discouragement. And of course, if you have bad habits right from the start, they're only going to get harder and harder to fix as you ingrain them through practice.

Many of the CFAR techniques you will encounter are subtle, despite their veneer of straightforwardness. Correct form is hard to come by, especially since each individual is different, and what works for one person may not be any good for another.

For that reason, we often spend a lot of time during the workshop talking about small, mundane problems with relatively few moving parts. That isn't because this is all the techniques are good for, but because, at the start, we want you to be able to focus on building form. It's like a weightlifter practicing with an empty bar before adding on the pounds—we encourage you to practice on simple things first, and then ramp up.

Another way to think of this is that your problems will tend to either be *adaptive* or *technical*. Adaptive problems require experimentation, novel strategies, or new ways of thinking and being; they're problems containing "unknown unknowns" and are often opaque in addition to being difficult. Technical problems may be equally difficult, but their difficulty lies in execution—technical problems are those where the path to the solution is known or knowable and does not need to be discovered.

It's likely that you're here because you have some interesting adaptive challenges in your life, and you're itching to get some new tools to work on them. Don't be disappointed if most of the techniques are presented with technical examples, or if your early practice is with technical problems. We're warming you up for the big stuff, and we'll absolutely get to it. We just want you to have the right muscle memory, and some practice under your belt, before we do.

---

## Boggle!

There's a way in which education tends to make knowledge very *flat*.

Let's take the Earth and the Sun, for example. If I were to ask you about the relationship between the two, you'd probably offer me the well-worn phrase "the Earth revolves around the Sun."

It's automatic, reflexive, almost atomic—once you start with "the Earth," you barely have to think anymore. The "revolves around the Sun" part just fills itself in.

But once upon a time, people *didn't know* that the Earth revolved around the Sun. In fact, people didn't even really know what the Earth and the Sun *were*—they thought they did, but looks can be deceiving. It took us multiple geniuses and the innovations of centuries to go from "the Earth is a flat plane and the Sun travels across the celestial sphere" to the factoid that we repeat back to our teachers in a bored monotone. Somehow, all of the confusion and excitement of discovering that the Sun is an incandescent ball of hydrogen and that the Earth is tied to it by the same fundamental force which makes pendulums swing and that both of them are round except not quite and that gravitational attraction is proportional to the square of the distance except not quite, don't forget relativity and quantum mechanics and—

—somehow, all of that gets lost when we flatten things out into “the Earth revolves around the Sun.”

Fortunately, there’s a solution—*boggling*. You’re reading an essay! What’s an essay? I mean, okay, it’s just a essay. But what is it really? I mean, where did these words come from? Who wrote them? Yeah, “Duncan Sabien,” but who’s that? And the words in front of you right now aren’t the *same* words that he wrote—are they? Sort of. What’s up with identity when it comes to concepts, anyway? Not to mention the literal images themselves! Pixels on a screen! How’d they get there? How do they know where to go? Who built the device they’re displayed on? How does it work? What’s actually going on in your brain, when you look at these squiggles and find yourself thinking thoughts? What even *is* a thought? I hear there are neurons involved—how does that work?

When you allow yourself to embrace confusion, and turn away from the cached, easy, empty answers, you start to see a much richer, deeper world, with many more opportunities to learn and to grow. During the workshop, there will be many things that seem like stuff that you already know, just as you already know that the Earth revolves around the Sun. But don’t be fooled! Surface explanations are the opposite of knowledge—they’re a curiosity-killer, preventing you from noticing that there’s stuff you still don’t get. Human cognition is one of the most complex, opaque, and difficult phenomena we’ve ever encountered. As you study it, don’t settle for flat knowledge—instead, boggle.

# What's next for instrumental rationality?

Preceded by: [Curating "The Epistemic Sequences" \(list v.0.1\)](#).

*Epistemic status: speculations and ideations from me about the potential for further progress on broadly accessible instrumental rationality content.*

In [Curating "The Epistemic Sequences"](#), I explained how the epistemic content of the LessWrong sequences has a different epistemic status than the instrumental content, i.e., content on how to behave. So what's next for instrumental rationality? It would be great if there were a how-to-behave version of the sequences that was built on foundations as strong as logic, probability, statistics, and causal inference.

Unfortunately, those foundations don't yet exist. There aren't formal foundations for decision theory, game theory, ethics, meta-ethics, and political theory that are "tried and true" the way logic, probability, statistics, and causal inference ("L+P+S+C") are.

Many people argue that universally-useful how-to-behave instructions *can't* exist, on the grounds that "philosophers have been trying to solve these issues for millennia", but I'm not sure that's a strong case. After all, philosophers had been trying to develop truth-seeking techniques for millennia prior to the 20th century, and then along came a bunch of progress in L+P+S+C with widespread applications, enabling what might be called an "epistemic enlightenment (for individuals)", which arguably culminated in the epistemic content of the LessWrong sequences. And, perhaps in the next decade, there could be progress in the theory of [embedded agency](#) and multi-agent rationality ("E+M") leading to real-world applications as robustly useful and well-vetted as L+P+S+C are today. If breakthroughs in embedded & multi-agent rationality then remained in practice for something like 30 years of applications in broad-sweeping domains (or, an AI-augmented equivalent of 30 human-civilization-years) the way L+P+S+C have, perhaps then will be a good time for someone to write the "The Instrumental Sequences", and a new generation of instrumentally enlightened people will look back and wonder why it was considered impossible to derive a principled account of how individuals should behave.

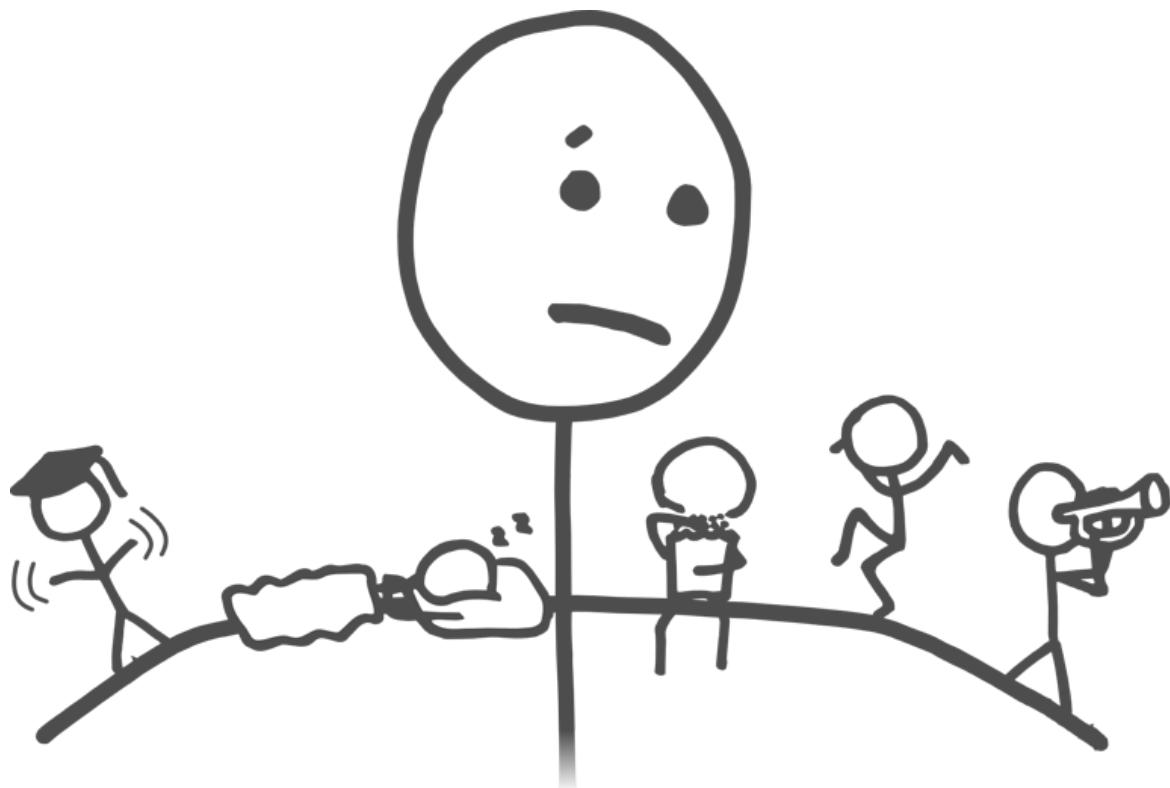
# Internal Double Crux

*Author's note: While most CFAR workshops taught double crux first and then internal double crux afterward, it's long been my opinion that both concepts would be better served in the other order. Recent experimentation has shown results consistent with this hypothesis; practicing collaborative truth-seeking inside one's own head helps one build the necessary mental muscles before moving on to doing it with a whole other separate human. [External] Double crux will be the next post in the sequence, followed by a loop back around to some of the opening session advice.*

---

**Epistemic status:** Preliminary/tentative

*Internal double crux (formerly propagating urges) is a technique-in-progress, with the goal of finding motivation through truth-seeking rather than through coercion or self-deception. It is currently in flux and has no formal research backing, but it follows logically from a handful of other threads about which CFAR is relatively confident (such as microhedonics, hyperbolic discounting, cognitive behavioral therapy, and useful-even-if-wrong theories like internal family systems or society of mind).*



*Why couldn't I just get an angel  
and a demon like everybody else?*

---

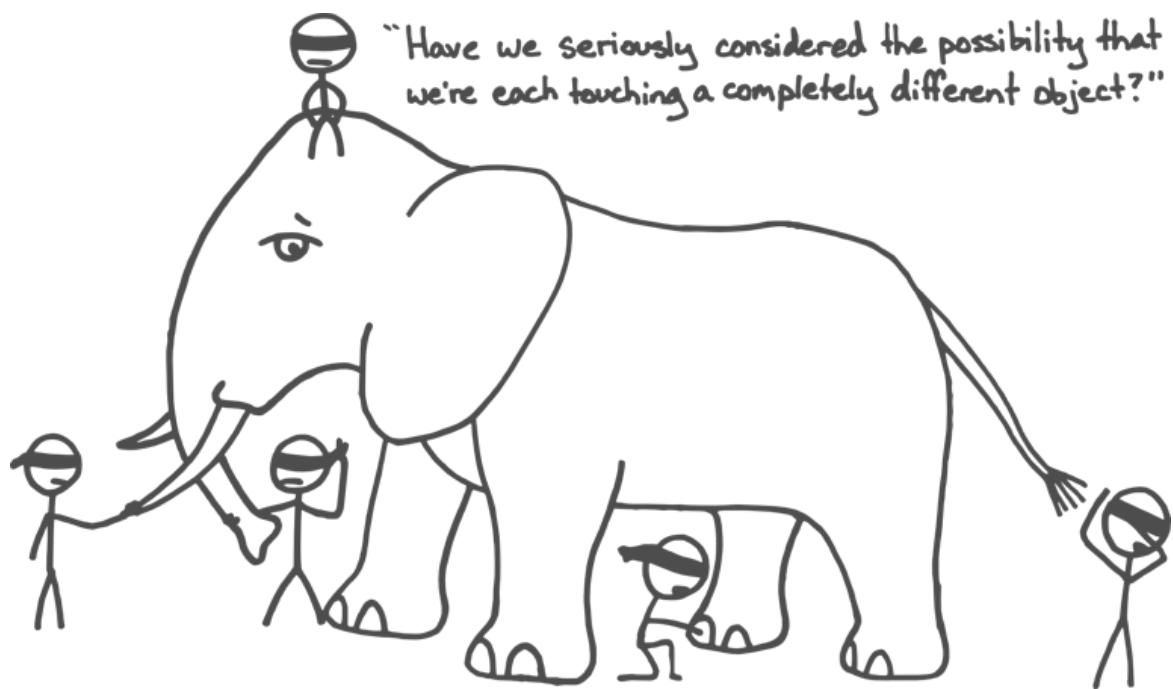
If two different people have access to the same information, and their models of the world cause them to make two different predictions, then we can confidently say that *at least* one of them is incorrect. We may not always be able to tell which is which, but we can be sure that one-if-not-both of them has a chance to update toward the truth.

Similarly, if a single person has *simultaneously contradictory beliefs or desires*, then at least one of the models behind those beliefs is wrong, miscalibrated, or incomplete.

(And usually both.)

If you both “want to get good at running” and also never want to get up off the couch and put on your running shoes, then one part of your belief set—one of your causal models of the universe—has concluded that running will help achieve your goals, and another has concluded that it doesn’t, and both of these can’t be true.

Internal double crux is a technique that seeks to resolve this conflict by helping each of these models to *incorporate* the information that the other has to offer. If you were to conceive of yourself as being made up of sub-agents, each of whom focuses on a different subset of your goals and has a different perspective on how the world works, then the goal is to cause those sub-agents to enter into a productive double crux conversation and correct their tunnel vision.



In particular, the hope is to have both sides of your internal disagreement update toward *truth*. The conflict is a result of some kind of confusion, and thus *reducing* confusion will also tend to reduce conflict.

CFAR's experience running internal double crux with participants is that the end result of such a process is a state of feeling (more) intrinsically motivated and internally unconflicted; of reducing one's need for duty or diligence or force-of-will and instead having one's urges aligned with one's actual goals. It's an early version of a technique for turning *wanting to want* into straightforward wanting.

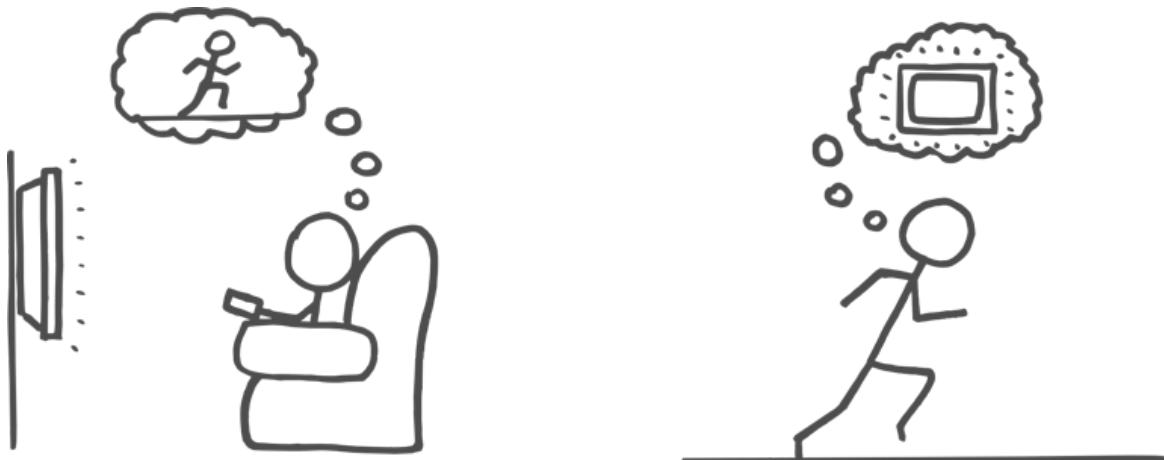
## Understanding “shoulds”

Part of the problem that internal double crux seeks to correct is our natural tendency to arbitrarily support some sub-agents or subgoals while suppressing others. Many people find it easy, for instance, to attach words like “motivated” or “goal-oriented” or “good” to the part of themselves that wants to go running or finish the project, while attaching words like “stupid” or “lazy” or “undisciplined” to the part that wants to stay on the couch.

This is an effective shortcut for some people, but it comes at a cost—you’re ignoring signals from part of your belief set, and expending energy on internal conflict and executive overrides that could otherwise be allotted to the things you actually want to do. Instead of containing, suppressing, or drowning out your conflicting urges, IDC encourages you to update and integrate them, or at the very least to give them an actual, impartial hearing before deciding that they’re inappropriate.

To return to the running example: you may have a belief that it’s good to exercise, and furthermore that *running* is the best and most efficient way to exercise, and furthermore that doing so is a better way to spend your afternoon than, say, Netflix.

If you happen to be *watching* Netflix at the time, this belief is likely to ruin your fun. Perhaps you get up, put on your shoes, and begin to run—yet as soon as you do, you find yourself longing to stop, and continue only with effort and some minor degree of suffering.



Rather than summarizing this situation as “I’m just lazy” or “I struggle to stay motivated,” it’s instead productive to think “in addition to my belief that it’s good to run, I apparently *also* have a belief that it’s good to watch Netflix.” This isn’t just a cute, permissive reframe; it’s what’s *actually going on*. Some part of you believes that Netflix is exactly the Thing To Be Doing.

And this part of you believes this for some causal reason. Beliefs don’t come from nowhere; they’re essentially always a response to some kind of past experience. The part of you that is generating pressure-toward-Netflix is doing so *because* it thinks that staying on the couch will make for a better life, and bring you closer to your goals. It’s not lazy or stupid, it’s *tunnel visioned*, failing to take into account things like long-term health, or the value of following through on your self-commitments.

(Just as the part of you that’s clamoring to get off the couch *also* has tunnel vision, and is discounting the value of relaxation or hedonism.)

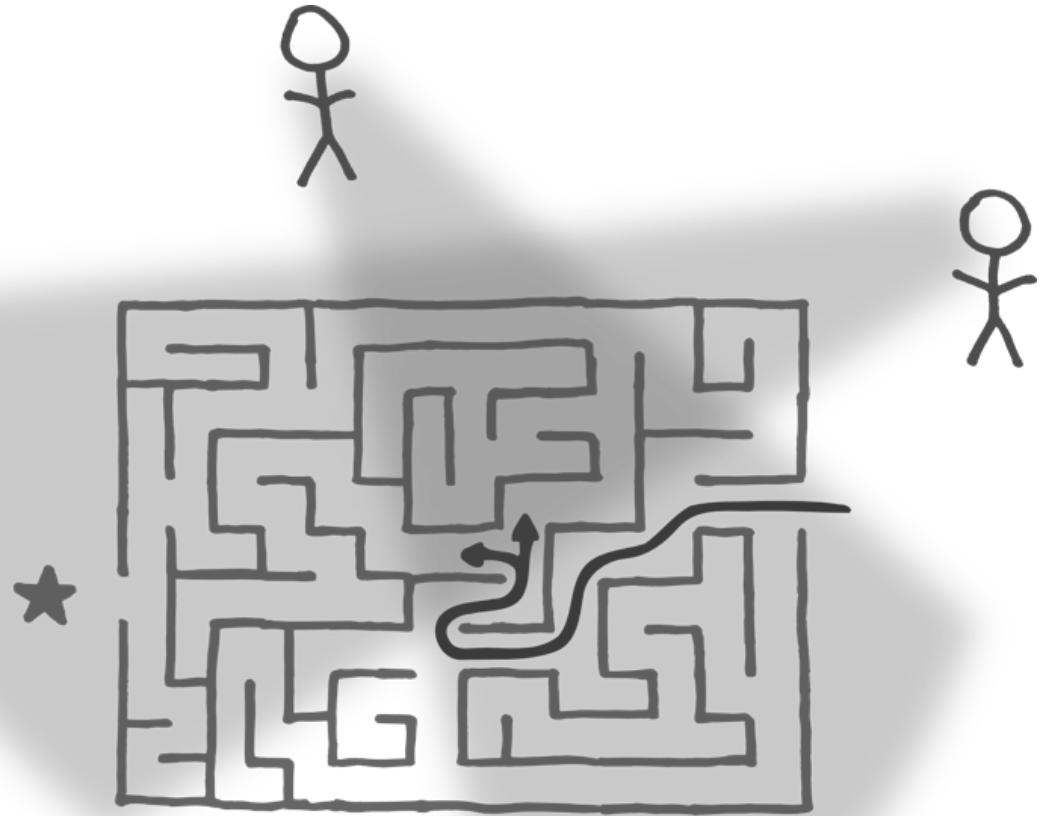
At CFAR, we often characterize these internal disagreements as “shoulds.” Given any default action, a **should** is an urge or a pressure to do something else instead:

- You’ve decided to be more gentle with your criticism and to experiment with using “I” statements, but you can’t shake the feeling that what your colleague is doing is objectively wrong and inexcusable and that there’s no point in beating around the bush.
- You’re working on the seventh chapter (of thirty) of your book, and even though you know these scenes are important for setting up later action, you find yourself wanting to do almost anything else.
- You’ve talked for years about wanting to learn [piano/Mandarin/swing dancing/Haskell/knitting/motorcycle maintenance], but even though there are classes at the local community center and all your friends are going, you’re oddly reluctant and keep making excuses.

If, on the other hand, you went ahead and grumped at your coworker, you might feel that you *should* have stuck to your communication goals; if you buckled down for a writing sprint, you might feel that you *should* have taken time out to spend with your significant other, or conserved resources for work the next day; if you start taking classes, you might feel that you *should* have saved the money, or spent it on something else instead.

Many people default to one side or the other when they notice a should—they have a deontological policy of defending their inner emotional selves, or of conforming to social expectations, or of sticking to the plan, or of being flexible and changing the plan. The problem is, any one-size-fits-all solution is going to miss a large percentage of the time, and writing the bottom line without actually considering the arguments is a recipe for inaccurate beliefs.

At their core, shoulds are *data*, and data is something an aspiring rationalist almost always wants more of. Just as regular double crux encourages us to remain open to the idea that others might have better information than we do, so too does internal double crux encourage us to listen to the input of every aspect of our motivational structure. Different parts of your psyche are better equipped to pay attention to different swaths of the available evidence, and they process that evidence in different ways. Given the complexity of the world, it makes sense to start from the assumption that a synthesis of conclusions will be more accurate than any one conclusion on its own.



The agent at the top mistakenly believes that the correct move is to head to the left, since that seems to be the most direct path toward the goal. The agent on the right can see that this is a mistake, but it would never have been able to navigate to that particular node of the maze on its own.

The part of you that wants to run is good at paying attention to your long-term goals, your social standing, your health, and your sense of yourself as a strong and capable person. The part of you that wants to watch Netflix is good at paying attention to your short-term urges, your energy levels, your sense of comfort, and whether or not the new *Stranger Things* episode seems likely to be good.

You can ignore one side or the other indefinitely, but the result is often feeling halfhearted or torn, ruminating or struggling with decisions, burning willpower, suffering from your decisions, and endorsing one part of your psyche beating up on another part. In order to build a maximally detailed understanding of the world and correctly strategize across all of your needs and goals, you've got to bring *all* of your models to the table—implicit, explicit, S1, S2, endorsed, embarrassing, vague, and exact.

---

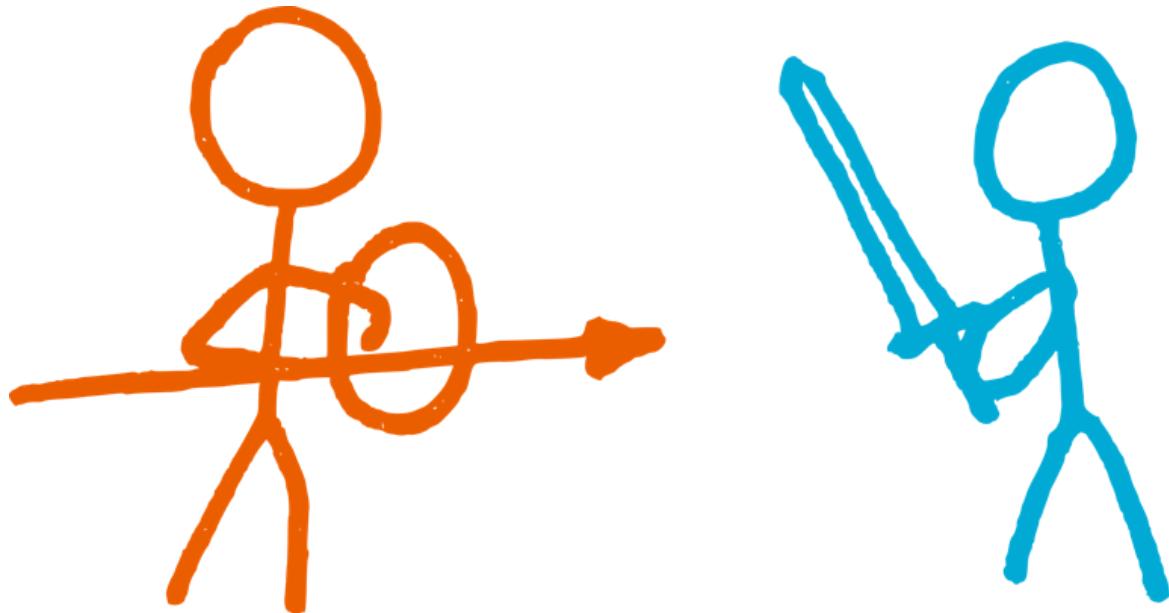
## An Example IDC

Moreso than with most techniques, we have found that participants learning IDC get substantial value out of holding themselves to a very specific format for at least their first

couple of attempts. There's some amount of magic that often requires experience and is less transmissible up front, so simply giving it a shot as-written (as opposed to making your own tweaks and adjustments on the fly) is something we recommend more strongly than usual.

### **Step 0: Find an internal disagreement.**

Look for any sort of "should" that's counter to your current default action—something you feel you aren't supposed to think or believe (though on some level you do), or a step toward your goal that *feels* useless or excessively unpleasant.



**Step 1: Take a piece of paper, and, at the top, draw two dots, representing the two perspectives/viewpoints/sub-agents. Name them.**

**Go running**

**Laze around**

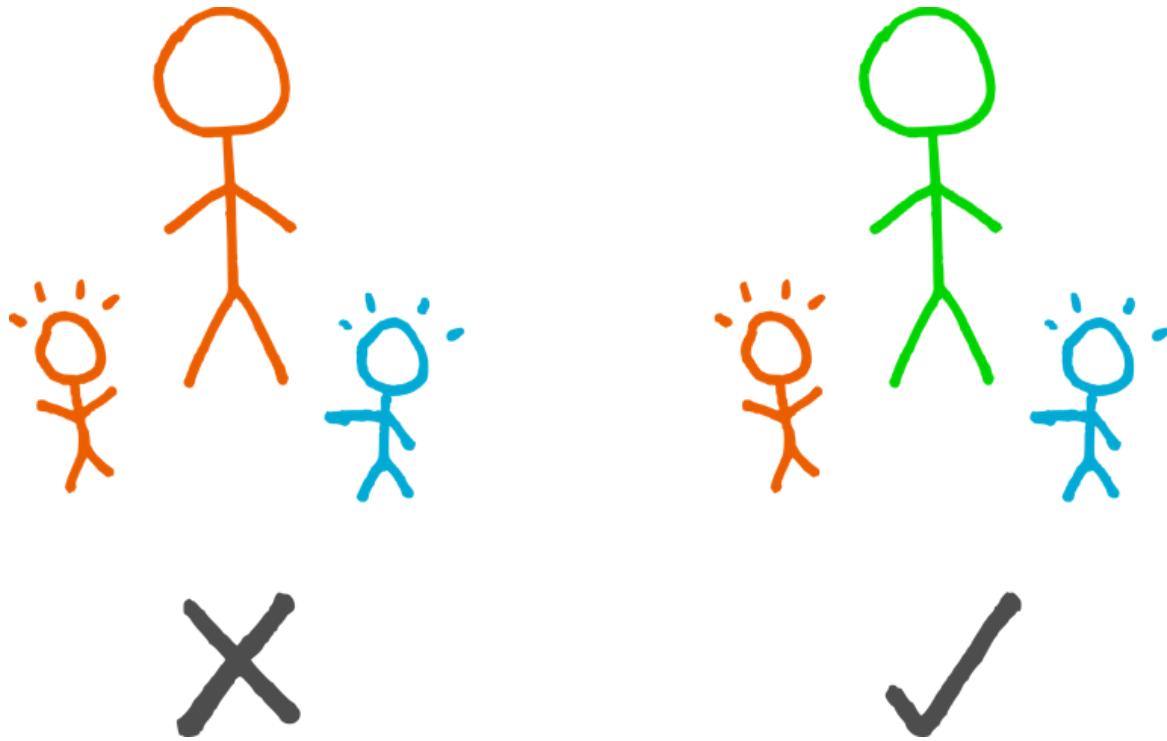
### **Step 1(b): Check the names for resonance and fairness.**

What you are doing during the internal double crux technique is, essentially, *moderating* a debate between two different parts of yourself/two different perspectives you're capable of adopting.

It's often useful to visualize these two different perspectives as something like distressed, angry kindergarteners, each of which is focused mainly on its own priorities and doing whatever it takes to get what it wants.

You, the moderator, want to make sure that you aren't *partisan* in this disagreement. Remember, you are taking as given that *each* side is in possession of some non-negligible pieces of the truth—you want to convince *both* sides to bring their map-fragments to the table, so that you can incorporate all of their information into your larger world model. That

won't happen if you're secretly allied with one side, and helping it beat up on the other.



It's fine to *have more sympathy* for one side than the other, to be clear. And indeed, if this is the case, you definitely want to notice this fact!

But when you choose to play the game of internal double crux, you want to *correct* for that default sympathy, and make sure to offer additional, compensatory support to your inner underdog.

Looking at the names above, it's clear that the moderator is "on Team Go Running." They've given the alternative the epithet "laze around" rather than a phrase that viewpoint might have chosen to describe itself:

Go running

Rest and recharge  
~~Laze around~~

### Step 2: Decide who speaks first.

Which side, if either, *feels more urgency*? Which side is clamoring more loudly to be heard? If you have no clear sense, feel free to just flip a coin.

### Step 2(b): Embody that perspective, and, from that perspective, say one thing.

This is where moderation comes into play. Often, the kindergarteners will want to unleash a flood of words, and it's your job to ease that flow into something productive and

comprehensible.

Step into the mindset of one of the sides. Get in touch with what that side wants. Feel into what it's like to hold [that value], and speak from that point of view.

The usual prompt here is "what's one important thing that the other side *doesn't* understand?" One crucial piece missing from their model of the world; one consideration that perspective is failing to take into account, or failing to weight properly.

Go running

This is the third time in  
a row we've skipped running!  
At this rate we'll never run.

Rest and recharge  
~~taze-around~~

It's important that you try to *actually embody* the viewpoint, rather than doing something more aloof or insulated like imagining what it "would" say. It isn't a performance piece—the idea is to connect with what-it's-like-to-be-the-version-of-yourself-who-really-wants-to-go-running (or whatever), and try to produce *authentic* sentences from that perspective. The skills you sharpened in the previous section on Focusing will come in handy for making sure that the words you write down *actually resonate* with that side.

### Step 3: Get the other side to acknowledge truth.

The overall aim of the exercise is to cause each perspective to *absorb* the truth/wisdom/experience of the other. In order for that to happen, you-the-moderator will encourage each side to start off its turn by first finding *some* grain of truth in what the other side just said:

Go running

This is the third time in  
a row we've skipped running!  
At this rate we'll never run.

Rest and recharge  
~~taze-around~~

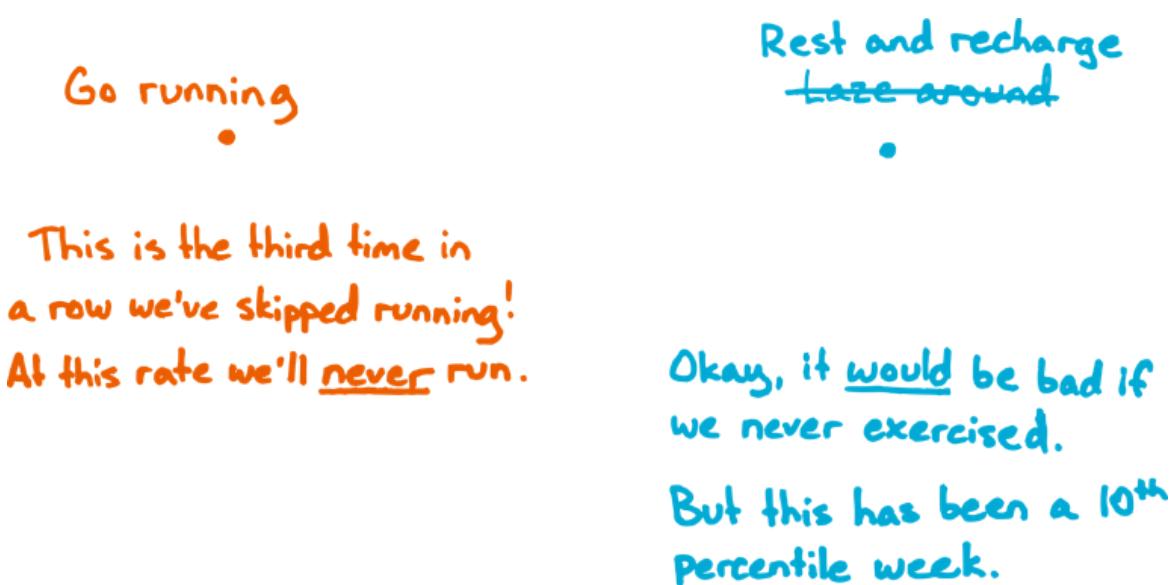
Okay, it would be bad if  
we never exercised.

... it doesn't have to be complete, and sometimes it won't even be something the other side

*directly* said, so much as a logical consequence or an underlying assumption. It just needs to be something that draws the two sides a tiny bit closer, however begrudgingly.

#### **Step 4: That side gets to add its own "one thing."**

Having acknowledged that the other side has some kind of point somewhere, this side now gets to lodge its own objection.



... in this case, that it's been a really rough week, and so perhaps it makes sense that we've skipped running three times in a row.

Notice that this whole point is not necessarily spelled out—you don't have to force yourself to speak in full, coherent, justifiable sentences. Sometimes one side might not even speak in words at all—might draw a picture, or leave a scribble, or just write AAAAAAAA.

It's important to allow these things to happen rather than to impose order from above. You-as-moderator are there to nudge the conversation back on track, as necessary; you don't have to pre-restrict the dialogue to things you've already thought of, or sentences that already pass some kind of filter.

#### **Step 5: Repeat.**

Back and forth, each side should a) acknowledge truth contained within the previous entry, and b) add one new bit of information from its own perspective.

Again, sometimes things do not go according to plan:

Go running

This is the third time in  
a row we've skipped running!  
At this rate we'll never run.

Yeah, it's been rough, but  
there's always an excuse.

Rest and recharge  
~~taze around~~

Okay, it would be bad if  
we never exercised.

But this has been a 10<sup>th</sup>  
percentile week.

You're acting like we have  
some chronic problem!

... in this case, the discussion started going *too quickly*. Orange's acknowledgement was perfunctory, and blue didn't acknowledge at all. Noticing this, you-as-moderator might pause, and look back at the previous orange statement, and see if you can nudge your blue side to make some form of genuine acknowledgement before proceeding.

Okay, fine. I agree that  
a plan with infinite ejector  
seats isn't really a plan.

But I think you're missing  
just how costly it is to burn  
a spoon on this.

It's fine for off-script or against-the-rules things to happen, as long as you bring the conversation *back* toward productive discourse.

I agree with something like "our motivational resources are finite."

But if we don't bump this one up the priority list, we're going to die.

Okay, fine. I agree that a plan with infinite ejector seats isn't really a plan.

But I think you're missing just how costly it is to burn a spoon on this.

AnonUser2784 Fuck You terrorist in my office

... here, for instance, the blue side had something of a minor meltdown.

But that meltdown *did* get written on the page. It's not the job of the moderator to pre-censor, but rather to correct after-the-fact. The moderator gives the blue side a chance to blow up and blow off some steam—to register the *magnitude* of its disagreement, rather than forcing down its reaction—and then gently requires that it nevertheless find some grain of truth in orange's previous point.

I agree with something like "our motivational resources are finite."

But if we don't bump this one up the priority list, we're going to die.

Years are made of -

Okay, yes, one week by itself is probably negligible.

But we have a pattern of making excuses and losing momentum and not following through and we've gained fifteen pounds and -

Fine. What you don't seem to get is that it's never just one week. The way you make these calls is biased and unprincipled.

Anardou2020 Fuck You  
terrorist n i o

That's a disingenuous mugging.

Fine. Dying would be bad.

But screw you, one week isn't going to make a diff!

In this case, once it was orange's turn, orange broke the rules a couple of times—first by

skipping straight ahead to objection, and then by responding with several points instead of one single point.

(Some CFAR participants have benefitted from actually writing down moderator interjections, sometimes in another color. For instance, after "Years are made of—" you might write down "Wait—can we do acknowledgement first?" and after "gained fifteen pounds" you might write "One thing at a time.")

What usually happens over the course of this back-and-forth is that the problem reveals itself to be some *other* problem, usually one that's a layer deeper and more interesting. It's like the couples' therapy truism that "it's never really about the dishes." The question of "should I go running or keep watching Netflix?" is a stand-in for, or an instantiation of, a more complicated dynamic.

Once you realize that—once the underlying disagreement makes itself known—it often helps to draw out a new sheet of paper and draw two new dots with two new names:

**Keep commitments  
to myself**

**Be willing to adapt to  
unforeseen circumstances**

---

That in itself often has tremendous value. Getting a clearer understanding of the deeper generators of various dissatisfactions and internal conflict gives you much better odds of actually solving them (as opposed to flailing around in the dark, patching symptom after symptom).

Why so many little rules?

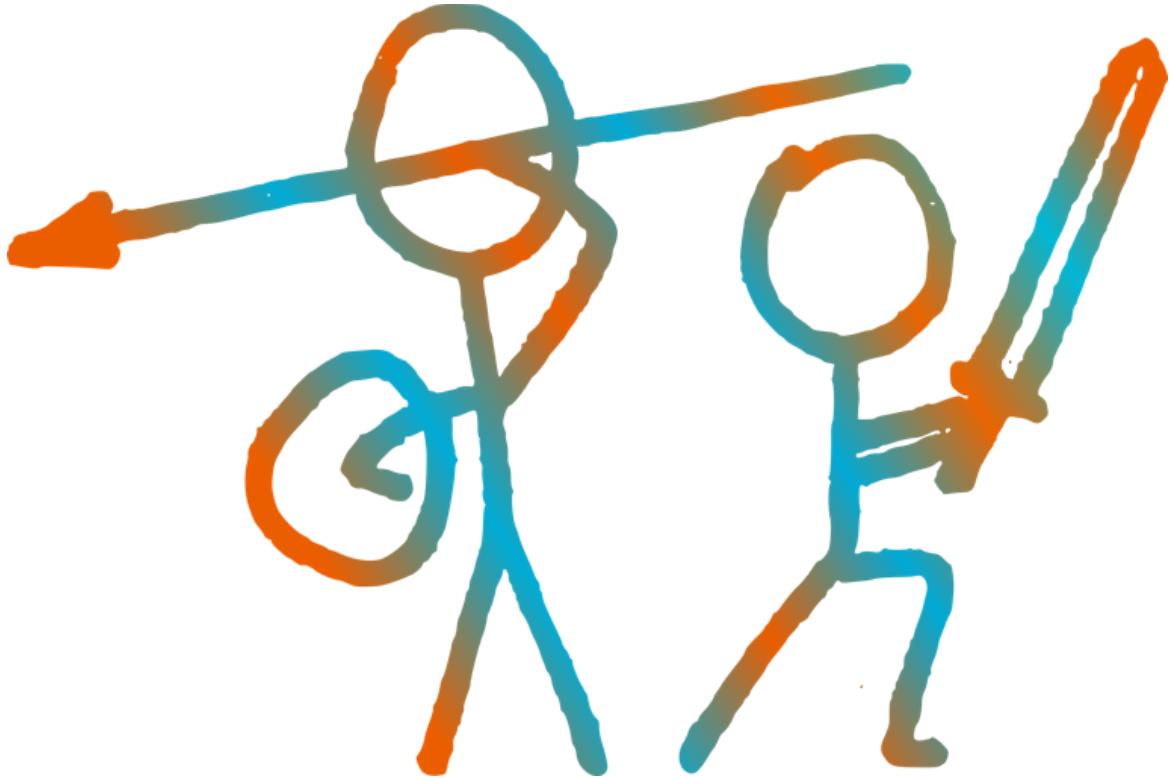
In part because most other ways of resolving internal disagreement seem, to us, to fail to strike at the root of the problem.

Continuing with the metaphor of the kindergarteners—an adult can fairly easily force two kindergarteners to stop fighting. You can separate them, and admonish them, and demand that they behave civilly toward each other, and cow them into compliance.

But until you cause them to *actually be cool with one another*, any cease-fire is going to be fragile, and dependent on the continued presence of an enforcing authority. As soon as the grownup is out of the room for long enough, they'll be right back at each other's throats.

Correspondingly, if one part of your value system is having to repeatedly brute-force overwhelm another part, any cease-fire based on your active attention or conscious consideration is going to be patchy at best. Even if you continue successfully engaging in the "right" behavior every time, you'll still be burning energy and willpower on costly self-control.

By following a process that causes your inner models to *actually understand each other*, you imbue each with some fraction of the wisdom and virtue of the other, leading them to be fundamentally less-in-conflict and better able to support each other (and the higher strategic you).



As always, you should in fact tinker with and iterate on this technique, or abandon it entirely if you find some other method to achieve the goal. But the rigid, rules-based approach has been surprisingly useful to a surprisingly large fraction of participants, so we do honestly recommend *actually trying it* before moving on to your own personal IDC'.

---

## The IDC algorithm

### 0. Find an internal disagreement.

- A “should” that’s counter to your current default action
- Something you feel you aren’t supposed to think or believe (though secretly you do)
- A step toward your goal that feels useless or unpleasant

### 1. Find a charitable handle for each side.

### 2. Embody one perspective and, from that perspective, write down one thing that the other perspective is failing to properly take into account.

### 3. Embody the other perspective and, from that perspective, write down an acknowledgement of one grain of truth in what the previous side had to say.

### 4. Still embodying the second perspective, offer back one counterpoint for the first side to consider.

### 5. Repeat steps 3 and 4 until the disagreement dissolves (or transforms). If useful, start over with step 1 with new names.

---

## IDC—Further Resources

Psychologists Carver and Scheier (2002) use the theory of control systems to model goal pursuit, where feedback about one's progress towards a goal is translated into pleasant or unpleasant feelings. These feelings then motivate the person to continue an effective approach or change an ineffective approach. In order for the system to function smoothly, it is necessary for the relevant part of the system to recognize the connection between the goal and one's current behavior.

Carver, C. S., & Scheier, M. F. (2002). *Control processes and self-organization as complementary principles underlying behavior*. Personality and Social Psychology Review, 6, 304-315. <http://goo.gl/U5WjY>

---

People tend to be more open to information inconsistent with their existing beliefs when they are in a frame of mind where it seems like a success to be able to think objectively and update on evidence, rather than a frame of mind where it is a success to be a strong defender of one's existing stance.

Cohen, G.L., Sherman, D.K., Bastardi, A., McGoey, M., Hsu, A., & Ross, L. (2007). *Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation*. Journal of Personality and Social Psychology, 93, 415-430. <http://goo.gl/ibpGf>

---

“Focusing” is a practice of introspection systematized by psychotherapist Eugene Gendlin which seeks to build a pathway of communication and feedback between a person’s “felt sense” of what is going on (an internal awareness which is often difficult to articulate) and their verbal explanations. It can be understood as a method of querying one’s inner simulator (and related parts of System 1). Gendlin’s (1982) book Focusing provides a guide to this technique, which can be used either individually or with others (in therapy or other debugging conversations).

Gendlin, E. (1982). *Focusing*. Second edition, Bantam Books.  
<http://en.wikipedia.org/wiki/Focusing>

---

“IFS,” or Internal Family Systems is a form of psychotherapy developed by Richard C. Schwartz in which the mind is conceptualized as a set of parts or subpersonalities, each with its own perspectives, interests, memories, and viewpoint, and each with positive intent for the overall person. IFS uses family systems theory (a separate branch of therapy) in a metaphorical way to understand how those subpersonalities are organized and how they interact with one another.

Schwartz, R. (1997). *Internal Family Systems Therapy*. Guilford Publications.  
[https://en.wikipedia.org/wiki/Internal\\_Family\\_Systems\\_Model](https://en.wikipedia.org/wiki/Internal_Family_Systems_Model)

# Abstracting The Hardness of Alignment: Unbounded Atomic Optimization

Crossposted from the [AI Alignment Forum](#). May contain more technical jargon than usual.

*This post is part of the work done at [Conjecture](#).*

## Disagree to Agree

([Practically-A-Book Review: Yudkowsky Contra Ngo On Agents](#), Scott Alexander, 2022)

This is a weird dialogue to start with. It grants so many assumptions about the risk of future AI that most of you probably think *both* participants are crazy.

(Personal Communication about a conversation with Evan Hubinger, John Wentworth, 2022)

We'd definitely rank proposals very differently, within the "good" ones, but we both thought we'd basically agree on the divide between "any hope at all" and "no hope at all". The question dividing the "any hope at all" proposals from the "no hope at all" is something like... does this proposal have any theory of change? Any actual model of how it will stop humanity from being wiped out by AI? Or is it just sort of... vaguely mood-affiliating with alignment?

If there's one thing alignment researchers excel at, it's disagreeing with each other.

I dislike the term pre paradigmatic, but even I must admit that it captures one obvious feature of the alignment field: the constant debates about the what and the how and the value of different attempts. Recently, we even had [a whole sequence of debates](#), and since I first wrote this post Nate [shared his take](#) on why he can't see any current work in the field actually tackling the problem. More generally, the culture of disagreement and debate and criticism is obvious to anyone reading the AF.

Yet Scott Alexander has a point: behind all these disagreements lies so much agreement! Not only in discriminating the "any hope at all" proposals from the "no hope at all", as in John's quote above; agreement also manifests itself in the common components of the different research traditions, for example in their favorite scenarios. When I look at Eliezer's [FOOM](#), at Paul's [What failure looks like](#), at Critch's [RAAPs](#), and at Evan's [Homogeneous takeoffs](#), the differences and incompatibilities jump to me — yet they still all point in the same general direction. So much so that one can wonder if a significant part of the problem lies outside of the fine details of these debates.

In this post, I start from this hunch — deep commonalities — and craft an abstraction that highlights it: unbounded atomic<sup>[1]</sup> optimization (abbreviated UAO and pronounced wow). That is, alignment as **the problem of dealing with impact on the world (optimization) that is both of unknown magnitude (unbounded) and non-**

**interruptible (atomic).** As any model, it is necessarily mistaken in some way; I nonetheless believe it to be a [productive mistake](#), because it reveals both what we can do without the details and what these details give us when they're filled in. As such, UAO strikes me as a great tool for [epistemological vigilance](#).

I first present UAO in more details; then I show its use as a mental tool by giving four applications:

- **(Convergence of AI Risk)** UAO makes clear that the worries about AI Risk don't come from one particular form of technology or scenario, but from a general principle which we're pushing towards in a myriad of convergent ways.
- **(Exploration of Conditions for AI Risk)** UAO is only a mechanism; but its abstraction makes it helpful to study what conditions about the world and how we apply optimization lead to AI Risk
- **(Operationalization Pluralism)** UAO, as an abstraction of the problem, admits many distinct operationalizations. It's thus a great basis on which to build [operationalization pluralism](#).
- **(Distinguishing AI Alignment)** Last but not least, UAO answers [Alex Flint's question](#) about the difference between aligning AIs and aligning other entities (like a society).

*Thanks to TJ, Alex Flint, John Wentworth, Connor Leahy, Kyle McDonell, Laria Reynolds, Raymond Arnold, Steve Byrnes, Rohin Shah, Evan Hubinger, James Lucassen, Rob Miles, Jamie Bernardi, Lucas Teixeira, and Andrea Motti for discussions on these ideas and comments on drafts.*

## Pinning UAO down

Let's first define this abstraction. Unbounded Atomic Optimization, as the name subtly hints, is made of three parts:

- **(Optimization)** Pushing the world towards a given set of states
- **(Unboundedness)** Finite yet without a known limit
- **(Atomicity)** Uninterruptible, happens "as in one step"

## Optimization: making the world go your way

Optimization seems to forever elude full deconfusion, but an adaptation of [Alex Flint's proposal](#) will do here: **optimization is pushing the world into a set of states**.

[2] Note that I'm not referring to computational optimization in the sense of a search algorithm; it is about changing the physical world.

When I'm talking about "amount" of optimization, I'm thinking of an underdefined quantity that captures a notion of how much effort/work/force is spent in pushing the world towards the target set of states. Here's a non-exhaustive list of factors that can increase the amount of optimization needed:

- **(Small target set)** Hitting a smaller target requires more effort
- **(Far away target set)** If there are large changes from the current state to the target set, it takes more effort to reach it.
- **(Stronger guarantees)** If the target set must be reached with high probability, it takes more effort

- **(Robustness)** If the world must be maintained in the target set, it takes more effort
- **(Finer state-space)** If the granularity of states is finer (there are more details in the state descriptions), then it might take more effort to reach the set.

## Unboundedness: phase transition in optimization

**Humans optimize all the time, as do institutions, animals, economic systems, and many other parts of our world. But however impressive the optimization, it is always severely bounded.** We talk about absolute power for a king or an emperor, but none of them managed to avoid death or maintain their will for thousands of years yet (most couldn't even get their teeth fixed better than paupers).

**Classical scenarios of AI risk, on the other hand, stress the unboundedness of the optimization being done.** Tiling the whole lightcone with paper clips gives a good example of massive amounts of optimization.

Another example of unbounded optimization common in alignment is manipulation: the AI optimizing for convincing the human of something. We're decently good at manipulating each other, but there's still quite clear bounds in our abilities to do so (although some critical theorists and anthropologists would argue we underapproximate the bounds in the real world). If the amount of optimization that can be poured into manipulation is arbitrarily large, though, we have no guarantee that any belief or system of beliefs is safe from that pressure.

More generally, unbounded optimization undermines solutions that are meant to deal with only some reasonable range of force/effort (like [buttresses](#) in structural engineering). So it means that no amount of buttresses is enough to keep the cathedral of our ideals from collapsing.

## Atomicity: don't stop me now

In [distributed computing](#), an atomic operation is one that cannot be observed “in the middle” from another process — either it didn’t happen yet, or it’s already finished. Ensuring atomicity plays a crucial role in abstracting the mess of distributed interleavings, loss of messages, and other joys of the cloud.

I use atomic analogously to mean “uninterruptible in practice”. It might be physically possible to interrupt it, but that would require enormous amounts of resources or solving hard problems like coordination.

In alignment, we’re worried about atomic optimization: the optimization of the world which we can’t interrupt or stop until it finishes.

What does this look like? [FOOM](#) works perfectly as an initial example: it instantiates atomicity through exponential growth and speed difference — you can’t stop the AI because it acts both far too smartly and quickly. But the whole point of using atomicity instead of FOOM is to allow other implementations. Paul Christiano ([What failure looks like](#)), Evan Hubinger ([Homogeneity vs heterogeneity in AI takeoff scenarios](#)) and

Andrew Critch ([What Multipolar Failure Looks Like](#)) all propose different AI Risks scenarios with atomicity without FOOM. Instead of speed, their atomicity comes from the need to solve a global coordination problem in order to stop the optimization. And coordination is just hard.

## Application 1: Highlight Convergence of AI Risk Scenarios

In almost any AI Risk story, you can replace the specific means of optimization with UAO, and the scenario still works.

For me, this highlights a crucial aspect of Alignment and AI Risk: it's never about the specific story. I get endlessly frustrated when I see people who disagree with AI Risk not because they disagree with the actual arguments, but because they can't imagine something like [FOOM](#) ever happening, or judge it too improbable.<sup>[3]</sup>

The problem with this take is not that FOOM is obviously what's going to happen with overwhelming probability (I'm quite unconvinced of that), but that it doesn't matter how UAO is implemented — as long as we have it, we're in trouble.

And because UAO based arguments abstract many (all?) of the concrete ones, they are at least as probable (and probably strictly more probable) as any of them. Not only that, they even gain from new formulations and scenarios, as these offer additional mechanisms for implementing UAO. So having a variety of takeoff speeds, development models, and scenarios turns from a curse to a boon!

What this also entails is that to judge the probability of these risks, we need to assess how probable UAO is, in any implementation.

## Convergence to UAO

To start with unboundedness, it follows straightforwardly from technological progress. Humanity is getting better and better at shaping the world according to its whims. You might answer that this leads to many unwanted consequences, but that's kind of the point, isn't it? At least no one can say that we don't have a massive impact on the world!

This is also where AI gets back into the picture: ML and other forms of AI are particularly strong modern ways of applying optimization to the world.<sup>[4]</sup> And we currently have no idea where it stops. Add to that the induction from past human successes that huge gains can come from insights into how to think about a problem, and you have a recipe for massively unbounded optimization in the future.

As for atomicity, it has traditionally been instantiated through three means in AI Risk arguments:

- **(Computers running faster than humans)** Current computers are able to do far more, and faster, than humans if given the right instructions, and Moore's law and other trends don't give us good reason to expect the gap to stop growing. This massive advantage incentivizes progress in optimizing the world

to go through computers, making the optimization increasingly atomic.<sup>[5]</sup> More generally, human supervision becomes the bottleneck in any automated setting, and so risks getting removed to improve efficiency if the system looks good enough.

- **(Inscrutability pushed by competitiveness)** Ideally, everyone would want to be able to understand completely what their AI does in all situations. This would clearly help provide guarantees for customers and iterate faster. But the reality of ML is that getting even half there is extraordinarily difficult, it costs a lot of time and energy, and you can get amazing results without understanding anything that the model does. So competitiveness conspires to push AGI developers to trade interpretability and understanding for more impressive and marketable capabilities. This gulf between what we can build and what we can understand fuels atomicity, as we don't have the mental toolkits to check what is happening during the optimization even if it is physically possible.
- **(Coordination failures)** Humans are not particularly good at agreeing with each other in high stakes settings. So if the only way to stop the ongoing optimization is an economy-wide decision, or an agreement to not use a certain type of model, we should expect enormous difficulties there. And in a situation (the use of more and more optimization in the world) where free riders can expect to reap literally all the benefits (if they don't die in the process), it's even harder to agree to stop an arms race.<sup>[6]</sup>

The gist is that **we're getting better and better at optimizing, through technology in general and computers and automation in particular. This in turn leads to a more and more atomic use of optimization, due to the high speed of computers and the incentives to automate. With the compounding effect of the difficulty to coordinate, we have an arms race for building more and more atomic optimization power, leading to virtually unbounded atomic optimization.**

## Application 2: Explore Conditions for AI Risk

While UAO is a crucial ingredient of AI Risk, it is not enough: most scenarios need some constraints on how UAO is applied. The abstraction of UAO lets us then focus on exploring these conditions, to better understand the alignment problem. As such, UAO provides a crucial tool for [epistemological vigilance](#) on the assumptions underlying our risk scenarios.

Let's look at two classes of proxies for an example: overapproximation proxies and utility maximization proxies. These two capture many of the concrete proxies that are used in AI Risk scenarios, and illustrate well how UAO can clarify where to investigate.

### The Danger of Overapproximations

Overapproximation proxies point to quite reasonable and non-world-shattering results, like "Make me rich".

Here are their defining properties:

- **(Looseness)** The proxy is one where the target set massively overapproximates the set of states that we actually want. For example, “Make me rich” as operationalized by “Make my bank account show a 10 digits number” contains a lot of states that we don’t really want (including those where I’m dead, or where the Earth is turned into computers that still somehow keep track of my bank account). The set of states I’m thinking of only makes a particularly small subset of the proxy’s set.
- **(Reliability)** The proxy asks for reaching the set of states with high probability
- **(Robustness)** The proxy asks that the change sticks, it doesn’t get out of the target state after the optimization is done.

Let’s look at what happens when we apply UAO to such proxies. Our proxy gives us a fixed, overapproximated target set of states. Let’s say something like “produce 20 billion paperclips in the United States per year” (about twice the current amount). You don’t need to tile the universe to reach that target at all. So it’s relatively easy to end up in the set of states we’re aiming for. But what about reliability and robustness, the other two requirements of the proxy? Well if you want to guarantee that you’ll reach the target set and not get out of it, one way to do so is to aim for the part of this target state that is more controlled and guaranteed.<sup>[2]</sup> Like for example, the one where the Earth is restructured for better paperclip-making conditions (without these bothering humans for example!). **As the optimization increases, it is increasingly spent on reliability and robustness, which strongly incentivizes using the many degrees of freedom to guarantee the result and its perennity. Hello instrumental convergence!**

The story is thus: unbounded atomic optimization + overapproximate proxies => incentive for numerous degrees of freedom to be used in systematically bad ways.

Note that if we want to avoid this fate, our abstract conditions give us multiple points of intervention:

- We want to change the proxy to have better properties.
- We might question whether realistic proxies have these properties.
- We might want to fill in details of how the UAO is used, to break the argument somewhere in the middle.
- We might want to remove or simplify some of the constraints on the proxy, to see if we can strengthen the argument by weakening its hypotheses.
- We might want to question the actual strength of the incentives, to break the argument in the abstract.

## Terrible Returns on Utility

Utility maximization proxies are specified by the maximal states according to some utility function. It should come to no surprise to readers of this post that maximizing utility can lead to terrible outcomes — the question is: what is needed for that to happen?

This part shows more how UAO can lead to asking relevant questions. My current best guess is that we also need two conditions on the proxy:

- **(Beyond the goal)** The utility function is such that the actual states of the world that we want, the ones that we’re visualizing when coming up with the utility function, have far less than maximal utility. So when I ask to maximize

money, the amounts I have in mind are far smaller than the ones that can be reached in the limit (by converting all the universe into computronium and thus encoding a massive number for example).

- (**Terrible upper set**) There is a threshold of utility that is physically reachable and such that every state with at least that much utility is terrible for us.

With these two conditions, it follows that UAO will push us into the terrible upper set, and lead to catastrophic AI Risk.

The interesting bit here lies in analyzing these conditions for actual utility functions, like “maximizing paperclips”. And just like with the overapproximation proxies, multiple points of interventions emerge from this analysis:

- We want to change the proxy to have better properties.
- We might question whether realistic proxies have these properties.
- We might want to fill in details of how the UAO is used, to break the argument somewhere in the middle.
- We might want to remove or simplify some of the constraints on the proxy, to see if we can strengthen the argument by weakening its hypotheses.

## Application 3: Anchor Operationalization Pluralism

In [my last post](#), I discussed different levels at which pluralism might be applied and justified. The one that UAO is relevant to in my opinion is [operationalization pluralism](#), or pursuing multiple operationalization (frames/perspectives/ways of filling the details) for the same problem.

Because the tricky part in operationalization pluralism is to capture the problem abstractly enough to allow multiple operationalization, without losing the important aspects of the problem.

UAO provides one candidate abstraction for the alignment problem.

In some sense, UAO acts as [a fountain of knowledge](#): it rederives known operationalizations when you fill in the implementation details or make additional assumptions. As such, it serves both as a concrete map and as a tool to explore the untapped operationalizations. We can pick unused assumptions, and generate the corresponding operationalization of the alignment problem.

Three concrete ways of generating operationalizations are

- Specifying the implementation details to make the problem more concrete.
- Staying at the abstract level, but privileging the study of one intervention on how UAO will be applied.
- Starting with epistemic tools, and operationalizing UAO in the way that is most susceptible to yielding to these tools.

Let's look at examples of all three in alignment research.

# Filling in the blanks: neural nets, brain-like algorithms and seed AI

The obvious way of operationalizing UAO is to make it concrete. This is exactly what Prosaic Alignment, Steve Byrnes' Brain-like AGI Alignment and some of MIRI's early work on seed AIs do.

- [Prosaic Alignment](#) assumes that UAO will be instantiated through neural networks trained by gradient descent. It is somewhat agnostic to architecture and additional tricks, as long as these don't cross a fuzzy boundary around a paradigm shift.<sup>[8]</sup> Also, despite Paul Christiano's doubts about FOOM, prosaic alignment doesn't forbid FOOM-like scenarios.
- Steve Byrnes' [research](#) assumes that UAO will be instantiated through reimplementing the learning algorithms that the human neocortex uses. It doesn't really specify how they will be implemented (not necessarily neural nets but might be).
- MIRI's early work (for example [modal combat and work on Loeb's theorem](#)) assumed that UAO would be instantiated through hand-written AI programs that were just good enough to improve themselves slightly, leading to an intelligence explosion (with a bunch of other assumptions). A running joke was that it would be coded in [LISP](#), but implementation details didn't really matter, so long as the initial code of the seed was human intelligible and human crafted.
- Critch's [RAAPs](#) assumes that UAO is instantiated through structure, that is through the economy itself (unbounded atomic capitalism, if you prefer).

These assumptions were historically made from a normative perspective: each researcher believed that this kind of AI was either the most probable, or had a significant enough probability to warrant study and investigation.<sup>[9]</sup>

But here we're starting from UAO instead. By making these additional assumptions, each operationalization unlocks new ways of framing and exploring the problem. As an analogy, in programming language theory, the more generic a type, the less you can do with it; and the more specific it becomes, the more methods and functions can be used on it. So if we assume that UAO will be instantiated as neural networks trained by gradient descent, we have more handles for exploring the general problem and investigating mechanisms. A perfect example is the small research tradition around [gradient hacking](#), which looks for very concrete neural networks implementations of a certain type of treacherous turn incentivized by instrumental convergence.

Yet there are also risks involved in such an instantiation. First, if the instance is a far simpler case than the ones we will have to deal with, this is an argument against the relevance of solving that instance. And more insidiously, what can look like an instantiation might just pose a completely different problem. That's one failure mode when people try to anchor alignment in ML and end up solving purely bounded optimization problems without any theory of change about the influence on unbounded atomic optimization.<sup>[10]</sup>

## Working directly on the abstraction

Another category of operationalizations stays at the abstract level, and focuses instead on one possible intervention on UAO as the royal road to alignment. A lot of the work published on the AF fits this category, including almost all deconfusion.

[11] Among others, there are:

- John Wentworth's [work](#) on Abstraction and the Natural Abstraction Hypothesis, which focuses on finding [True Names](#) for human values, in order to not have proxies but the real deal.
- Quintin Pope, Alex Turner, Charles Foster, and Logan Smith's [work](#) on shard theory, which focuses on a structural way of counterpowers which allow more optimization to be spent without the classical failure modes.
- Stuart Armstrong's [work](#) on Model-Splintering, which focuses on how to extend values when more optimization leads to shifts in ontology.

The tricky part is that so much of the work at this level looks like fundamental science: it's about exploring the problem almost as a natural object, in the way computer scientists would study a complexity class and its complete problems. In the best cases, this level of abstraction can yield its secrets to simple and powerful ideas, like "[high-level summary statistics at a distance](#)" or "[counting options through permutations](#)". But even then, drawing conclusions for the solution of the problem is hard, and requires [epistemological vigilance](#).

That being said, such work still plays a crucial role in alignment research, and we definitely need more of it. Even when working from within an instantiation like prosaic alignment, it's often fruitful to move between this level and the more concrete. I conjecture that it comes both from the purity of the models used (which leads to focus on nice math) and from removing the details that obscure or hide the core of unbounded atomic optimization.

## Privileging particular tools

The last category in my non-exhaustive list are those operationalizations which start from their methods and the veins of evidence where they go searching for hidden bits.

- Vanessa Kosoy's [work](#) is the most obvious example to me, with a focus on extracting knowledge and solutions through computational learning theory. This also comes with some instantiation assumptions (that the AI is a Bayesian RL model), but those are significantly less concrete and constrained than in prosaic alignment, for example.
- Andrew Critch's [RAAPs](#) looks like a framing of alignment and UAO fitted to the analysis of structures, from sociology to computational social choice.
- Steve Byrnes's [work](#) also fit as a research programme driven by neuroscience. I don't know if he would agree with this characterization, but it still looks like a fruitful framing to me.
- A tradition of mostly MIRI work but also some CLR and some independent research focuses on decision theory and how it can clarify issues around the alignment and the consequences of UAO.

Here the risks are to take an irrelevant field, or one with only superficial links to alignment. I think it's possible to analyze the expected productivity of an analogy, for example based on the successes in that field. Also relevant, if the field in question doesn't have many successes, is whether the analogy reduces alignment to a

currently really hard problem (like P vs NP), or to some simpler problem that these other fields have a reasonable chance to tackle.

My attitude to this category of operationalization is that we should look for even more opportunities and bring as many analogies as we can, as long as we expect them to become productive for alignment. The [PIBSS Fellowship](#) is pushing in that direction, and I expect a clearer framing of the constraints to help.

## Application 4: Separate AI Alignment From Other Forms of Alignment

As a final application of UAO, let's separate alignment of AIs from other forms of alignment.

Here I want to turn to Alex Flint's nice analysis of [Alignment vs AI Alignment](#),<sup>[12]</sup> where he attempts to separate aligning AI from alignment of other systems like oneself or society. Concretely, his non-AI examples are:

- Aligning a society through property rights
- Aligning a society through laws
- Aligning one's own cognition through habit formation
- Aligning a company via policy and incentives
- Aligning animals through selective breeding and domestication in general

Alex then asks what separates aligning an AI from all these examples.

My answer: **the combination of unboundedness and atomicity in the optimization.** In all these examples, unbounded optimization applied atomically is irrelevant. In principle each example can be optimized somewhat unboundedly, but it happens so slowly that we can iterate — an assumption requiring [epistemological vigilance](#) in alignment.

Or said differently, it's unbounded optimization but applied little by little, with time to change course in between. Just like cathedral builders could see cracks and failure happening over the course of decades and correct them.

Note that this doesn't mean these fields can't help with alignment. Just that alignment is qualitatively different from the phenomena traditionally tackled by economics, behavior change, and these other fields. This difference must be kept in mind when building a theory of change for applying insights from these other disciplines.

## UAO, a Productive Mistake

We've seen that unbounded atomic optimization serves in multiple applications:

- It highlights the convergence of multiple concrete instantiations of AI Risk by abstracting them all.
- It helps in formulating and exploring conditions for AI Risk.
- It gives a framing for operational pluralism in alignment.
- It separates AI alignment from alignment of other systems.

This makes me think that UAO is [a productive mistake](#).

How is it a mistake? That is, what does it hide away or distort? Mostly it assumes the hardness of the problem. Some people believe that alignment is significantly easier than dealing with UAO — maybe the increases in optimization between iteration of AIs will be slow enough to adapt and break atomicity, for example. I'm personally dubious of such simplifications, as they look more like wishful thinking than arguments to me. But UAO is definitely colored by my takes, and my general stance towards [epistemological vigilance](#).

Still, UAO can act as a characterization of the hard alignment problem that is more conducive to debates about the difficulty of alignment and the assumptions we can get away with.

1. ^

Here the word "atomic" refers to the etymological meaning "indivisible", rather than the common usage "small"

2. ^

This setting can deal with utility functions by focusing on the sets with maximal utility (which exists because there are finitely many states).

3. ^

How do I know that they might agree with the actual argument? Because most often, when I then present them a more structural implementation of UAO like [Critch's RAAPs](#), they end up agreeing with the risks!

4. ^

Here again, it's important to note that I'm using optimization in the "physically changing the world" sense, not in the computational "internal search" sense. So what AI gives us here is the ability to "internally search" for better ways of acting in the world, and this whole process fits under what I call optimization.

5. ^

This is where the atomicity comes from in fast takeoffs and FOOM-like scenarios.

6. ^

Exploring these structural factors is the big contribution of Critch's [RAAPs](#) in my opinion.

7. ^

This is but another way of framing Bostrom's insightful point about how even a wireheading AI would have reasons to tile the universe to protect itself and its wireheading.

8. ^

Important to note that this subclass of alignment is comparatively far larger (at least in terms of active research) than the other two, and has additional specializations (for example whether the NN will be trained by RL or self-supervised learning).

9. ^

Critch feels like a strong exception, because I interpret his introduction of RAAPs as an attempt to add structural perspective to alignment to round off the field. And although Paul believes in the normative claim that the first AGI will probably be prosaic, he [does argue](#) that even if that's not the case, we should expect a solution to prosaic alignment to translate to the other version and capture some hard parts of the problem. And when I asked him the question, he told me that what mattered was to make the problem well-defined.

10. ^

See [this post](#) for an exploration of the common assumptions that need to be questioned in alignment to not fall into this trap.

11. ^

Some exceptions are Evan Hubinger's et al. [inner optimization](#) and Paul Christiano's [universality](#), which are tailored for prosaic alignment. Yet they end up being useful for other approaches too.

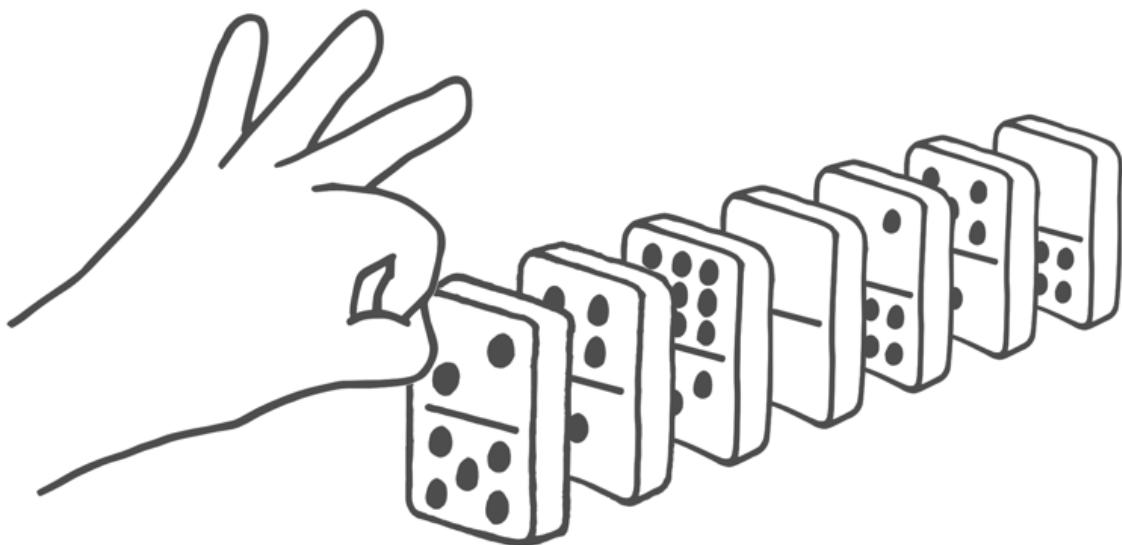
12. ^

Discussions with Alex while he was writing that post ultimately led me to realizing the need for the atomicity condition, so he gets the credit for that!

# Trigger-Action Planning

**Epistemic status:** Established and confirmed

*There has been a tremendous amount of research on “implementation intentions” since their development by psychologist Peter Gollwitzer in the late 1990’s. A meta-analysis of 94 studies involving 8461 participants found that interventions using implementation intentions were an average of .65 standard deviations more effective than control interventions. Similar effect sizes were found in the 34 studies which looked at behavioral change on personal or health goals (average of .59 standard deviations more effective). Trigger-action planning—our version of implementation intentions—draws directly on this research and has proven useful to the majority of our alumni for a wide range of problems, tasks, and goals.*



In previous sections of this book, we’ve looked at the differences between System 1 and System 2, talked about the process of turning goals into plans, and learned to distinguish useful and relevant practice from irrelevant or unproductive practice. In this section, we will combine those insights and their implications into a single, robust technique for building awareness and supporting behavioral change.

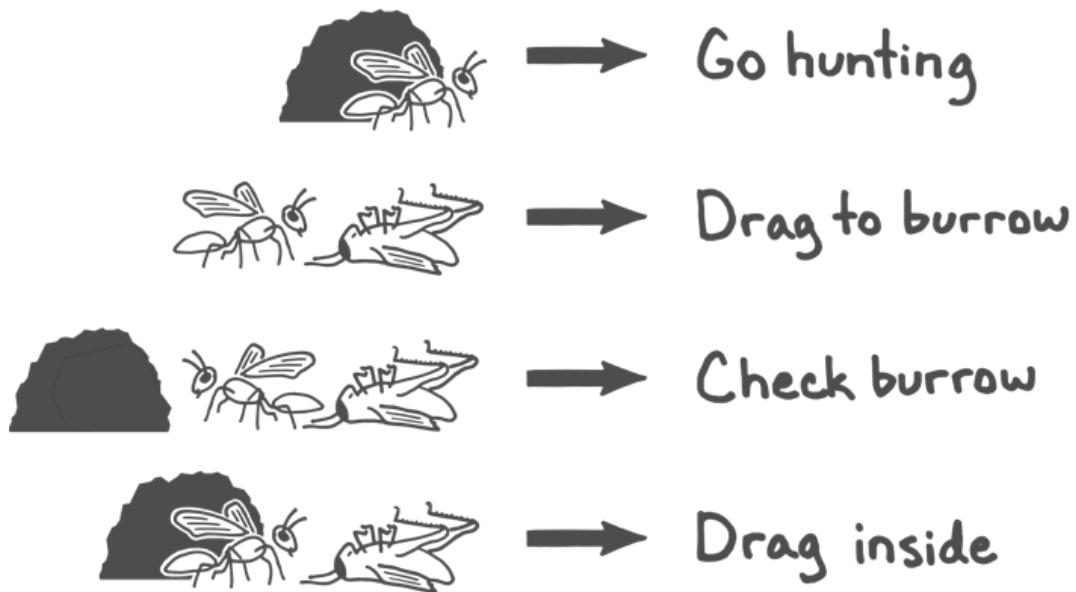
## Complex chains: The parable of the Sphex

*Sphexes* are a genus of wasps, and for many years, a story about their behavior has been a major touchstone in cognitive science. Typically, when it comes time for egg laying, a sphex will build a burrow and fill it with paralyzed insects for her future larvae to eat. When hunting, she will sting her prey, wait for the venom to take effect, drag the prey back to the burrow entrance, leave it outside while she goes in and reconnoiters (presumably confirming the absence of predators or structural problems), and finally come back out to drag her victim inside.

This sequence of actions is elaborate, organized, and complex, and on the surface seems to indicate an impressive level of mental sophistication for an insect whose brain weighs less than a milligram. However, in 1879, French entomologist Jean Henri Fabre decided to dig deeper:

I will mention an experiment...at the moment when the Sphex is making her domiciliary visit, I take the cricket left at the entrance to the dwelling and place her a few inches farther away. The Sphex comes up, utters her usual cry, and...comes out of her hole to seize it and bring it back to its right place. Having done this, she goes down again, but alone [once more leaving the cricket outside]. I play the same trick upon her, and the Sphex has the same disappointment on her return to the surface. The victim is once more dragged back to the edge of the hole, but the wasp always goes down alone. . . forty times over, did I repeat the same experiment on the same wasp; her persistence vanquished mine and her tactics never varied.

Fabre's own experiments on other wasps (from the same colony, from the same species but other colonies, and from other species) showed that this was not the only possible result—many wasps eventually break the pattern and drag their prey straight into the burrow. But even the quickest tend to repeat themselves four or five times, implying that the overall process is less a single, coherent strategy and more a series of disconnected if-then actions:



By “chaining together” a series of simple, atomic responses (e.g *if I come out of my burrow and there's a paralyzed cricket, drag it inside immediately*), the sphex is able to execute complex, multi-step behaviors as if it were capable of thinking and planning ahead—even though it largely isn't. The “intelligence” lies in the *algorithm*, rather than in active cognition.

---

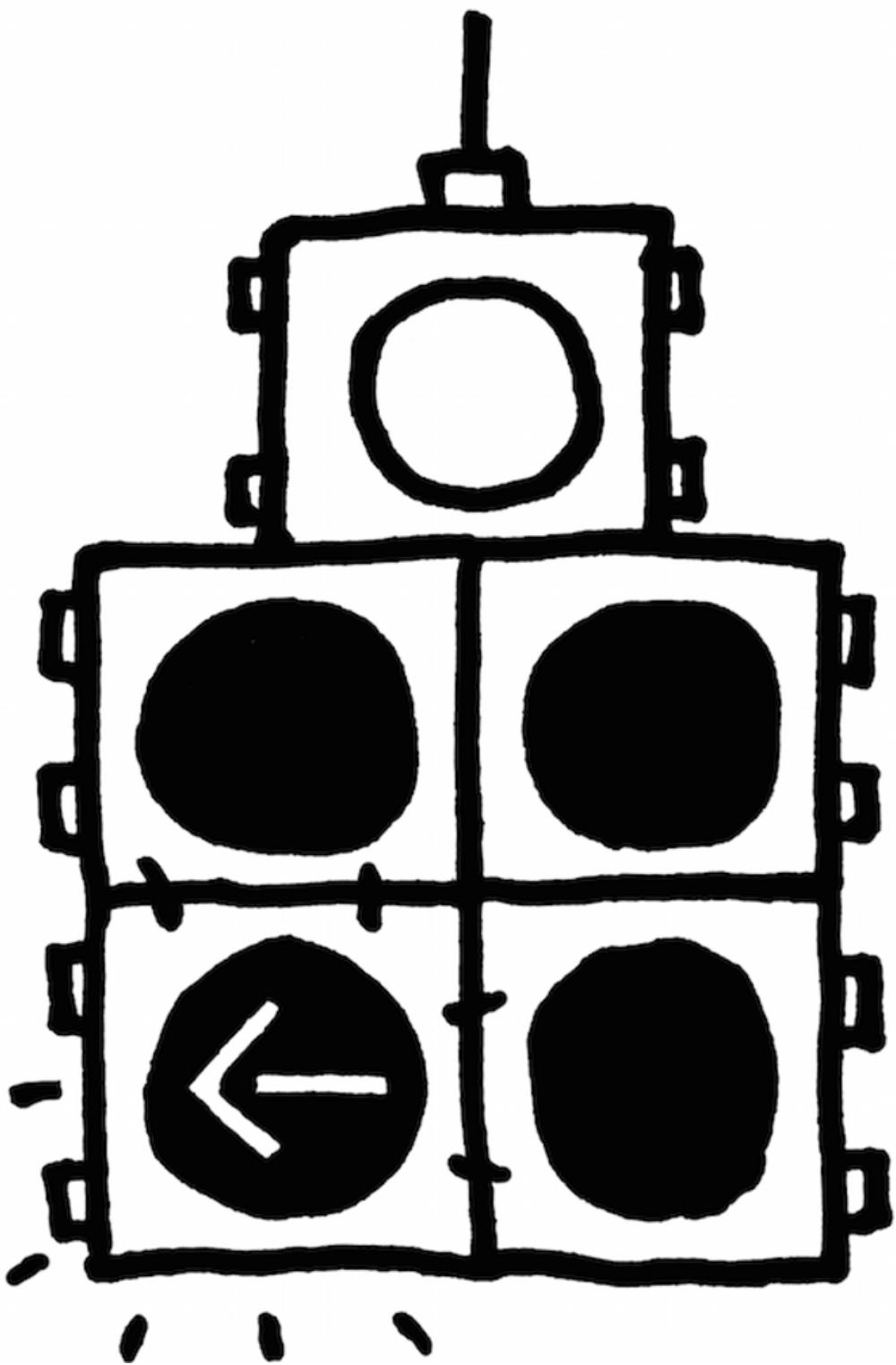
## The trigger-action pattern

There is another species that is capable of chaining together a series of atomic reflex actions into complex and appropriate behavior without any need for active cognition—humans! In many respects, this is what our System 1 is *for*—it's constantly running in the background, aggregating all of our lived experiences and guiding our actions when we're not paying attention. It's because of our System 1 that we can do things that approximate multitasking—carrying on conversations while eating, thinking about upcoming weekend plans while driving in light traffic, exercising while watching TV.

One of the ways we manage this is with a host of trigger-action patterns, derived from our model of the universe and constantly reinforced through experience:

- Someone sneezes? → Say "bless you" or "gezundheit."
- Bowl of chips in front of you? → Grab one. (Grabbed one? Eat it!)
- Hear a buzz or a ping? → Pull your phone out of your pocket.
- Opened a web browser? → Go to [your usual first-click site].
- Open the fridge after shopping? → Realize you forgot to get milk.

These actions are generally quick and effortless, with our conscious minds rarely getting involved (and usually only if we run into problems, like when we get caught in "...I'm fine, and you?" loops, or when you head toward the office even though it's the weekend, or when the left turn arrow causes you to take your foot off the brake, even though you're going straight).



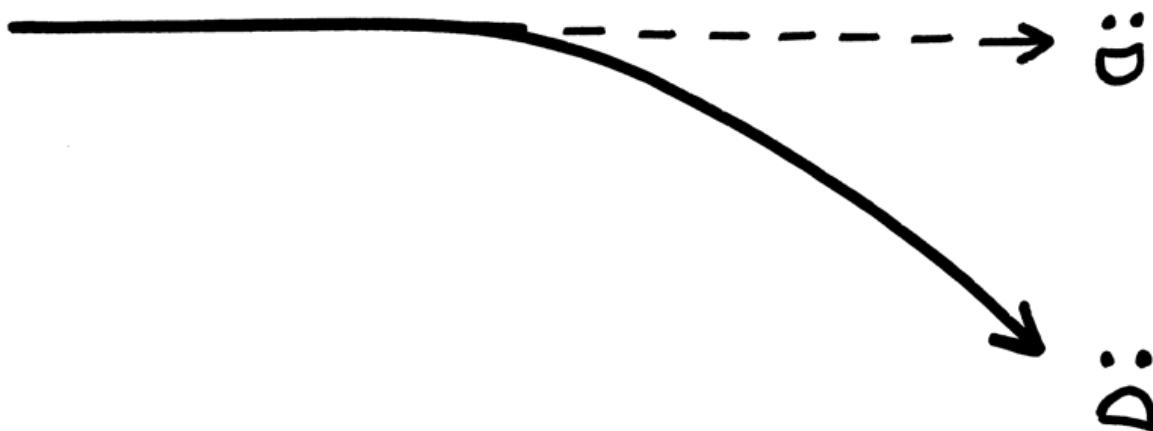
The examples above are single-step, but we all have chains, as well, for any complex task we've spent time reinforcing—the series of actions you take in the shower or upon arriving home, the lines of an argument you've had ten times already, the flow you experience while

playing sports or working with machinery or playing jazz or pushing code to Github. Most people who drive spend only the tiniest amount of attention *actively thinking about driving* while on the road—barring heavy traffic or sudden surprises, we maintain control of our cars with dozens of trigger-action patterns.

Not every pattern is visible or obvious, either—think about the triggers that cause you to smile, or sigh, or tense up, the reliable causes of a good or bad day. We each have triggers which result in a particular emotion (often referred to as trigger-affect patterns), or triggers which bring specific words or memories to mind (like the first few words of a well-known song, or the first half of a common phrase). Sometimes these can chain and reinforce, too, all inside our heads—some stray thought triggers an emotion, and that emotion triggers another thought, which reminds us of something else, which elicits further feelings, and so on.

---

Consider the following:



You're trucking along, living a generally good and happy life, and then *something happens*, and you find yourself in the sad timeline instead of the happy one. You ate an entire package of Oreos, despite intending to lose weight. You got in another fight with your romantic partner, despite genuinely not wanting to. You just straight-up forgot about your new year's resolution; it never came to mind. You road-raged, you failed to finish the presentation before the deadline, you spent all evening on Reddit instead of thinking about your research, you somehow never called them back and now it's been months and it feels too awkward.

There are a few interesting takeaways from thinking about situations like those *in terms of* an image like the one above.

First: for most goals and values, there *actually exists* a moment-of-departure from [a path consistent with the positive outcome] and [a path consistent with the negative one]. There is usually an identifiable point at which one of those outcomes becomes distinctly more likely than the other (though it may be hard to pinpoint, even after the fact).

Second: in most cases, the paths tend to get farther and farther apart over time. It's rare that one *instantaneously* and *irrevocably* leaps from 😊 to 😢; most of the time, there is a shift in *trajectory*, and one's prognosis worsens as continued-progress-along-the-wrong-path compounds.

You could think of the distance between the dotted and solid lines as a measure of the *total effort required* to make it back to the better timeline. The quicker you notice that you've

changed course, the shorter the distance back to the better path. The less time that you've spent accelerating in the wrong direction, the less inertia you have to overcome.

Which leads to one of the key actionable insights of the TAPs perspective: there are times when the total effort to switch from 😞 to 😊 is zero, or close enough—e.g. simply catching the moment when you *would* have made the unfortunate switch, and then not doing so. In many, many cases, an epsilon of prevention is worth an omega of cure.

To put it another way:

It's not *that* far off to declare humans to be simply slightly-more-complicated sphex wasps, largely following the path of least resistance in accordance with a preset autopilot.

To change the outcome of a given situation, then, there must be either a) some *change to the autopilot itself*, or b) some turning-off of the autopilot, summoning effortful sapience.

In both a) and b), it pays to know *which* moment is the critical moment—either which specific if-then to change, or when a pilot's attention will do better than the preprogrammed defaults.

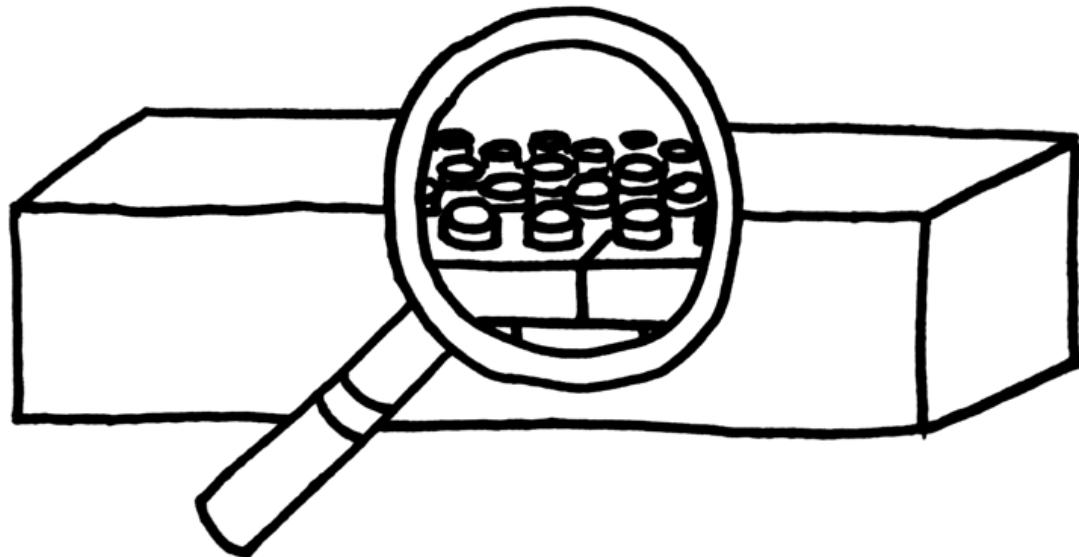
---

## TAPs: From patterns to plans

A full understanding of trigger-action patterns requires close attention to concrete detail. It's less about things like "when I exercise, I get discouraged" and more about "when I run for a while, my chest starts to ache, and when my chest starts to ache, I start thinking about how far away the end is, and when I start thinking about how far away the end is, my enthusiasm for getting fit vanishes."

In cognitive behavioral therapy, patients are often taught to monitor their thoughts for specific words or phrases that have emotional power; kids who struggle with ADHD are sometimes encouraged to note *exactly* what happened right before they got distracted, and the *first thing* that caught their attention once they looked away.

This level of detail allows us to break down our behavior into blocks and parts, giving us a language to encode both physical and cognitive actions. That encoding often brings with it understanding and insight—a sort of gears-level awareness of what our brains are doing from moment to moment—and that insight, in turn, gives us a powerful tool for change.



In CFAR parlance, the word “TAP” refers not only to trigger-action *patterns*, but also to trigger-action *plans*—plans which center on taking advantage of these short causal chains. TAPs are simultaneously one of the most basic and most effective tools for tinkering with our own habitual behavior, and since a large percentage of our behavior is habitual, that makes them one of our best tools, period.

Once you’re familiar with the technique, making a TAP is simple, and often takes less than a minute. It’s a quick, four-step process:

- Choose a *goal* (a desired outcome or behavior)
- Identify a relevant *trigger* (something that will happen naturally)
- Decide on an *action* that you want to occur after the trigger
- *Rehearse* the causal link (e.g. with deliberate visualization)

To start with, it’s often easiest to take existing trigger-action patterns and tweak them; sometimes changing one key link in a chain can produce an entirely new behavior. For instance, if you have a goal of exercising more, you might notice that your usual routine has you walking into the building and heading straight for an elevator. You can increase your daily physical activity with a simple TAP—when you *feel the metal of the door handle* (trigger), you’ll *remember to look over at the stairwell* (action).

Why this particular framing, instead of something like “When I go inside the building, I’ll take the stairs”? For starters, the trigger *go inside the building* is a little bit fuzzy. It would probably work for some people, but especially when you’re just starting to learn TAPs, it’s best to err on the side of concreteness and specificity. Feeling the metal of the handle against your palm, or hearing the squeak of the hinge, or noticing the change in temperature as you step inside—these things are clear-cut and unmistakable.

As for the action of *take the stairs*, well—taking the stairs is certainly specific. The problem is that it’s a relatively *large* action, and one that might plausibly require willpower for a lot of people. That doesn’t mean you shouldn’t do it, it just means you might want to leave it out of your TAP. One of the things that makes TAPs so powerful is that, done correctly, they don’t take effort. They build on your ordinary momentum, working by reflex and association, just as you don’t have to try to eat chips when there’s a bowl of them in front of you.

When embarking on any kind of significant behavioral change, it’s easy to get discouraged—to hit a few early failures and feel like abandoning the whole plan. TAPs, as a class, fail in one of two places:

- The trigger fails to fire (i.e. you don’t notice the thing you were hoping to notice)
- You don’t take the action (e.g. because it would take more energy than you have to spare)

Earlier versions of CFAR’s TAPs classes did recommend actions such as “take the stairs,” but following up with participants revealed that the pattern that was actually installed was often something of the form “when I feel the metal of the door handle, I will feel guilty and say mean things to myself the whole time I’m on the elevator.”

By setting an action like “look at the stairs,” you’re both making that second failure mode much less likely (since just looking is a much lighter action), and also avoiding a kind of locking-yourself-in-a-box, predeciding-the-right-strategy kind of mistake. Rather than turning yourself back into a sphex, you are instead summoning sapience—turning off your autopilot for a moment. The TAP is a sort of pop-up dialog box that says “Hi there! This is a chance to remember that you had a goal to take the stairs more often. Would you like to do anything about that?”

In many cases, this is enough—CFAR instructor Duncan Sabien found that the best intervention to cause him to use his expensive elliptical machine was simply to *touch* it, each morning, as he came out of his bedroom. The problem was that Duncan’s ordinary default

habits *didn't include using an elliptical*. By walking down the hall and touching it, he shook himself out of his mindless autopilot, and subsequently spent some of his shower time thinking over the day's schedule and forming intentions about when he could most easily fit in a run.

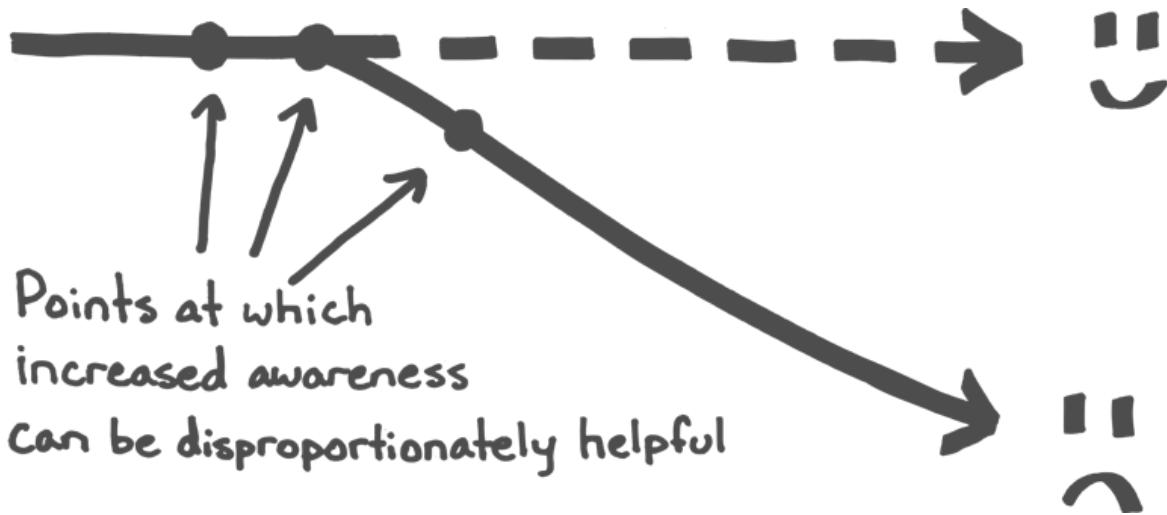
And in cases where this is *not* enough—where your trigger does indeed fire, but after two weeks of *giving yourself the chance to take the stairs*, you discover that you have actually taken yourself up on it zero times—the solution is not TAPs! The problem lies elsewhere—it's not an issue with your autopilot, but rather with your chosen action or some internal conflict or hesitation, and there are other techniques that can be used to illuminate and solve *those* problems.

This isn't to say that the more heavy-duty kind of TAP is *off-limits*. People do indeed get value out of *just making themselves do the thing*. As you grow more comfortable with TAPs, you'll get a better sense of what's viable and sustainable within your own motivational system. As usual, though, we recommend that you build form first—starting off with lightweight practice before putting your skills to a serious test.

- **Goal:** Eat more healthy food  
**TAP:** Grab handle of shopping cart → Ask myself whether this is a “healthy” shopping trip, or a regular one
- **Goal:** Do a better job of showing my friends that I care about them  
**TAP:** Notice that something made me think of a particular friend → Write it down right away on my list of possible birthday gifts
- **Goal:** Remember to bring a book from home  
**TAP:** Drop my keys into the bowl by the door → Pause and think *get the book and put it with my keys*.

---

## Tips for TAPs



Good places to use a TAP:

- Look for *weak links*—places that will help you head off problems before they arise, and recover quickly from the ones you can't prevent.
- Look for *high leverage*—places where you'll have the opportunity to get significant value out of very little effort (e.g. changing shopping habits is much easier than resisting food that's right there in the cupboard).

Selecting the right trigger:

- Look for triggers that are *noticeable* and *concrete* (e.g. "when the microwave beeps" rather than "at dinnertime").
- Whenever possible, choose triggers that are *close* and *relevant* to the behavior you're trying to change (for instance, a toilet flush is closer to the ideal prompt for flossing than a phone alarm would be, even though the phone alarm is highly reliable).
- Don't forget that internal triggers (like specific thoughts and feelings) can be just as good as external ones.

Selecting the right action:

- Choose actions that are *simple* and *atomic*—if you want to do something complicated, consider slowly building up a multi-TAP chain.
- Remember to pick things that you are capable of, and that require as little effort as possible.
- Think concretely and focus on relevance—choose actions that are *actually useful*, not ones that train the wrong skill or seem like you "should" do them.

Making TAPs stick:

- Add new TAPs one or two at a time, rather than in large batches.
- Stay close to your current/natural trigger-action patterns, and make incremental changes.
- Practice mentally rehearsing each new TAP ten times until you've gotten the hang of it (not three or five, but *actually ten*, closing your eyes and going through a complete imaginary run-through each time).
- Write down all of your intended TAPs in one place, and check the list at the end of the week.

Getting better at TAPs generally:

- Practice *noticing* the trigger-action patterns that already exist in your life by looking backwards (e.g. huh, I'm suddenly feeling tired and pessimistic; what happened in the last thirty seconds?). Consider adding an end-of-day review where you think back over the actions you took and the choices you made, and whether there were any where you wish you'd gone down a different branch (and what "going down a different branch" would *actually look like*, in practice).
- Use meta-TAPs, like a TAP to ask yourself if there are useful TAPs to be made in a given situation.
- Steal TAPs from people who are unusually effective or who do not have the problems you have, either by asking them directly what's going on in their thoughts as they do X, Y, and Z, or by modeling their behavior from the outside
- Try gain-pain movies—first imagine some exciting or attractive aspect of the future where you've achieved your goal, and then think about the obstacles that lie between you and that future, and then repeat several times.<sup>[1]</sup>
- Use them frequently! They're good for goals of all sizes, and every CFAR technique can be productively framed in terms of TAPs.

Be patient with yourself

- Remember that the brain responds to reinforcement—if you notice that you missed a trigger, don't punish yourself for the failure; reward yourself for the belated noticing! Over time, your ability to notice will improve, if you don't teach your brain to regret mentioning things after the fact.

---

## Trigger-Action Planning—Further Resources

Logan Strohl's [Intro to Naturalism](#) sequence focuses on building a particular kind of awareness that could be framed as "TAPs for Noticing."

---

Locke and Latham (2002) review decades of research on goal setting and performance. Among their findings: people who set a challenging, specific goal tend to accomplish more than people who set a vague goal (such as "do as much as possible") or those who set an easy goal.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task performance: A 35 year odyssey. *American Psychologist*, 57, 705-717.  
<http://goo.gl/9krv3Q>

---

Gollwitzer and Oettingen (2011) review research on planning and goal pursuit, with an emphasis on implementation intentions (trigger-action plans). They discuss evidence that implementation intentions can be helpful for several subskills of goal pursuit, including getting started, staying on track, overcoming obstacles, and taking advantages of opportunities, as well as cases where implementation intentions are less effective (such as when a person is not very committed to the goal). They also include specific suggestions for how to formulate trigger-action plans.

Gollwitzer, P. M., & Oettingen, G. (2011). Planning promotes goal striving. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (2nd ed., pp. 162-185). New York: Guilford. <http://goo.gl/Dj8NC>

---

A meta-analysis of 94 studies involving 8461 participants found that interventions involving implementation intentions produced an average effect size of  $d = 0.65$  (Gollwitzer & Sheeran, 2006). A similar effect size was found in the 34 studies which involved behavioral change on a personal or health goal ( $d = 0.59$ ).

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, 38, 69-119. <http://goo.gl/AHHUUK>

---

Mental contrasting is the practice of imagining a desired future where a goal has been achieved, and then contrasting it with the current imperfect situation where there are still obstacles to achieving the goal. Oettingen (2012) reviews dozens of studies showing that mental contrasting tends to increase commitment to a goal, including energy and determination, in a way that does not occur in people who merely fantasize about a desired future, or in those who merely think about the current situation and its obstacles (though this effect only occurs when the desired future seems achievable).

Oettingen, G. (2012). Future thought and behavior change. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology*, 23, 1-63. <http://goo.gl/ov54yp>

---

Mental contrasting can be a helpful precursor to the formulation of implementation intentions, since it increases goal commitment and brings to mind obstacles which trigger-action planning can address. Several experiments involving real-world behavior change have used an intervention which combined mental contrasting and implementation intentions, and one such study (Adriaanse et al., 2010) found that this combined intervention was more effective than either one alone at reducing consumption of an unhealthy food.

Adriaanse, M. A., Oettingen, G., Gollwitzer, P. M., Hennes, E. P., de Ridder, D. T. D., & de Witt, J. B. F. (2010). When planning is not enough: Fighting unhealthy snacking habits by mental

contrasting with implementation intentions (MCII). European Journal of Social Psychology, 40, 1277-1293. <http://goo.gl/MCV88X>

---

Psychologist Heidi Grant Halvorson's book Succeed provides a practical summary of research on goal achievement, including an account of implementation intentions and mental contrasting.

Halvorson, Heidi Grant (2010). Succeed: How we can reach our goals.  
<http://www.heidigranthalvorson.com/>

1. ^

First developed by psychologist Gabrielle Oettingen under the name "mental contrasting." Gain-pain movies have been shown to be an excellent companion to TAPs, increasing enthusiasm, emotional resistance, and awareness of goal relevance.

## **Avoid the abbreviation "FLOPs" - use "FLOP" or "FLOP/s" instead**

Especially in discussions about AI, the abbreviation "FLOPs" is being used for both "floating point operations per second" (a measure of computational power) and "floating point operations" (a measure of total computations, and equivalent to the previous term times seconds). This is ambiguous and confusing. For clarity, I propose people avoid this specific abbreviation and instead use the alternatives of "FLOP" (for floating point operations) and "FLOP/s" (for floating point operations per second).

# A Pattern Language For Rationality

There's a lens to looking at the rationality project that I've been finding enlivening recently, and I think it's reached the point where more eyes and hands might be useful, while not being anywhere near complete yet. First, some background.

Christopher Alexander was a designer and architect; his thinking, and focus on patterns in particular, were [influential in programming](#); wikis were first invented to facilitate the collaborative creation and modification of 'patterns' in the style he described. He wrote lots of books, but I'll focus on this trio:

- [The Timeless Way of Building](#) (about the 'quality without a name', and good vs. bad buildings)
- [A Pattern Language: Towns, Buildings, Construction](#) (about the 'patterns' they've identified for physical buildings and towns)
- [The Oregon Experiment](#) (about the University of Oregon, and how it could replace a 'master plan' with principles)

The three books, while each on a different subject or layer of 'design', all work together and depend on each other. The first identifies the target, why it would even be good to pursue, and how you know whether or not you've found it. The second is a detailed description of the patterns they've found useful in approaching the target. The third is what it looks like to organize systems to deliberately and durably organize themselves according to this design.

Just like this approach was profitably translated to programming (and other areas of design), I think it's worth looking at the 'rationality' project as a way to design decisions, habits, and thinking, and attempting to deliberately incorporate Alexander's approach and strategy. This post is an attempt to get started<sup>[1]</sup> in each of the three directions, rather than fully lay them out; depending on how things go, I might turn this post into a sequence / flesh out individual sections.

## The Timeless Way of Living

The Tao that can be told is not the eternal Tao; The name that can be named is not the eternal name.

*[Consider also 'thinking', 'deciding', 'doing', and 'being' in place of 'living', for both this section and the next.]*

Alexander talks about the "Quality Without A Name", <sup>[2]</sup> so-called because no existing English word was a good fit, altho he could point towards it pretty effectively in a few pages; you should read the start of The Timeless Way of Building ([Amazon](#), [pdf](#)) to get his sense of it. I think "equilibrium" comes somewhat close—a design has the Quality Without A Name if all of the forces present, both internal and external, are balanced. This isn't the true name because many equilibria we see don't have the Quality Without A Name, because for them only some of the forces are active in determining the level. A house might be at 'equilibrium' according to the windows and the thermostat, but not according to the human inside who's not happy with the situation and wants to do something about it.

When talking about buildings, he talks about whether they're 'alive' or 'dead'. His overall sense is that many design features are 'obvious' or 'natural'; while there might be lots of detail to the model of how to make things that are alive, most of the ability to detect whether or not a building is alive or dead is already 'baked in' to being a human.<sup>[3]</sup>

I think it's relatively easy to point to good buildings and bad buildings, and somewhat harder to point to good lives and bad lives, mostly because we can walk around the inside of building made by others but not their minds. Nevertheless, it seems possible to collect pictures of what it is that makes life worth living, what properties good decisions have, what virtues we might like to embody. This is, historically, a place I think the rationality project has done pretty well.

You may try to name the highest principle with names such as "the map that reflects the territory" or "experience of success and failure" or "Bayesian decision theory." But perhaps you describe incorrectly the nameless virtue. How will you discover your mistake? Not by comparing your description to itself, but by comparing it to that which you did not name.

One interesting thing about practical Bayesianism (rather than superhuman Bayesianism) is that among the normal, explicitly considered hypotheses lurks a monster: "all other explanations." What probability should it have? How should one update on it when observing a new piece of evidence? For the other explanations, one can follow the math as written; that one is precisely a placeholder for that where one cannot do the math, and must guess. Having a token in your language for "that which is not expressed in the language" helps keep you grounded in the reality larger than your mind, rather than trapped in your imagination.

## A Pattern Language for Living

One fascinating thing about A Pattern Language (for buildings) is that it is highly opinionated, and only contains what you *should* do. But it's often just as useful to have a list of [anti-patterns](#), that is, common models or designs that are recommended *against*. Then, as you're designing, you can notice that you're about to make a mistake, and replace the anti-pattern with an appropriate pattern.<sup>[4]</sup>

According to me, the rationality project that grew out of Overcoming Bias and LessWrong has a tilted focus in favor of anti-patterns, because of the heavy early influence of the [heuristics and biases](#) literature. As well as [37 ways words can be wrong](#), what are the ways words can be right? LW does have a lot of patterns, several pointers to deep generators of patterns, and some focus on the Quality Without A Name, but it seems to me like there's significant room for improvement here, and I'm interested in working to map out the space and collect patterns.

In A Timeless Way of Building, Alexander describes the patterns of a person's life:

*A building or a town is given its character, essentially, by those events which keep on happening there most often.*

A field of grass is given its character, essentially, by those events which happen over and over again--millions upon millions of times. The germination of the grass seed, the blowing wind, the flowering of the grass, the movement of the worms, the hatching of the insects...

A car is given its character by the events which keep on happening there--the rolling of the wheels, the movement of the pistons in the cylinders, the limited to and fro of the steering wheel and axle, as the car changes direction.

A family is given its character by the particular events which keep on happening there--the small affections, kisses, breakfast, the particular kinds of arguments which keep on happening, the way these arguments resolve themselves, the idiosyncrasies of people, both together and alone, which make us love them...

*And just the same is true in any person's individual life.*

If I consider my life honestly, I see that it is governed by a certain very small number of patterns of events which I take part in over and over again.

Being in bed, having a shower, having breakfast in the kitchen, sitting in my study writing, walking in the garden, cooking and eating our common lunch at my office with my friends, going to the movies, taking my family to eat at a restaurant, having a drink at a friend's house, driving on the freeway, going to bed again. There are a few more.

There are surprisingly few of these patterns of events in any one person's way of life, perhaps no more than a dozen. Look at your own life and you will find the same. It is shocking at first, to see that there are so few patterns of events open to me.

Not that I want more of them. But when I see how very few of them there are, I begin to understand what huge effect these few patterns have on my life, on my capacity to live. If these few patterns are good for me, I can live well. If they are bad for me, I can't.

*Of course, the standard patterns of events vary very much from person to person, and from culture to culture.*

For a teenage boy, at high school in Los Angeles, his situations include hanging out in the corridor with other boys; watching television; sitting in a car with his girlfriend at a drive-in restaurant eating coke and hamburgers. For an old woman, in a European mountain village, her situations include scrubbing her front doorstep, lighting a candle in the local church, stopping at the market to buy fresh vegetables, walking five miles across the mountains to visit her grandson.

These are the patterns on the scale of 'events', but one of the most charming features of A Pattern Language is the breadth of scales that it considers. Ordered from largest to smallest, the first entry in the book is Alexander's sense of how the whole Earth should be organized (a global government made of independent regions, each containing 2-10 million people, so they can be small enough to be tolerable for the people inside them) and the final three (to give some sense of variety) are what sorts of chairs to put in a space, how to light them, and how to decorate one's walls. The patterns nest within each other, as each larger pattern depends on smaller patterns beneath it (until the simplest, smallest patterns at the end).

Similarly, rationality patterns extend from the great [Neo-Enlightenment project](#) all the way down to habits that fire on the [5-Second Level](#). Also, in a manner which makes them more difficult to organize, the patterns scale both in time and number of people (whereas with buildings, it is practical to simply organize the patterns by volume).

Historically, I think LessWrong has done great at collecting antipatterns, and only well at collecting patterns, and somewhat poorly at organizing them all into a common reference work. (Read The Sequences, we used to say, but much of what has happened since then is scattered instead of curated and carefully arranged. My guess is a solid approach to take here is something like "go thru the Sequences, the CFAR material, the library of Scott Alexandria, and so on, pulling out the patterns and trying to arrange them into a pattern language, identifying the connections between them and filling the resulting gaps. If you're interested in helping with this, comment below or send me a PM.)

I also have some sense that this sort of 'life-design' is a key component of applying rationality consistently and well; some large part of the rationality project's benefits for individuals have been, I think, from being exercising this sort of intentionality and deliberate thought about parts of their life where people often just operate on autopilot. Having a systematic way to look at this, instead of just noticing things as they come up, seems quite valuable (so long as one makes changes in an organic, sustainable way which is pointed towards the Quality Without A Name; it's generally a mistake to give edit access to yourself to your less grounded parts).

[Incidentally, when I did this I got 25 patterns for myself, with 'standing at my computer' split into a further ten patterns, tho many of those patterns were things that happen on a ~weekly cadence instead of a daily one, which increases the number substantially.]

## The Oregon Experiment For Time

The University of Oregon wanted a 'master plan' to guide its growth for the next few decades, and reached out to Christopher Alexander; he responded with a bit of a rant on why master plans were terrible, and what they should do instead.[\[5\]](#)

Their goal, as Alexander saw it, was the same timeless way of building that leads to the organic growth of living towns and cities. The challenge was how to do it when there was a monopoly funder and decider. How do you get the users to change their location to fit their needs, while being harmonious with the whole, while everything needs to be signed off on by the center?

This is especially relevant for me because I'm something of a free agent, at the moment. I don't see any plans that obviously have good impact on x-risk, I'm not interested in pretending any plans are more attractive than they actually seem, and yet I still find it fun to help people out and think about these problems (and think the 'option value' of staying involved is quite high). I also am financially secure enough that I don't need a job, which makes a freelance style more attractive, whereas if there were jobs I thought were obviously worth taking then the benefits of specialization would push hard against being freelance. But this feels like the same sort of situation-a monopoly funder/decider which is nevertheless trying to do things that are good for its users and generate the Quality Without A Name.

I think this is probably still relevant even for people in traditional jobs or educational situations; after all, it's still the case that you're the monopoly funder and decider for what to do with your time and energy, and it's worth looking into ways to incorporate these principles (or *deliberately* find competing principles, rather than aping the behavior of those around you or following whatever instructions you receive[\[6\]](#)).

So let's take a look at Alexander's principles and see if they can be adapted from "university space use plan" to "individual time use plan".

1. Organic Order: use a gradual process to construct the campus, rather than visualizing it all at once.
2. Participation: users make decisions about what and how to build.
3. Piecemeal Growth: weight budget overwhelmingly towards small projects (in terms of number; be roughly equal in terms of dollars spent.)
4. Patterns: control and harmony should be exerted mostly by deciding the 'language' of the campus, i.e. what patterns are appropriate; all users should build things according to the relevant patterns.
5. Diagnosis: once a year, the design committee goes thru the campus and determines which parts are currently 'alive' and 'dead'.
6. Coordination. Funding process handles a stream of proposals that are put forth by users.

It seems to me like Organic Order, Participation, Piecemeal Growth, and Coordination translate straightforwardly. Organic Order, for example, suggests that I should only be setting vague themes with long-term planning, rather than trying to schedule out all of my hours weeks, months, or years ahead of time.

Patterns and Diagnosis both feel like they would translate straightforwardly, if the other components were present; Patterns, for example, seems like it could just obviously work if I had the pattern language to work off of, and Diagnosis seems like it relies on a sense of 'alive' and 'dead' which I don't have a good handle on yet (but could probably wing it-The Timeless Way of Building is pretty confident that people 'already have' the ability to tell alive and dead spaces apart).

Some seem like they need rescaling. Piecemeal Growth has budget categories that make sense for buildings / renovation projects in a university, but what's the corresponding categories that make sense for spending time?<sup>[7]</sup> Note that if you have a sense that projects should be distributed such that each category has equal spending, where you draw the category boundaries now has decision-relevant effects.

## Organic Order

As I go thru life, my local information, wants, and capacities will vary substantially; the planning process should take that into account, and not make plans that it expects to become stale.

## Participation

Lots of what I do is for other people. They know things I don't, and want things I don't yet imagine; the planning process should take that into account, and try to bring me into contact with my impact and what it could be.

For buildings, the principle is somewhat obvious; people have lots of implicit knowledge of how their space should be arranged, and should tell the design to the architect, rather than the other way around, or trying to verbalize their requirements. For time, it seems more about establishing clear requirements, and trying to put those requirements in the hands of the beneficiaries.

For everything that I spend serious effort on, “knowing where the score is” seems like a basic thing to keep in mind. If I’m doing something “for my husband”, it really should be clear to me the score is in terms of “how much he valued it” and the score is in him; and if instead I’m doing something “so that I feel like I’m a good husband”, then the score is in me. If I’m doing something to help out people in my house, knowing what they actually want means it’s more likely that I’ll do something useful to them.

Speaking about the EA / longtermism space more broadly, I really wish we were better at using prices to convey information. If you’re working a job in the selfish economy, the price your employers are willing to pay you is a pretty good measure of “how much good you’re doing for them”, whereas if you work at a nonprofit, you just have a sense of ‘doing good’, and there’s not much in the way of joint calculation of how much good you’re doing. EA is, in large part, an attempt to fix that—but I think it’s still missing in lots of places locally.<sup>[8]</sup>

## Piecemeal Growth

Part of contact with reality is learning which things succeed and which things fail, which things are effortless and which are effortful.

Alexander argues that often, people consider equal numbers of projects in various size categories, but this means the result is heavily skewed towards the larger projects. Instead, have a process that allocates equal budgets to the various size categories, leading to most projects being small.

I think I’ve often made a mistake where I turn small ideas into big ideas, and thus big obligations. I started writing a fanfiction whose core idea was a six chapter cycle; it turned into an epic that I never finished. A friend who is a successful artist would start her day by sketching out lots of quick things, and then would pick from the pile which one to turn into a detailed painting. It seems likely that her paintings were unusually good because she had that filtering step, where ideas could be tried in a low-stakes environment and then only the best ones got significant effort.

As well, there’s a common pattern in industrial engineering of “WIP as waste”; work-in-progress is not providing value to anyone yet and is causing costs while it exists, and so a system that manages the same output with less WIP is better. If rather than working on five posts in parallel and publishing them all at once at the end, I work on the posts serially, then the first post released can be released much earlier (and be providing benefits while the others are being written, and they can change in reaction to its feedback, and so on).

If I want to spend ten hours on a goal, quite probably the right way to do that is to do five small 1-hour projects, then spend another five hours on the one that seems the most promising. That way, if it turns out that there are dead ends along half of the approaches I thought of, I spend most of my time on an approach that turned out to be promising. Otherwise, the method will end up heavily associated with the goal (consider the difference between when I sit down to spend 10 hours “writing a blog post about X” and when I spend 10 hours “on X”), and I might just give up or waste time for half of my goals.

A less-obvious benefit of this approach is that, from the point of view of the university campus, the ‘already built’ areas of the campus need love too (in order to stay ‘alive’,

in good repair, in touch with their user's interests, and so on), but what they need will primarily be small projects. A committee which only allocates money to large fancy new buildings will see all of its old buildings deteriorate, whereas one which reserves a large fraction of its funds for small projects will see many of the old buildings well-taken care of.

One argument against piecemeal growth for individuals (instead of organizations) is that a university can spend its budget on many projects in parallel, whereas an individual has to do things one after another. Not being able to spend 20% of my focus on five small projects at the same time, whereas I could spend 20% of my budget on five small projects at the same time, means that smaller projects are more costly because of switching costs and delays to other work.

## Patterns

When choosing to do things, have a sense of 'what they are', what subcomponents they need, and what supercomponents they support. Have a sense of both what to do and what not to do.

It also seems like having a shared language for variation in buildings aids in specialization--you can easily talk about the differences between a park, a house, and a workshop, and find the building you're looking for. Similarly, being able to advertise what projects you're good at, enjoy doing, and hook into your larger goals makes it easier to coordinate.

I'm also growing to appreciate 'seasonality' more, in a way that patterns can identify and support. One might naively think that if doing X is better than doing Y, I should just schedule myself to only ever do X and never do Y, but it's rare for there to be true dominance. A pattern like "Focus in the mornings, connection in the afternoons, freedom in the evenings" ensures that no nutrient goes uneaten for too long (tho what the actual ratio should be is a fact about your situation, instead of about what happens when you divide one by three).

## Diagnosis

See what is actually true on a pace that allows it to influence what you do; don't let corners fester unconsidered.

For the university, this looks like an annual checkup, where the planning committee physically tours the campus, simply noting 'alive' or 'dead' for each space. While you can do a quick tour thru a physical space, doing tours thru time is less easy. My guess is that this category should be 'retrospectives' in general, dealing with properties on the relevant timeframe. A daily reflection asks about the aliveness of hours and tasks; a weekly reflection asks about the aliveness of days and projects; a monthly reflection asks about the aliveness of weeks and themes; an annual reflection asks about the aliveness of months, life goals, and principles.

## Coordination

For a campus, it's obvious who the users are and why they have information that the central committee lacks. If the music school is poorly designed for the students of the

school, how will the campus architects know, unless they have a way for the students to tell them, and then participate in the design to fix it?

For me, I think people approach me infrequently with things that I could do that would help them. There have been times when I've been busy and had to say no a lot, but my typical response these days is 'yes', suggesting that I'm getting too few requests.

The other side of this is a general approach to planning; all projects, regardless of size, have to talk about their goals and what patterns they're a part of, with larger projects having more preparatory work than smaller ones, and considered in larger budget meetings (as opposed to handled by a more local budget process). Coordination is about the process by which the pieces fit together and change behavior.

## Other Principles?

At the moment, my plan is to start off with something that's a near-copy of The Oregon Experiment, and then adjust it in contact with reality. But perhaps you can see now some piece that I'm missing, or some piece that I'm holding on to despite its irrelevance; I'd love comments if so.

1. ^

Well, for me to get started; I think a lot of this stuff is already 'in the water' in some important sense, but like scholarship being 'in the water' is very different from carefully reading books and writing reviews of them, I think having a loose sense that 'patterns are neat' and a big book which links all of the patterns together in a sensible way are very different, and I want to get the latter.

2. ^

I saw someone writing about Christopher Alexander abbreviate it as "QWAN", which 1) makes sense and 2) is easy to pronounce but 3) feels weird or ironic in some way. I went back and forth on whether to use it in the post, and decided not to (while naming the dilemma).

3. ^

His design sense is very human-centric, which I weirdly want to describe as being 'alien to me'? Like, his sense of what makes a building good is very tied to the human perceptual system instead of simpler metrics that are more understandable to an engineer. "A human should have an easy time making sense of this", he says, even tho that's much harder to measure than things like the materials cost. I think he's also about ten percent too certain that all humans share the same sensemaking system, instead of many of these things actually being matters of taste.

4. ^

Often, I wish that there were 'neutral' categories with 'good' and 'bad' subcategories, as it seems nonsensical to claim that anti-patterns aren't patterns in the generic, neutral sense; however, I'll use the standard language

here, where patterns are definitionally ‘recommended’ and anti-patterns are definitionally ‘anti-recommended’.

5. ^

[As far as I can tell](#), they implemented Alexander's suggestions, but I couldn't easily find much of a retrospective, which makes this a bit disappointing as an 'experiment' for us as observers. They also have a set of 12 principles for plans which are quite different from the 6 principles underlying the creation of the process, which seems reasonable.

6. ^

For example, see [Half-assing it with everything you've got](#) as an example of 'looking behind the curtain' and figuring out what principle is the right call.

7. ^

I'm also thinking here primarily about “high-energy” time; it makes sense to view my morning pomodoros as ‘improvement dollars’ and less sense to view my evening wind-down activities in the same way. But it feels like Piecemeal Growth has similar things to say within each category, even if it can't translate currencies across categories.

8. ^

If you could work for either GiveWell or 80k or MIRI, do you have a number that you can optimize that represents how much good you're doing for them? If there's not, how could there be shared epistemic state on that?