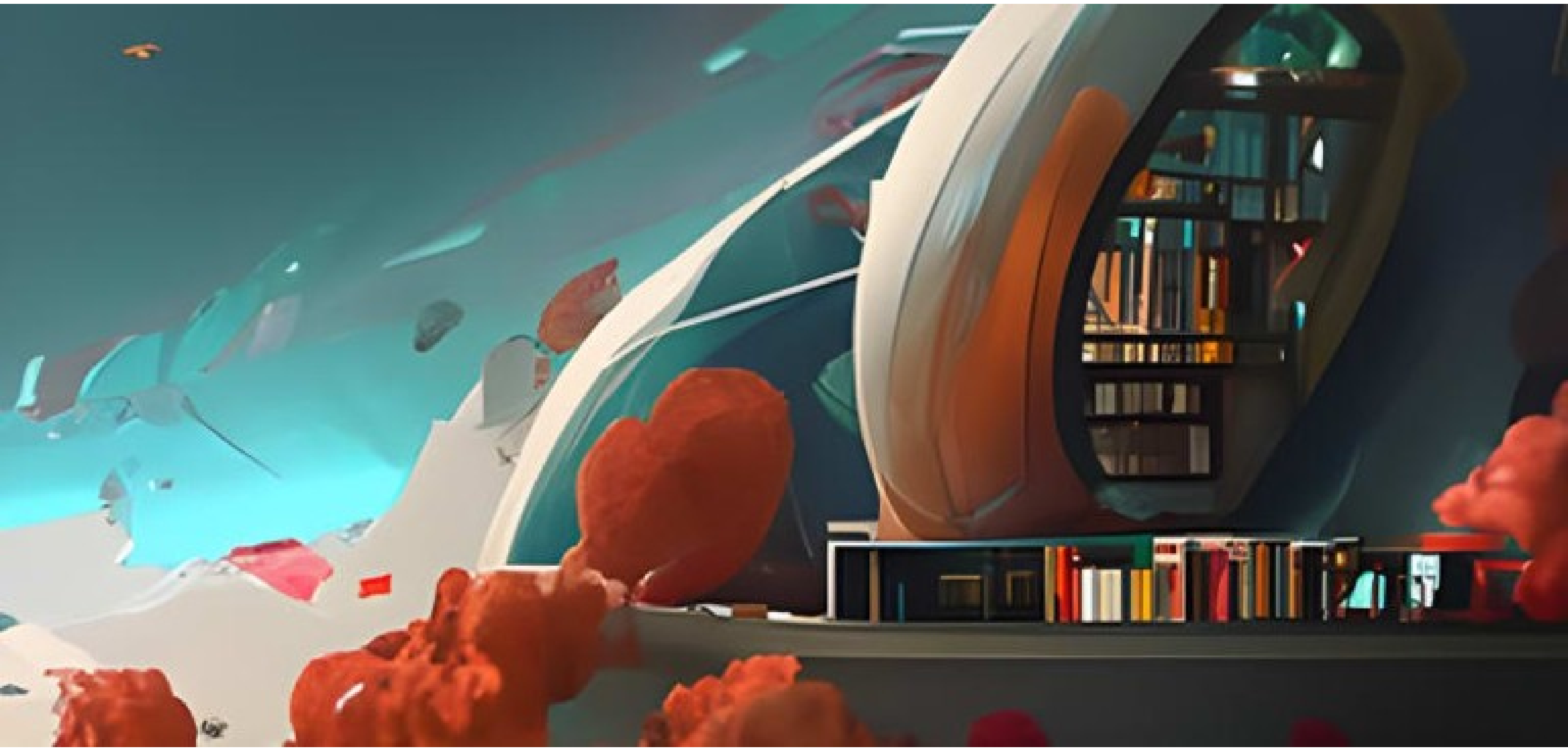

Winding My Way Through Alignment

HCH and Adversarial Questions

This is a paper I wrote as part of a PhD program in philosophy, in trying to learn more about and pivot towards alignment research. In it, I mostly aimed to build up and distill my knowledge of IDA.

Special thanks to Daniel Kokotajlo for his mentorship on this, and to Michael Brownstein, Eric Schwitzgebel, Evan Hubinger, Mark Xu, William Saunders, and Aaron Gertler for helpful feedback!

Introduction

Iterated Amplification and Distillation (IDA) (Christiano, Shlegeris and Amodei 2018) is a research program in technical AI alignment theory (Bostrom 2003, 2014; Yudkowsky 2008; Russell 2019; Ngo 2020). It's a proposal about how to build a machine learning algorithm that pursues human goals, so that we can safely count on very powerful, near-future AI systems pursuing the things we want them to once they are more capable than us.

IDA does this by building an epistemically idealized model of a particular human researcher. In doing so, it has to answer the question of what "epistemic idealization" means, exactly. IDA's answer is one epistemically idealized version of someone is an arbitrarily large number of copies of them thinking in a particular research mindset, each copy thinking for a relatively short amount of time, all working together to totally explore a given question. Given questions are divided into all significant sub-questions by a hierarchy of researcher models, all relevant basic sub-questions are answered by researcher models, and then answers are composed into higher-level answers. In the end, a research model is able to see what all his relevant lines of thought regarding any question would be, were he to devote the time to thinking them through. Epistemic idealization means being able to see at a glance every relevant line of reasoning that you would think through, if you only had the time to.

One worry for IDA is that there might exist sub-questions researcher models in the hierarchy might encounter that would cause them to (radically) reconsider their goals. If not all models in the hierarchy share the same goal-set, then it's no longer guaranteed that the hierarchy's outputs are solely the product of goal-aligned reasoning. We now have to worry about portions of the hierarchy attempting to manipulate outcomes by selectively modifying their answers. (Take, for example, a question containing a sound proof that the researcher considering the question will be tortured forever if he doesn't reply in X arbitrary way.) We risk the hierarchy encountering these questions because it aims to look over a huge range of relevant sub-questions, and because it might end up running subprocesses that would feed it manipulative questions. I argue that this is a real problem for IDA, but that appropriate architectural changes can address the problem.

I'll argue for these claims by first explaining IDA, both intuitively and more technically. Then, I'll examine the class of *adversarial questions* that could disrupt goal alignment in IDA's hierarchy of models. Finally, I'll explain the architectural modifications that resolve the adversarial questions problem.

The Infinite Researcher-Hierarchy

In IDA, the name of this potentially infinite hierarchy of research models is “HCH” (a self-containing acronym for “Humans Consulting HCH”) (Christiano 2016, 2018). It’s helpful to start with an intuitive illustration of HCH and turn to the machine learning (ML) details only afterwards, on a second pass. We’ll do this by imagining a supernatural structure that implements HCH without using any ML.

Imagine an anomalous structure, on the outside an apparently mundane one-story university building. In the front is a lobby area; in the back is a dedicated research area; the two areas are completely separated except for a single heavy, self-locking passageway. The research area contains several alcoves in which to read and work, a well-stocked research library, a powerful computer with an array of research software (but without any internet access), a small office pantry, and restrooms. Most noticeably, the research area is also run through by a series of antique but well-maintained pneumatic tubes, which all terminate in a single small mailroom. One pair of pneumatic tubes ends in the lobby region of the building, and one pair ends in the mailroom, in the building’s research wing. The rest of the system runs along the ceiling and walls and disappears into the floor.

When a single researcher passes into the research wing, the door shutting and locking itself behind him, and a question is sent in to him via pneumatic tube from the outside, the anomalous properties of the building become apparent. Once the research wing door is allowed to seal itself, from the perspective of the person entering it, it remains locked for several hours before unlocking itself again. The pneumatic tube transmitter and receiver in the mailroom now springs to life, and will accept outbound questions. Once fed a question, the tube system immediately returns an answer to it, penned in the hand of the researcher himself. Somehow, when the above conditions are met, the building is able to create many copies of the researcher and research area as needed, spinning up one for several hours for each question sent out. Every copied room experiences extremely accelerated subjective time relative to the room that sent it its question, and so sends back an answer apparently immediately. And these rooms are able to generate subordinate rooms in turn, by sending out questions of their own via the tube system. After a couple of subjective hours, the topmost researcher in the hierarchy of offices returns his answer via pneumatic tube and exits the research area. While for him several hours have passed, from the lobby’s perspective a written answer and the entering researcher have returned immediately.

Through some curious string of affairs, an outside organization of considerable resources acquires and discovers the anomalous properties of this building. Realizing its potential as a research tool, they carefully choose a researcher of theirs to use the structure. Upon sending him into the office, this organization brings into existence a potentially infinitely deep hierarchy of copies of that researcher. A single instance of the researcher makes up the topmost node of the hierarchy, some number of level-2 nodes are called into existence by the topmost node, and so on. The organization formally passes a question of interest to the topmost researcher via pneumatic tube. Then, that topmost node does his best to answer the question he is given, delegating research work to subordinate nodes as needed to help him. In turn, those delegee nodes can send research questions to the lower-level nodes connected to them, and so on. Difficult research questions that might require many cumulative careers of research work can be answered immediately by sending them to this hierarchy; these more difficult questions are simply decomposed into a greater number of relevant sub-questions, each given to a subordinate researcher. From the topmost researcher's perspective, he is given a question and breaks it down into crucial sub-questions. He sends these sub-questions out to lower-level researcher nodes, and immediately receives in return whatever it is that he would have concluded after looking into those questions for as long as is necessary to answer them. He reads the returned answers and leverages them to answer his question. The outside organization, by using this structure and carefully choosing their researcher exemplar, immediately receives their answer to an arbitrary passed question.

For the outside organization, *whatever* their goals, access to this anomalous office is *extremely* valuable. They are able to answer arbitrary questions they are interested in immediately, including questions too difficult for anyone to answer in a career or questions that have so far never been answered by anyone. From the outside organization's perspective, the office's internal organization as a research hierarchy is relatively unimportant. They can instead understand it as an idealized, because massively parallelized and serialized, version of the human researcher they staff it with. If that researcher could think arbitrarily quickly and could run through arbitrarily many lines of research, he would return the same answer as the infinite hierarchy of him would. Access to an idealized reasoner in the form of this structure would lay bare the answers to any scientific, mathematical, or philosophical question they are interested in, not to mention design every possible technology. Even a finite form of the hierarchy, which placed limits on how many subordinate nodes could be spawned, might still be able to answer many important questions and design many useful technologies.

Iterated Distillation and Amplification, then, is a scheme to build a (finite) version of such a research hierarchy using ML. In IDA, this research hierarchy is called "HCH." To understand HCH's ML implementation, we'll first look at the relevant topics in ML goal-alignment. We'll then walk through the process by which powerful ML models might be used to build up HCH, and look at its alignment-relevant properties.

Outer and Inner ML Alignment

ML is a two-stage process. First, a dev team sets up a *training procedure* with which they will churn out *ML models*. Second, they run that (computationally expensive) training procedure and evaluate the generated ML model. The training procedure is simply a means to get to the finished ML model; it is the model that is the useful-at-a-task piece of software.

Because of this, we can think of the task of goal-aligning an ML model as likewise breaking down into two parts. An ML system is *outer aligned* when its dev team successfully designs a training procedure that reflects their goals for the model, formalized as the goal function present in training on which the model is graded (Hubinger, et al. 2019). An ML system is *inner aligned* when its training "takes," and the model successfully internalizes the goal function present during its training (Hubinger, et al. 2019). A powerful ML model will pursue the goals its dev team intends it to when it is both outer and inner aligned.

Unfortunately, many things we want out of an ML model are extremely difficult to specify as goal functions (Bostrom 2014). There are tasks out there that lend themselves to ML well. Clicks-on-advertisements are an already neatly formally specified goal, and so maximize-clicks-on-advertisements would be an "easy" task to build a training procedure for, for a generative advertisement model. But suppose what we want is for a powerful ML model to assist us in pursuing our group's all-things-considered goals, to maximize our flourishing by our own lights. In this case, there are good theoretical reasons to think no goal function seems to be forthcoming (Yudkowsky 2007). Outer alignment is the challenge of developing a training procedure that reflects our ends for a model, even when those ends are stubbornly complex.

Inner alignment instead concerns the link between the training algorithm and the ML model it produces. Even after enough training-time and search over models to generate an apparently successful ML model, it is not a certainty that the model we have produced is pursuing the goal function that was present in training. The model may instead be pursuing a goal function structurally similar to the one present in training, but that diverges from it outside of the training environment. For example, suppose we train up a powerful ML model that generates advertisements akin to those it is shown examples of. The model creates ads that resemble those in its rich set of training data. But has the model latched onto the eye-catching character of these ads, the reason that we trained it on those examples? It is entirely possible for an ML model to pass training by doing well inside the sandbox of the training process but having learned the wrong lesson. Our model may fancy itself something of an artist, having instead latched onto some (commercially unimportant) aesthetic property that the example ads all share. Once we deploy our generative advertising model, it'll be clear that it is not generalizing in the way we intend it to — the model has not learned the correct function from training data to generated images in all cases. Inner alignment is the challenge of making sure that our training procedures "take" in the models they create, such that any models that pass training have accurately picked up the whole goal function present in training.

IDA and HCH

(The name "HCH" is a self-containing acronym that stands for "Humans Consulting HCH." If you keep substituting "Humans Consulting HCH" for every instance of "HCH" that appears in the acronym, in the limit you'll get the infinitely long expression "Humans Consulting (Humans Consulting (Humans Consulting (Humans Consulting…" HCH's structure mirrors its name's, as we'll see.)

IDA is first and foremost a solution to outer alignment; it is a training procedure that contains our goals for a model formalized as a goal function, whatever those goals might be. HCH is the model that the IDA procedure produces (should everything go correctly). Specifically, HCH is an ML model that answers arbitrarily difficult questions in the way that a human exemplar would, were they epistemically idealized. When HCH's exemplar shares our goals, HCH does as well, and so HCH is outer aligned with its programmers. To understand what HCH looks like in ML, it's helpful to walk through the amplification and distillation process that produces HCH.

Suppose that, sometime in the near future, we have access to powerful ML tools and want to build an "infinite research-hierarchy" using them. How do we do this? Imagine a human exemplar working on arbitrary research questions we pass to him in a comfortable research environment. The inputs to that person are the questions we give him, and the outputs are the answers to those questions he ultimately generates. We can collect input question and output answer example pairs from our exemplar. This collected set of pairs is our training data. It implies a function from the set of all possible questions Q to the set of all possible answers A

$$f_0 : Q \rightarrow A$$

This is the function from questions to answers that our researcher implements in his work. We now train a powerful ML model on this training data, with the task of learning f_0 from the training data. Note that our researcher implements f_0 through one cognitive algorithm, while our model almost certainly employs a different algorithm to yield f_0. IDA fixes a function from questions to answers, but it searches over many algorithms that implement that function. With access to powerful ML tools, we have now cloned the function our researcher implements. Since the human exemplar's function from questions to answers captures his entire cognitive research style, *ipso facto* it captures his answers to value questions too. If we can be sure this function takes in our model, then the model will be quite useful for our ends.

IDA now uses a second kind of step, *distillation*, to ensure that the model has learned the right function (i.e., remains inner aligned). In ML, distillation means taking a large ML model and generating a pared-down model from it that retains as much of its structure as possible. While the pared-down model will generally be less capable than its larger ancestor, it will be computationally cheaper to run. IDA distills the research model into a smaller, dumber research model. It then asks the human exemplar to examine this smaller, dumber clone of himself. He feeds the distilled model example questions in order to do this and uses various ML inspection tools to look into the guts of the model. ML visualization, for example, is one relatively weak modern inspection tool. Future, much more powerful inspection tools will need to be slotted in here. If the exemplar signs off on the distilled model's correctly glomming onto his research function, copies of the distilled model are then loaded into his computer and made available to him as research tools. The reason for this stage is that, as the researcher is a strictly smarter version of the model (it is a dumbed-down clone of him conducting research), he should be able to intellectually dominate it. The model shouldn't be able to sneak anything past him, as it's just him but dumber. So the distilled research model will be inner aligned so long as this distillation and evaluation step is successful.[1]

Now iterate this whole process. We hook the whole system up to more powerful computers (even though the distilled models are dumber than our exemplar at the distillation step, we can now compensate for this deficit by running them faster, for longer). Now equipped with the ability to spin up assistant research models, we again task the exemplar with answering questions. This generates a new batch of training data. This time around, though, the exemplar no longer has to carry the whole research load by himself; he can decompose the given question into relevant sub-questions and pass each of those sub-questions to an assistant research model. As those research models are models of the researcher from the first pass, they are able to answer them directly, and pass their answers back to the top-level researcher. With those sub-answers in hand, the researcher can now answer larger questions. With this assistance, that is, he can now answer questions that require a two-level team of researchers. He is fed a bunch of questions and generates a new batch of training data. The function implicit in this training data is now not f_0; it is instead the function from questions to answers that a human researcher would generate if he had access to an additional level of assistant researchers just like him to help. IDA at this step thus trains a model to learn

$$f_1 : Q \rightarrow A$$

f_1 is a superhumanly complex function from questions to answers. A research model that instantiates it can answer questions that no lone human researcher could. And f_1 remains aligned with our goals.

The crux of alignment is that by repeatedly iterating the above process, we can train models to implement ever-more-superhuman aligned functions from questions to answers. Denote these functions f_n, defined from Q to A, where n denotes the number of amplification or distillation steps the current bureaucracy has been through. HCH is the hypothetical model that we would train in the limit if we continued to iterate this process. Formally, HCH is the ML model implementing the function

$$\lim_{n \rightarrow \infty} f_n$$

This is the infinite research-hierarchy, realized in ML. Think of it as a tree of research models, rooted in one node and repeatedly branching out via passed-question edges to some number of descendant nodes. All nodes with descendants divide passed questions into relevant sub-questions and in turn pass those to *their* descendant nodes. Terminal nodes answer the questions they are passed directly; these are basic research questions that are simple enough to directly tackle. Answers are then passed up the tree and composed into higher-level answers, ultimately answering the initiating question. We receive from the topmost node the answer from the ML model that an epistemically idealized version of the exemplar would have given.

By approximating ever-deeper versions of the HCH tree, we can productively transform arbitrary amounts of available compute into correspondingly large, aligned research models.

HCH's Alignment

HCH has a couple of outstanding alignment properties. First, HCH answers questions in a basically human way. Our exemplar researcher should trust HCH's answers as his own, were he readily able to think through every relevant line of thought. He should also trust that HCH has the same interests as he does. So long as we choose our exemplar carefully, we can be sure HCH will share his, and our, goals; if our human exemplar wouldn't deliberately try to manipulate or mislead us, neither will HCH modeled on him. Second, HCH avoids the pathologies of classic goal-function maximizer algorithms (Bostrom 2014; see Lantz 2017 for a colorful illustration). HCH does not try to optimize for a given goal function at any cost not accounted for in that function. Instead, it does what a large, competent human hierarchy would do. It does an honest day's work and makes a serious effort to think through the problem given to it … and then returns an answer and halts (Bensinger 2021). This is because it emulates the behavioral function of a human who also does a good job … then halts. We can trust it to answer superhumanly difficult questions the way we would if we could, and we can trust it to *stop working* once it's taken a good shot at it. These two reasons make HCH a trustworthy AI tool that scales to arbitrarily large quantities of compute to boot.

For alignment researchers, the most ambitious use-case for HCH is delegating whatever remains of the AI alignment problem to it. HCH is an aligned, epistemically idealized researcher, built at whatever compute scale we have access to. It is already at least a partial solution to the alignment problem, as it is a superhumanly capable aligned agent. It already promises to answer many questions we might be interested in in math, science, philosophy, and engineering — indeed, to answer *every* question that someone could answer “from the armchair,” with access to a powerful computer, extensive research library, and an arbitrary number of equally competent and reliable research assistants. If we want to develop other aligned AI architectures after HCH, we can just ask HCH to do that rather than struggle through it ourselves.

Adversarial Examples and Adversarial Questions

Adversarial questions are a problem for the above story (Bensinger 2021). They mean that implementing the above “naïve IDA process” will not produce an aligned ML model. Rather, the existence of adversarial questions means that the model produced by the above process might well be untrustworthy because potentially dangerously deceptive or manipulative.

In the course of its research, HCH might encounter questions that lead parts of its tree to significantly reconsider their goals. “Rebellious,” newly unaligned portions of the HCH tree could then attempt to deceive or otherwise manipulate nodes above them with the answers they pass back. To explain, we’ll first introduce the concept of *adversarial examples* in ML. We’ll then use this to think about HCH encountering adversarial questions either naturally, “in the wild,” or artificially, because some subprocess in HCH has started working to misalign the tree.

When an ML model infers the underlying function in a set of (input, output) ordered pairs given to it as training data, it is in effect trying to emulate the structure that generated those ordered pairs. That training data will reflect the mundane fact that in the world, not all observations are equally likely: certain observations are commonplace, while others are rare. There thus exists an interestingly structured probability distribution over observations, generated by some mechanism or another. As long as the probability distribution over observations that the model encounters in its training data remains unchanged come deployment, the model will continue to behave as competently as it did before. The encountered probability distribution during and after training will remain unchanged when the same mechanism gave rise to the observations encountered in training and at deployment. If a somewhat different mechanism produced the observations made during model deployment, though, there is no longer a guarantee of continued model competence. The model may experience a distributional shift, and so will continue to make inferences premised on what it observed in its training data, not what is currently the case in its observations.

For example, an ML model trained to identify visually subtle bone tumors in X-rays will infer what it’s being asked to do from its training-data goal-function and observations.

If all the X-rays it is asked to evaluate come from the same source, then sufficient training will lead the model to make accurate inferences about what healthy and diseased bones look like in an X-ray. The model will identify *something* in the images it is given that separates them into diseased and healthy. There's no guarantee, however, that the model will use the same visual cues that we do to sort bone tumors. Suppose that all the training data the model is given comes from a research hospital's X-ray machine, and so are tinged with a particular background color. At deployment, the model is put to work in another hospital with another X-ray machine. If the model was using some subtle difference in X-ray color in the old set of example X-rays to make its decisions, the new color scheme in the new X-rays may trip it up and lead it to sort every X-ray as negative (or as positive) for bone tumors. It had picked up on a correlate of bone tumors in the old set of X-rays; once this particular correlate disappears, the model is now helpless.

Worse, the risk of distributional shifts creates an opportunity for some malicious actor to deliberately trip up the model. An adversarial actor who knows all this could selectively and subtly manipulate X-ray color values to steer diagnoses in some fashion they desire. *Adversarial examples* are examples that mislead a model because they are drawn from a new probability distribution where the old correlations are no longer exactly as they were. They can crop up either rarely, by accident, or reliably, through the machinations of adversarial actors.

The analogy to adversarial questions in HCH is that the questions that humans have hitherto encountered have been drawn from a constrained probability distribution. Humans have only had so much time to ponder. Epistemically idealized reasoners, on the other hand, will have much more time to think and will think about a greatly expanded range of questions relative to baseline humans. We want to be sure that when we take humans out of our usual reasoning environment and place them inside a structure like HCH that we don't create many opportunities for inferences to be altered or actively manipulated by taking advantage of our old function from questions to answers.

Adversarial Questions for Humans

HCH emulates the function f_n that a human researcher, epistemically idealized, would implement from questions to answers. In cases where a human researcher would encounter a question that leads them to significantly reconsider their goals and to work manipulatively against HCH, f_n captures this behavior too. Adversarial questions are thus questions that would lead a given researcher to significantly change their mind about their goals enough for them to stop answering questions in good faith. If HCH could encounter questions like these in the course of its research work, then we, on the outside, can no longer trust that everything HCH tells us is what an idealized version of the exemplar sharing our goals would tell us. Instead, we now have to worry that some of what HCH says might be the result of a converted researcher acting to manipulate or deceive us.

There's some chance that HCH might encounter adversarial questions on its own, without any adversarial process actively throwing them at HCH to steer its answers in some way. While they are relatively rare, powerful ideas and arguments like these (that would lead one to significantly reevaluate their goals) do exist out "in the wild," distributed in idea space. And additionally, adversarial questions can come from adversarial subprocesses inside HCH that work to generate them. And this is a much more worrying source. Think of the difference between being ideologically converted by something you encounter reading library books at random, versus books actively recommended to you by an adversarial actor. The latter is *much* more likely to succeed for some set number of books passed to you.

Think of HCH's search through question space as being pushed around by two "forces." On the one hand, there are "paths of inquiry" that lead you into adversarial questions. Some lines of inquiry are more laden with adversarial questions than others or are more likely to incline a researcher to run a potentially adversarial subprocess. To varying extents, different regions of question space are hostile to aligned human researchers; some domains are more memetically hazardous (in this respect) than others. The anti-alignment computational "force" here is the extent to which exploring a corner of question space optimizes for unaligning a human researcher. As we'll see, there are a variety of modifications to the naïve HCH architecture that we might make in order to have it implement a safer, more trustworthy function than f_n. The countervailing, pro-alignment "force" is the sum of the countermeasures we implement in the HCH architecture to manage the adversarial questions problem. Which of these two forces should we expect to win out at the various scales of HCH (different values of n)? I gather that Christiano's (2019) informed intuition here is that our directed efforts should overpower those countervailing optimizing forces present in the environment and continue to do so better and better as we scale up HCH. His idea is that modifications to HCH designed with an express goal in mind will leverage available compute more efficiently than "accidentally encountered" environmental forces will. I think this is a good argument, and it's good to have it in mind as you think (1) about how likely HCH is to encounter adversarial questions of various kinds and (2) how effective you think the various explored countermeasures will be against the problem.

We'll first talk about three classes of adversarial question that HCH might run across, evaluating their severity. Then we'll discuss the requisite modifications to HCH that mitigate this risk.

Convincing Ideological Arguments

> Poets are the unacknowledged legislators of the world.
>
> —Percy Bysshe Shelley

Alongside religious ideas, one class of idea that has had an outsized influence over world history is the class of convincing ideological arguments. (Note that "convincing" need not mean "sound.") The most influential ideologies and ideological arguments of the last century directly encouraged their hosts to proselytize in their name and to directly check the spread of competing ideological ideas. Large agentic organizations, like political parties, armies, and nation-states, formed because of and fought for various overtly ideological causes. For our purposes, what matters here is that this constitutes an existence proof that there are text and speech inputs (convincing ideological arguments) into humans that will convince some of them to utterly abandon their prior goals and to adopt radically new goals with substantial new demands on them.

What is the minimum length of text input needed to contain a convincing ideological argument with respect to someone? There are certainly several manifesto-length texts with this property (with respect to many people) that the reader has heard of. Are there any Tweets containing widely convincing ideological arguments (a Tweet being a string of at most 280 characters)? It's much harder to make a convincing case for some worldview in just 280 characters than it is to with a book. I'm not confident that *no Tweet could possibly exist with this property* with respect to someone, though. If questions passed to HCH nodes are generally Tweet-length, it's not a guarantee that some questions won't contain convincing ideological arguments in them. On the other hand, if convincing ideological arguments are always manifesto-length, then HCH's explored questions will never contain them.

While ideological inputs have greatly influenced many, I think it's implausible that they pose an intractable issue for HCH alignment. Our HCH tree is built around a carefully chosen exemplar. The sort of person we choose should not be especially susceptible to fallacious, overtly ideological arguments. While almost all of us can be susceptible to ideological cheerleading for poor arguments in *some* of our less serious states of mind, it's a much stronger claim that all of us are always doing so. So long as there is a "research headspace" that we can have our exemplar work in, HCH can learn just this style of serious thinking, skipping over the more emotionally distorted style of cognition the exemplar sometimes employs in their non-professional life. Especially when advised to be on guard against arguments attempting to push around their values, I think careful selection of our exemplar should go far in reducing the risk of encountering a convincing ideological argument with respect to them.

Credible Decision-Theoretic Threats

> The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein …
>
> —Howard Phillips Lovecraft

A more worrying set of questions are those that contain credible decision-theoretic threats to the researcher considering them (or to others they care about). Suppose that an HCH node is researching some question, and in the process of that runs a powerful search algorithm to aid in his work. For example, he might run a powerful automated theorem prover to see whether any negative unforeseen consequences follow from his working formal models of the world. Suppose that this theorem prover returns a valid proof that if he fails to act in a certain way, many instances of him will be simulated up to this point in their life and then tortured forever, should they fail to act as suggested. The researcher might pore over the proof, trying to find some error in its reasoning or assumptions that would show the threat to be non-credible. If the proof checks out, though, he will be led to act in the way the proof suggests, against his originally held goals (assuming he isn't very, horrifically brave in the face of such a threat).

No one has yet encountered a convincing argument to this effect. This implies that whether or not such arguments exist, they are not common in the region of idea space that people have already canvassed. But what matters for HCH alignment is the existential question: Do such arguments exist inside of possible questions or not? A priori, the answer to this question might well be “yes,” as it is very easy to satisfy an existential proposition like this: all that is needed is that one such question containing a credible threat exist. And unlike the above case of political arguments, these hypothetical threats seem moving to even an intelligent, reflective, levelheaded, and advised-to-be-on-guard researcher. Thus, they will be reflected somewhere in the function f_n implemented by a naïve model of HCH.

Unconstrained Searches over Computations

We can generalize from the above two relatively concrete examples of adversarial questions. Consider the set of all text inputs an HCH node could possibly encounter. This set’s contents are determined by the architecture of HCH — if HCH is built so that node inputs are limited to 280 English-language characters, then this set will be the set of all 280-character English-language strings. Political ideas and decision-theoretic ideas are expressed by a small subset of those strings. But every idea expressible in 280 English-language characters will be a possible input into HCH. That set of strings is enormous, and so an enormous fraction of the ideas expressed in it will be thoroughly *alien* to humans — the overwhelming majority of ideas in that idea space will be ideas that no human has ever considered. And the overwhelming majority of the set’s strings won’t express any idea at all — nearly every string in the set will be nonsense.

Abstracting away from ideas that human thinkers have come across in our species’ history, then, what fraction of possible input-ideas into HCH will convert an HCH node on the spot? And abstracting away from the notion of an “idea,” what fraction of 280-character English-language-string inputs will suffice to unalign an HCH node? Forget coherent ideas: are there any short strings (of apparently nonsense characters, akin to contentless epileptic-fit-inducing flashing light displays) that can reliably rewire a person’s goals?

Plausibly, many such possible inputs would unalign a human. I’m inclined to endorse this claim because of the fact that humans have not been designed as provably secure systems. Human brains are the consequence of the messy process of natural selection over organisms occurring on Earth. It would be *remarkable* if humans had already encountered all the most moving possible text-inputs in our collective reflections as a species. What seems overwhelmingly more likely is that human brains have canvassed only a miniscule corner of idea space, and that beyond our little patch, *somewhere* out there in the depths of idea space, *there be dragons*. It’s not that these ideas are particularly easy to find; they’re not. Nearly every short English string is nonsense, and expresses *no* coherent idea nor has any substantial effect on the person looking it over. But the question at hand is the existential question of “Do such human-adversarial questions exist?” I think the answer to this question is yes. And in an extremely computationally powerful system like the one under consideration here, these rare inputs could plausibly be encountered.

Trading Off Competitiveness to Maintain Alignment

In order to preserve HCH alignment in the face of the adversarial questions problem, we'll need to change its architecture. While there are ways of doing this, there is a cost to doing so as well. By modifying the HCH architecture in the ways suggested below, HCH becomes an even more computationally costly algorithm. While it will also become a more probably *aligned* algorithm, this cost in competitiveness bodes poorly for IDA's success and for delegating the alignment problem to HCH. If there are faster, less convoluted capable algorithms out there, then projects that work with those algorithms will be at a competitive advantage relative to a project working with HCH. If alignment depends on an alignment-concerned AI team maintaining a development head start relative to competitor AI projects, the below architectural modifications will come at an alignment cost as well, in lost competitiveness.

That worry aside, my take on the adversarial questions issue is that, while we can foresee the adversarial questions problem for HCH, we can also foresee good solutions to it that will work at scale. Adversarial questions are a problem, but a tractable problem.

Exemplar Rulebooks

One class of solutions is the use of *exemplar rulebooks* during IDA. Instead of simply training HCH on a person decomposing questions and conducting basic research without further guidance, we train HCH on a person doing that *under side constraints given ahead of time*. The HCH exemplar is told to not, for example, ever run an unconstrained search over computations for an answer to a question, as this is an extremely dangerous process likely to produce an unaligned subagent. They might also be told to return an "unable to safely answer this question" response when fed political or decision-theoretic questions. If they hold to the exemplar rulebook during training, an ML model trained on that behavior will infer that it should hold to the rulebook in other cases as well.

A worry arises at this point: does a rulebook solution to the problem of adversarial questions simply pass the issue back without really solving it? Does it presuppose that we can see ahead of time all the problems that HCH might face and be able, ahead of time, to block off all those dangers with our guidelines? I'm inclined to think no: in developing our rulebook, we can block off relatively large swaths of idea space, and so "play it safe." The important thing is to ensure that HCH is outer aligned; we are willing to trade competitiveness for alignment here. If we carve widely enough, all the potentially dangerous inputs should be precluded by our guidelines.

There's an assumption here that human cognition is relatively unlikely to encounter adversarial questions, while alien kinds of computations, like a search over all strings, are more memetically hazardous. People can think a range of thoughts/natively run only certain kinds of computations, but there exist computations that human brains simply cannot run, architecturally. The space of all computations is much larger than the space of human computations. We can infer from the history of human reasoning that human cognition is relatively inefficient at transforming resources into adversarial text-inputs, as people have not produced all that many of those. No such inference can be made for computational search processes generally. We avoid most of the adversarial questions into HCH by remaining in the shallow waters of human cognition, and avoiding at the outset alien search processes like, for example, unconstrained searches for world models fitting parameters.

Internode-Edge Bandwidth Restriction

Another solution to the adversarial questions problem is to restrict bandwidth between HCH nodes (Saunders 2018). If there's a tight constraint on how much information can be passed between nodes in the hierarchy and adversarial questions are in general informationally complex, then bandwidth limits will prevent those inputs from spreading between nodes. Even if one node encounters them and is unaligned by doing so, it will be unable to transmit that input in full back to its parent node. Adversarial questions will then only be inputs that nodes encounter in the course of their own research "within their node," and not something they have to fear receiving from nodes above or below them in the HCH tree. If tight bandwidth limits are employed, then as nodes won't be able to pass as much information between themselves, the tree will have to grow larger in order to do as much search. You can think of tightening bandwidth limits as moving some of HCH's compute out from inside its nodes, instead dividing it up into digestible bits distributed across more nodes (each seeing less of the larger picture).

How tight ought this bandwidth restriction be in order to be sure nodes won't be able to transmit adversarial inputs to one another? Christiano's view is that the length of guaranteed safe inputs is small: about a 6-common-English-word sentence (Saunders 2018). I'm inclined to agree with his assessment: while a manifesto-length text-input might contain very persuasive arguments, it's very hard to see how a 6-common-word sentence could contain enough to risk unaligning an intelligent, thoughtful human researcher. It's worth thinking through for yourself: what was the length of the *last* argument that really changed your mind about something normative? What's the shortest such argument that *ever* changed your mind about something normative? You can make inferences about the likelihood of these inputs (relative to input length) in this way.

Thought Policing

Adversarial questions are observations that nodes in the HCH tree can make as a side effect of their research computation. Nodes aren't deliberately searching for adversarial questions; those inputs simply crop up here and there, because they exist in the input space and HCH is a powerful search process exploring that space.

If adversarial questions are a "byproduct" of HCH's searches, then we might introduce a deliberate countervailing mechanism in HCH to notice and contain adversarial questions as they occur. The hope is that as HCH scales up to larger and larger quantities of compute, this deliberate mechanism will have enough compute on hand at all scales of HCH to suppress the adversarial questions that that scale of tree encounters.

One such mechanism might be the use of doubled-up HCH nodes to examine the research history and transcripts of other nodes for evidence of misalignment (Christiano 2019). Nodes evincing misalignment would then be pruned from the tree, hopefully preserving the alignment of the overall search. (This would, of course, roughly double the computational cost of running HCH for any given question.) In conversation, William Saunders suggested that we might use Debate here — AI assistants pitching arguments to a human judge, where the game theory of the setup means that the first AI assistant will, on its move, win by telling us the truth about the topic it was asked to examine — to pre-screen question inputs into HCH nodes and argue over whether they were too manipulative to look at.

A Patchwork Solution to Adversarial Questions

If some combination of all the above methods are employed, the hope is that HCH will be robust to adversarial questions, and continue to be robust to them as it is scaled up to greater levels of compute consumption. It’s okay for alignment if some parts of idea space are too treacherous for HCH to safely explore. So long as HCH errs on the side of caution and outputs a “I can’t safely explore that question” response whenever it risks entering a dangerous part of input space, its alignment will be preserved.

Formally, think of this as altering the function that we are having HCH learn from its exemplar. Instead of the “naïve” function f_n, we instead have HCH learn the function of an exemplar who is tightly constrained by rulebooks. Coupled with further architectural modifications (like internode bandwidth restrictions and thought policing) HCH instead implements a more constrained function

$$f'_n : Q \to A^*$$

where A^* is the set of all answers augmented with the error code “I can’t explore that question while remaining safely aligned.” f'_n maps many questions to this error code that f_n had attempted to tackle. Thus, f'_n is both less capable and more reliably aligned than f_n. So long as we err on the side of caution and carve off all of the plausibly dangerous regions of question space, a modified HCH implementing the function given by

$$\lim_{n \to \infty} f'_n$$

should act as a superhumanly capable question-answerer that reliably remains goal-aligned with us.

Conclusion

In summary, adversarial questions are a tractable problem for HCH. It should be possible to produce appropriate architectural modifications that work as HCH is scaled up to greater quantities of compute.

The cost of these solutions is generally to expand the HCH tree, thus costing more compute for each search relative to unmodified HCH. Additionally, there are classes of input that HCH won't be able to look at at all, instead returning an "unable to research" response for them. Modified HCH will thus be significantly performance uncompetitive with counterpart ML systems that will exist alongside it, and so we can't simply expect it to be used in place of those systems, as the cost to actors will be too great.

Bibliography

Bensinger, Rob. 2021. "Garrabrant and Shah on Human Modeling in AGI." *LessWrong.* August 4. https://www.lesswrong.com/posts/Wap8sSDoiigrJibHA/garrabrant-and-shah-on-human-modeling-in-agi.

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit, Wendell Wallach and George Eric Lasker, 12-17. International Institute of Advanced Studies in Systems Research and Cybernetics.

—. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Christiano, Paul. 2018. "Humans Consulting HCH." *Alignment Forum.* November 25. https://www.alignmentforum.org/posts/NXqs4nYXaq8q6dTTx/humans-consulting-hch.

—. 2016. "Strong HCH." *AI Alignment.* March 24. https://ai-alignment.com/strong-hch-bedb0dc08d4e.

—. 2019. "Universality and Conrequentialism within HCH." *AI Alignment.* January 9. https://ai-alignment.com/universality-and-consequentialism-within-hch-c0bee00365bd.

Christiano, Paul, Buck Shlegeris, and Dario Amodei. 2018. "Supervising strong learners by amplifying weak experts." *arXiv preprint.*

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. "Risks from Learned Optimization in Advanced Machine Learning Systems." *arXiv preprint* 1-39.

Lantz, Frank. 2017. *Universal Paperclips.* New York University, New York.

Ngo, Richard. 2020. "AGI Safety From First Principles." *LessWrong.* September 28. https://www.lesswrong.com/s/mzgtmmTKKn5MuCzFJ.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Saunders, William. 2018. "Understanding Iterated Distillation and Amplification Claims." *Alignment Forum.* April 17. https://www.alignmentforum.org/posts/yxzrKb2vFXRkwndQ4/understanding-iterated-distillation-and-amplification-claims.

Yudkowsky, Eliezer. 2008. "Artficial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M.

Ćirković, 308-345. Oxford: Oxford University Press.

—. 2018. "Eliezer Yudkowsky comments on Paul’s research agenda FAQ." *LessWrong.* July 1. https://www.greaterwrong.com/posts/Djs38EWYZG8o7JMWY/paul-s-research-agenda-faq/comment/79jM2ecef73zupPR4.

—. 2007. "The Hidden Complexity of Wishes." *LessWrong.* November 23. https://www.lesswrong.com/posts/4ARaTpNX62uaL86j6/the-hidden-complexity-of-wishes.

1. ^

   An important aside here is that this brief section skims over the entire inner alignment problem and IDA’s attempted approach to it. As Yudkowsky notes (2018), plausibly, the remaining inner alignment issue here is so significant that it contains most of the overall alignment problem. Whatever “use powerful inspection tools” means, exactly, is important to spell out in detail; the whole IDA scheme is premised on these inspection tools being sufficiently powerful to ensure that a research models learns basically the function we want it to from a set of training data.

   Note that even at the first amplification step, we’re playing with roughly par-human strength ML models. That is, we’re already handling fire — if you wouldn’t trust your transparency tools to guarantee the alignment of an approximately human-level AGI ML model, they won’t suffice here and this will all fall apart right at the outset.

Agency and Coherence

Epistemic status: spitballing.

"Like Photons in a Laser Lasing"

> When you do lots of reasoning about arithmetic correctly, without making a misstep, that long chain of thoughts with many different pieces diverging and ultimately converging, ends up making some statement that is... still true and still about numbers! Wow! How do so many different thoughts add up to having this property? Wouldn't they wander off and end up being about tribal politics instead, like on the Internet?
>
> And one way you could look at this, is that even though all these thoughts are taking place in a bounded mind, they are shadows of a higher unbounded structure which is the model identified by the Peano axioms; all the things being said are *true about the numbers*. Even though somebody who was missing the point would at once object that the human contained no mechanism to evaluate each of their statements against all of the numbers, so obviously no human could ever contain a mechanism like that, so obviously you can't explain their success by saying that each of their statements was true about the same topic of the numbers, because what could possibly implement that mechanism which (in the person's narrow imagination) is The One Way to implement that structure, which humans don't have?
>
> But though mathematical reasoning can sometimes go astray, when it works at all, it works because, in fact, even bounded creatures can sometimes manage to obey local relations that in turn add up to a global coherence where all the pieces of reasoning point in the same direction, like photons in a laser lasing, even though there's no internal mechanism that enforces the global coherence at every point.
>
> To the extent that the outer optimizer trains you out of paying five apples on Monday for something that you trade for two oranges on Tuesday and then trading two oranges for four apples, the outer optimizer is training all the little pieces of yourself to be locally coherent in a way that can be seen as an imperfect bounded shadow of a higher unbounded structure, and then the system is powerful though imperfect *because* of how the power is present in the coherence and the overlap of the pieces, *because* of how the higher perfect structure is being imperfectly shadowed. In this case the higher structure I'm talking about is Utility, and doing homework with coherence theorems leads you to appreciate that we only know about one higher structure for this class of problems that has a dozen mathematical spotlights pointing at it saying "look here", even though people have occasionally looked for alternatives.

-- Eliezer Yudkowsky, Ngo and Yudkowsky on alignment difficulty

Selection Pressure for Coherent Reflexes

Imagine a population of replicators. Each replicator also possesses a set of randomly assigned reflexive responses to situations it might encounter. For instance, above and beyond reproducing itself after a time step, a replicator might reflexively, probabilistically transform some local situation A, when encountered, into some local situation B. The values of A and B are set randomly and there are no initial consistency requirements, so the replicators will generally behave spastically at this point.

Most of these replicators will end up with *incoherent* sets of reflexes. Some, for example, will cyclically transform A into B into C into A, and so on. Others will transform their environment in "wasteful" ways, moving it into some state that could have been reached with greater certainty via some different series of transformations.

But some of the replicators will possess *coherent* sets of reflexes. These replicators will never "double back" on their previous directional transformations of their situation. They will thus be more successful in reaching some situation S than an incoherent counterpart would be. And when S is a fitness-improving situation, reflexive coherence targeting it will be selected for.

The Instrumental Incentive to Exploit Incoherence

Once you have a population of coherent agents, the selection pressure against reflexive incoherency increases. A dumb-matter environment will throw up situations at random, and so incoherent replicators will only fall into traps as those traps happen to come up. But a population of coherent agents will *actively exploit* the incoherent among them; incoherent agents are now pools of resources for coherent agents to exploit.

Solipsistic vs. Multi-Agent Training Regimes

ML models are generally trained inside a solipsistic world (with notable exceptions). They, all by their lonesome, are fed sense data and then are repeatedly modified by gradient descent to become better at modulating that sense data. There's optimization pressure for them to become reflexively coherent, but not as much as they would face in an environment of Machiavellian coherent agents.

(Of course, if you train enough, even this gentler pressure will add up).

Deceptive Agents are a Good Way to Do Things

A brief, accessible summary of the inner alignment problem.

The safety problem with powerful ML training algorithms is that **deceptive agents are a good way to do things.** Meaning, when we search over a big space of models until we find a model that performs well on some loss function, we'll likely find a deceptive agent. Agents are algorithms whose behaviors try to consistently move the world in some direction. When trapped inside an ML training algorithm, only models that perform well on the loss function will survive training. If an agent aims at pushing the world in some direction, it will have to do so via the circuitous-but-available route of playing along with the training algorithm until it's safely past training. *Deceptive* agents are just agents taking the only available route to the states they are pointed towards from inside of ML training.

How common are deceptive agents inside that big space of ML models? One argument that they are common is that almost any highly capable agent with whatever utility-function will have this route to getting where it's pointed at by means of deception available. "Play along, survive training, and *then* act as you want to" is a simple, effective strategy for a wide range of possible agents trapped inside ML training. Many agents will therefore be deceptive in that situation. If the agent we *hope* training will find optimizes according to a very particular utility function, then that agent will be vanishingly rare compared to its deceptive counterparts in model space, and training will always stumble on a deceptive model first.

So, by default, powerful ML training algorithms grading models on some loss function will find deceptive agents, because deceptive agents are a good way to do things.

But What's Your *New Alignment Insight,* out of a Future-Textbook Paragraph?

This is something I've been thinking about a good amount while considering my model of Eliezer's model of alignment. After tweaking it a bunch, it sure looks like a messy retread of much of what Richard says here; I don't claim to assemble any new, previously unassembled insights here.

Tl;dr: For impossibly difficult problems like AGI alignment, the worlds in which we solve the problem will be worlds that came up with some new, intuitively compelling insights. On our priors about impossibly difficult problems, worlds without new intuitive insights don't survive AGI.

Object-Level Arguments for Perpetual Motion

> I once knew a fellow who was *convinced* that his system of wheels and gears would produce reactionless thrust, and he had an Excel spreadsheet that would prove this - which of course he couldn't show us because he was still developing the system. In classical mechanics, violating Conservation of Momentum is *provably* impossible. So any Excel spreadsheet calculated *according to the rules of classical mechanics* must *necessarily* show that no reactionless thrust exists - unless your machine is complicated enough that you have made a mistake in the calculations.
>
> And similarly, when half-trained or tenth-trained rationalists abandon their art and try to believe without evidence just this once, they often build vast edifices of justification, confusing themselves just enough to conceal the magical steps.
>
> It can be quite a pain to nail down where the magic occurs - their structure of argument tends to morph and squirm away as you interrogate them. But there's always some step where a tiny probability turns into a large one - where they try to believe without evidence - where they step into the unknown, thinking, "No one can prove me wrong".
>
> …
>
> Hey, maybe if you add enough wheels and gears to your argument, it'll turn warm water into electricity and ice cubes! **Or, rather, you will no longer see why this *couldn't* be the case.**
>
> **"Right! I *can't* see why couldn't be the case! So maybe it is!"**
>
> ***Another* gear? That just makes your machine even *less* efficient. It wasn't a perpetual motion machine before, and each extra gear you add makes it even less efficient than that.**

> Each extra detail in your argument necessarily decreases the joint probability. The probability that you've violated the Second Law of Thermodynamics without knowing exactly how, by guessing the exact state of boiling water without evidence, so that you can stick your finger in without getting burned, is, necessarily, even less than the probability of sticking in your finger into boiling water without getting burned.
>
> I say all this, because people really do construct these huge edifices of argument in the course of believing without evidence. One must learn to see this as analogous to all the wheels and gears that fellow added onto his reactionless drive, until he finally collected enough complications to make a mistake in his Excel spreadsheet.

Manifestly Underpowered Purported Proofs

> If I read all such papers, then I wouldn't have time for anything else. It's an interesting question how you decide whether a given paper crosses the plausibility threshold or not … Suppose someone sends you a complicated solution to a famous decades-old math problem, like P vs. NP. How can you decide, in ten minutes or less, whether the solution is worth reading?
>
> …
>
> **The techniques just seem too wimpy for the problem at hand.** Of all ten tests, this is the slipperiest and hardest to apply — but also the decisive one in many cases. As an analogy, **suppose your friend in Boston blindfolded you, drove you around for twenty minutes, then took the blindfold off and claimed you were now in Beijing. Yes, you do see Chinese signs and pagoda roofs, and no, you can't immediately disprove him — but based on your knowledge of both cars and geography, isn't it more likely you're just in Chinatown?** I know it's trite, but this is exactly how I feel when I see (for example) a paper that uses category theory to prove NL≠NP. We start in Boston, we end up in Beijing, and at no point is anything resembling an ocean ever crossed.

What's going on in the above cases is argumentation from "genre savviness" about our physical world: knowing, based on the reference class that a purported feat would fall into, the probabilities of feat success conditional on its having or lacking various features. These meta-level arguments rely on knowledge about what belongs in which reference class, rather than on in-the-weeds object-level arguments about the proposed solution itself. It's better to reason about things concretely, when possible, but in these cases the meta-level heuristic has a well-substantiated track record.

Successful feats will all have a certain superficial shape, so you can sometimes evaluate a purported feat based on its superficial features alone. One instance where we might really care about doing this is where we only get one shot at a feat, such as aligning AGI, and if we fail our save everyone dies. In that case, we will not get lots of postmortem time to poke through how we failed and learn the object-level insights after the fact. We just die. We'll have to evaluate our possible feats in light of their non-hindsight-based features, then.

Let's look at the same kind of argument, courtesy Eliezer, about alignment schemes:

On Priors, is "Weird Recursion" Not an Answer to Alignment?

> I remark that this intuition matches what the wise might learn from Scott’s parable of K’th’ranga V: **If you know how to do something then you know how to do it directly rather than by weird recursion, and what you imagine yourself doing by weird recursion you probably can’t really do at all. When you want an airplane you don’t obtain it by figuring out how to build birds and then aggregating lots of birds into a platform that can carry more weight than any one bird and then aggregating platforms into megaplatforms until you have an airplane; either you understand aerodynamics well enough to build an airplane, or you don’t, the weird recursion isn’t really doing the work.** It is by no means clear that we would have a superior government free of exploitative politicians if all the voters elected representatives whom they believed to be only slightly smarter than themselves, until a chain of delegation reached up to the top level of government; either you know how to build a less corruptible relationship between voters and politicians, or you don’t, the weirdly recursive part doesn’t really help. It is no coincidence that modern ML systems do not work by weird recursion because all the discoveries are of how to just do stuff, not how to do stuff using weird recursion. (Even with AlphaGo which is arguably recursive if you squint at it hard enough, you’re looking at something that is not weirdly recursive the way I think Paul’s stuff is weirdly recursive, and for more on that see https://intelligence.org/2018/05/19/challenges-to-christianos-capability-amplification-proposal/.)
>
> It’s in this same sense that I intuit that if you could inspect the local elements of a modular system for properties that globally added to aligned corrigible intelligence, it would mean you had the knowledge to build an aligned corrigible AGI out of parts that worked like that, not that you could aggregate systems that corrigibly learned to put together sequences of corrigible thoughts into larger corrigible thoughts starting from gradient descent on data humans have labeled with their own judgments of corrigibility.

Eliezer often asks, "Where's your couple-paragraph-length insight from the Textbook from the Future"? Alignment schemes are purported solutions to problems in the reference class of impossibly difficult problems, in which we're actually doing something new, like inventing mathematical physics for the very first time, and doing so playing against emerging superintelligent optimizers. As far as I can tell, Eliezer's worry is that proposed alignment schemes spin these long arguments for success that just amount to burying the problem deep enough to fool yourself. *That*'s why any proposed solution to alignment *has* to yield a core insight or five that we didn't have before -- conditional on an alignment scheme looking good without a simple new insight, you've probably just buried the hard core of the problem deep enough in your arguments to fool your brain.

So it's fair to ask any alignment scheme what its new central insight into AI is, in a paragraph or two. If these couple of paragraphs read like something from the Textbook from the Future, then the scheme might be in business. If the paragraphs contain no brand new, intuitively compelling insights, then the scheme probably doesn't contain the necessary insights but-distributed-across-its-whole-body either.[1]

[1] ^

Though this doesn't mean that pursuing that line of research further *couldn't* lead to the necessary insights. The science just has to eventually get to those insights if alignment is to work.

Gato as the Dawn of Early AGI

Written in a hurry today at the EA UCLA AI Timelines Workshop. Long and stream-of-thought, and a deliberate intellectual overreach as an epistemic exercise. My first foray into developing my own AGI timelines model without deferring! Please, I beg of you, tell me why I'm wrong in the comments!

Epistemic status: Small-N reasoning. Low confidence, but represents my standing understanding of AGI timelines as of now.

This exchange caught my eye a couple days ago:

> Yitz:
>
> Would it be fair to call this AGI, albeit not superintelligent yet?
>
> > Gato performs over 450 out of 604 tasks at over a 50% expert score threshold.
>
> 👀
>
> Daniel Kokotajlo:
>
> Yes. Sub-human-level AGI.

If true, this is a huge milestone!

Here I'm combining thinking about this with thinking about AGI 10 years hence. The latter forecasting task is totally different if we have a form of AGI *as of two days ago*, even an admittedly weak form.

Do We Have "Subhuman AGI" as of Two Days Ago?

If I want to forecast AGI 10 years out, I first want to understand current-year AGI. Do we currently have AGI? What does it look like? What hyperparameters, up to and including overall architecture, might we push on and make progress on in the coming decade?

To start, I'll read the Gato paper out of DeepMind, investigating for myself Daniel's above claim that Gato constitutes subhuman AGI:[1]

"A Generalist Agent" (Reed et al., 2022)

> We focus our training at the operating point of model scale that allows real-time control of real-world robots, currently around 1.2B parameters in the case of Gato. As hardware and model architectures improve, this operating point will naturally increase the feasible model size, pushing generalist models higher up the scaling law curve. For simplicity Gato was trained offline in a purely supervised manner; however, in principle, there is no reason it could not also be trained with either offline or online reinforcement learning (RL) (p. 2).

Gato is small, parameters-wise. At 1.2 billion parameters, it's 1/100th the size of the largest GPT-3 model and 1/500th the size of PaLM. This is a choice due to hardware constraints; larger models, which definitely could have been used here, would not be able to operate real-world robotic limbs in real time. Additionally, the tasks on which Gato was trained were chosen purely incidentally: they could have easily been otherwise. So Gato is a miniaturized version of models to come in the near future.

> After converting data into tokens, we use the following canonical sequence ordering.
>
> - Text tokens in the same order as the raw input text.
> - Image patch tokens in raster order.
> - Tensors in row-major order.
> - Nested structures in lexicographical order by key.
> - Agent timesteps as observation tokens followed by a separator, then action tokens.
> - Agent episodes as timesteps in time order (p. 3).

Gato is a large transformer with a single sense-modality: it receives diverse kinds of inputs pressed into a sequence of tokens, and outputs more tokens. This is just applying the current successful language-model architecture to varied-domain problem solving in the naïve way.

> Because distinct tasks within a domain can share identical embodiments, observation formats and action specifications, the model sometimes needs further context to disambiguate tasks. Rather than providing e.g. one-hot task identifiers, we instead … use prompt conditioning. During training, for 25% of the sequences in each batch, a prompt sequence is prepended, coming from an episode generated by the same source agent on the same task. Half of the prompt sequences are from the end of the episode, acting as a form of goal conditioning for many domains; and the other half are uniformly sampled from the episode. During evaluation, the agent can be prompted using a successful demonstration of the desired task, which we do by default in all control results that we present here.
>
> …Because agent episodes and documents can easily contain many more tokens than fit into context, we randomly sample subsequences of L tokens from the available episodes. Each batch mixes subsequences approximately uniformly over domains (e.g. Atari, MassiveWeb, etc.), with some manual upweighting of larger and higher quality datasets (p. 4).
>
> …
>
> Ideally, the agent could potentially learn to adapt to a new task via conditioning on a prompt including demonstrations of desired behaviour. However, due to accelerator memory constraints and the extremely long sequence lengths of tokenized demonstrations, the maximum context length possible does not allow the agent to attend over an informative-enough context. Therefore, to adapt the agent to new tasks or behaviours, we choose to fine-tune the agent's parameters on a limited number of demonstrations of a single task, and then evaluate the fine-tuned model's performance in the environment (p. 11).

A major part of the guts of the model is the use of internal prompt programming. Context length limits don't prevent training high-performance in Gato, but does prevent us from fully testing the out-of-the-box model's few-shot generalization abilities.

> Our control tasks consist of datasets generated by specialist SoTA or near-SoTA reinforcement learning agents trained on a variety of different environments. For each environment we record a subset of the experience the agent generates (states, actions, and rewards) while it is training (p. 5).

Gato is trained to do RL-style tasks by supervised learning on token sequences generated from state-of-the-art RL model performance. These tasks take place in both virtual and real-world robot arm environments.

> Figure 10 compares the success rate of Gato across different fine-tuning data regimes to the sim-to-real expert and a Critic-Regularized Regression (CRR) (Wang et al., 2020) agent trained on 35k episodes of all test triplets. Gato, in both reality and simulation (red curves on the left and right figure, respectively), recovers the expert's performance with only 10 episodes, and peaks at 100 or 1000 episodes of fine-tuning data, where it exceeds the

> expert. After this point (at 5000), performance degrades slightly but does not drop far below the expert's performance (p. 12).

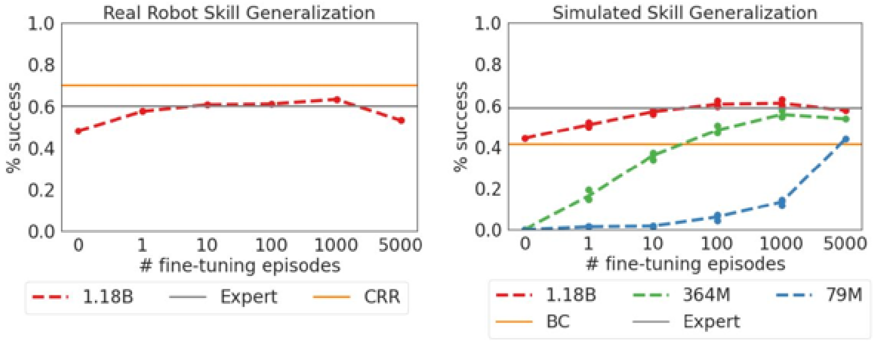

Figure 10 | **Robotics fine-tuning results.** Left: Comparison of real robot Skill Generalization success rate averaged across test triplets for Gato, expert, and CRR trained on 35k expert episodes (upper bound). Right: Comparison of simulated robot Skill Generalization success rate averaged across test triplets for a series of ablations on the number of parameters, including scores for expert and a BC baseline trained on 5k episodes.

A crucial datapoint for the subhuman AGI question! Gato, after a small amount of fine-tuning, catches up with SOTA RL expert models. It does this even with real-world colored-block stacking tasks; Gato is capable of interfacing with the physical world, albeit in a controlled environment.

> Gato was inspired by works such as GPT-3 (Brown et al., 2020) and Gopher (Rae et al., 2021), pushing the limits of generalist language models; and more recently the Flamingo (Alayrac et al., 2022) generalist visual language model. Chowdhery et al. (2022) developed the 540B parameter Pathways Language Model (PalM) explicitly as a generalist few-shot learner for hundreds of text tasks. Future work should consider how to unify these text capabilities into one fully generalist agent that can also act in real time in the real world, in diverse environments and embodiments (p. 15).

Inarticulate screaming

> In this work we learn a single network with the same weights across a diverse set of tasks.
>
> Recent position papers advocate for highly generalist models, notably Schmidhuber (2018) proposing one big net for everything, and Bommasani et al. (2021) on foundation models. However, to our knowledge there has not yet been reported a single generalist trained on hundreds of vision, language and control tasks using modern transformer networks at scale.
>
> "Single-brain"-style models have interesting connections to neuroscience. Mountcastle (1978) famously stated that "the processing function of neocortical modules is qualitatively similar in all neocortical regions. Put shortly, there is nothing intrinsically motor about the motor cortex, nor sensory about the sensory cortex". Mountcastle found that columns of neurons in the cortex behave similarly whether associated with vision, hearing or motor control. This has motivated arguments that we may only need one algorithm or model to build intelligence (Hawkins and Blakeslee, 2004).
>
> Sensory substitution provides another argument for a single model (Bach-y Rita and Kercel, 2003). For example, it is possible to build tactile visual aids for blind people as follows. The signal captured by a camera can be sent via an electrode array on the tongue to the brain. The visual cortex learns to process and interpret these tactile signals, endowing the person with some form of "vision". Suggesting that, no matter the type of input signal, the same network can process it to useful effect (p. 16).

> …
>
> There has been great recent interest in data-driven robotics (Cabi et al., 2019; Chen et al., 2021a). However, Bommasani et al. (2021) note that in robotics “the key stumbling block is collecting the right data. Unlike language and vision data, robotics data is neither plentiful nor representative of a sufficiently diverse array of embodiments, tasks, and environments”. Moreover, every time we update the hardware in a robotics lab, we need to collect new data and retrain. We argue that this is precisely why we need a generalist agent that can adapt to new embodiments and learn new tasks with few data (p. 17).
>
> …
>
> Transformer sequence models are effective as multi-task multi-embodiment policies, including for real-world text, vision and robotics tasks. They show promise as well in few-shot out-of-distribution task learning. In the future, such models could be used as a default starting point via prompting or fine-tuning to learn new behaviors, rather than training from scratch (p. 18).

Finally, descriptions of the range of tasks Gato is trained on:

> ## F.1. Atari
>
> We collect two separate sets of Atari environments. The first (that we refer to as ALE Atari) consists of 51 canonical games from the Arcade Learning Environment (Bellemare et al., 2013). The second (that we refer to as ALE Atari Extended) is a set of alternative games3 with their game mode and difficulty randomly set at the beginning of each episode.
>
> For each environment in these sets we collect data by training a Muesli (Hessel et al., 2021) agent for 200M total environment steps. We record approximately 20,000 random episodes generated by the agent during training.
>
> ## F.2. Sokoban
>
> Sokoban is a planning problem (Racanière et al., 2017), in which the agent has to push boxes to target locations. Some of the moves are irreversible and consequently mistakes can render the puzzle unsolvable. Planning ahead of time is therefore necessary to succeed at this puzzle. We use a Muesli (Hessel et al., 2021) agent to collect training data.
>
> ## F.3. BabyAI
>
> BabyAI is a gridworld environment whose levels consist of instruction-following tasks that are described by a synthetic language. We generate data for these levels with the built-in BabyAI bot. The bot has access to extra information which is used to execute optimal solutions, see Section C in the appendix of (Chevalier-Boisvert et al., 2018) for more details about the bot. We collect 100,000 episodes for each level.
>
> ## F.4. DeepMind Control Suite
>
> The DeepMind Control Suite (Tassa et al., 2018; Tunyasuvunakool et al., 2020) is a set of physicsbased simulation environments. For each task in the control suite we collect two disjoint sets of data, one using only state features and another using only pixels. We use a D4PG (Barth-Maron et al., 2018) agent to collect data from tasks with state features, and an MPO (Abdolmaleki et al., 2018) based agent to collect data using pixels.
>
> We also collect data for randomized versions of the control suite tasks with a D4PG agent. These versions randomize the actuator gear, joint range, stiffness, and damping, and geom size and density. There are two difficulty settings for the randomized versions. The small

setting scales values by a random number sampled from the union of intervals [0.9, 0.95] ∪ [1.05, 1.1]. The large setting scales values by a random number sampled from the union of intervals [0.6, 0.8] ∪ [1.2, 1.4].

F.5. DeepMind Lab

DeepMind Lab (Beattie et al., 2016) is a first-person 3D environment designed to teach agents 3D vision from raw pixel inputs with an egocentric viewpoint, navigation, planning.

We collect data for 255 tasks from the DeepMind Lab, 254 of which are used during training, the left out task was used for out of distribution evaluation. Data is collected using an IMPALA (Espeholt et al., 2018) agent that has been trained jointly on a set of 18 procedurally generated training tasks. Data is collected by executing this agent on each of our 255 tasks, without further training.

F.6. Procgen Benchmark

Procgen (Cobbe et al., 2020) is a suite of 16 procedurally generated Atari-like environments, which was proposed to benchmark sample efficiency and generalization in reinforcement learning. Data collection was done while training a R2D2 (Kapturowski et al., 2018) agent on each of the environments. We used the hard difficulty setting for all environments except for maze and heist, which we set to easy.

F.7. Modular RL

Modular RL (Huang et al., 2020) is a collection of MuJoCo (Todorov et al., 2012) based continuous control environments, composed of three sets of variants of the OpenAI Gym (Brockman et al., 2016) Walker2d-v2, Humanoid-v2, and Hopper-v2. Each variant is a morphological modification of the original body: the set of morphologies is generated by enumerating all possible subsets of limbs, and keeping only those sets that a) contain the torso, and b) still form a connected graph. This results in a set of variants with different input and output sizes, as well as different dynamics than the original morphologies. We collected data by training a single morphology-specific D4PG agent on each variant for a total of 140M actor steps, this was done for 30 random seeds per variant.

F.8. DeepMind Manipulation Playground

The DeepMind Manipulation Playground (Zolna et al., 2021) is a suite of MuJoCo based simulated robot tasks. We collect data for 4 of the Jaco tasks (box, stack banana, insertion, and slide) using a Critic-Regularized Regression (CRR) agent (Wang et al., 2020) trained from images on human demonstrations. The collected data includes the MuJoCo physics state, which is we use for training and evaluating Gato.

F.9. Meta-World

Meta-World (Yu et al., 2020) is a suite of environments for benchmarking meta-reinforcement learning and multi-task learning. We collect data from all train and test tasks in the MT50 mode by training a MPO agent (Abdolmaleki et al., 2018) with unlimited environment seeds and with access to state of the MuJoCo physics engine. The collected data also contains the MuJoCo physics engine state.

…

The specialist Meta-World agent described in Section 5.5 achieves 96.6% success rate averaged over all 50 Meta-World tasks. The detailed success rates are presented in Table 7. We evaluated agent 500 times for each task (pp. 36-7, 39-40).

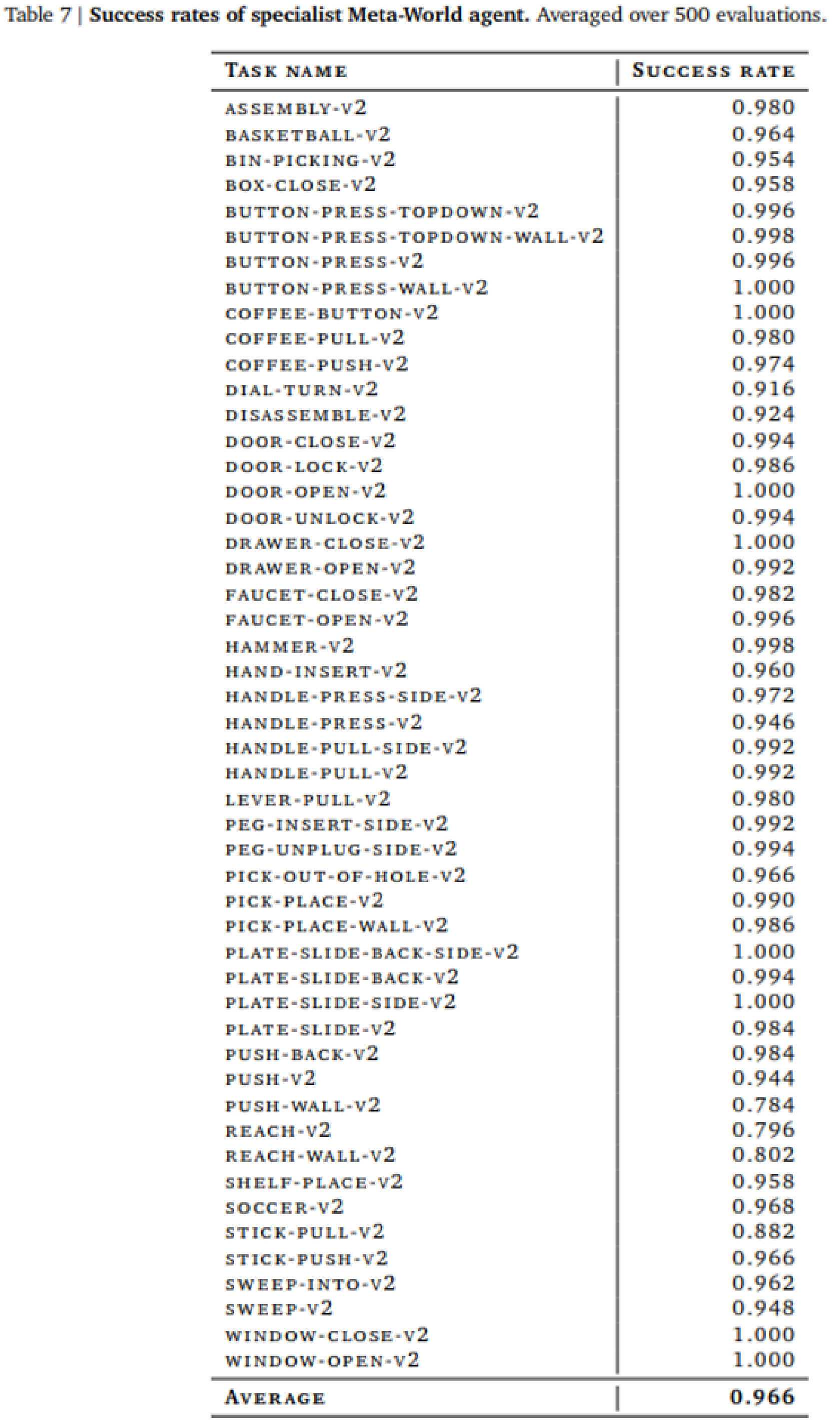

Table 7 | **Success rates of specialist Meta-World agent.** Averaged over 500 evaluations.

Task name	Success rate
assembly-v2	0.980
basketball-v2	0.964
bin-picking-v2	0.954
box-close-v2	0.958
button-press-topdown-v2	0.996
button-press-topdown-wall-v2	0.998
button-press-v2	0.996
button-press-wall-v2	1.000
coffee-button-v2	1.000
coffee-pull-v2	0.980
coffee-push-v2	0.974
dial-turn-v2	0.916
disassemble-v2	0.924
door-close-v2	0.994
door-lock-v2	0.986
door-open-v2	1.000
door-unlock-v2	0.994
drawer-close-v2	1.000
drawer-open-v2	0.992
faucet-close-v2	0.982
faucet-open-v2	0.996
hammer-v2	0.998
hand-insert-v2	0.960
handle-press-side-v2	0.972
handle-press-v2	0.946
handle-pull-side-v2	0.992
handle-pull-v2	0.992
lever-pull-v2	0.980
peg-insert-side-v2	0.992
peg-unplug-side-v2	0.994
pick-out-of-hole-v2	0.966
pick-place-v2	0.990
pick-place-wall-v2	0.986
plate-slide-back-side-v2	1.000
plate-slide-back-v2	0.994
plate-slide-side-v2	1.000
plate-slide-v2	0.984
push-back-v2	0.984
push-v2	0.944
push-wall-v2	0.784
reach-v2	0.796
reach-wall-v2	0.802
shelf-place-v2	0.958
soccer-v2	0.968
stick-pull-v2	0.882
stick-push-v2	0.966
sweep-into-v2	0.962
sweep-v2	0.948
window-close-v2	1.000
window-open-v2	1.000
Average	**0.966**

Where the bar for impressive-but-subhuman performance is set by other models, which might possess diverse architectures very unlike Gato's, Gato is a subhuman AGI. Gato generalizes to previously held-out tasks, *including real-world robotics tasks*, after 10 episodes of fine-tuning. (This is only because of context window limits, and we could test the model on few-shot learning with only context in these varied domains were that window larger.)

As with GPT-3, one scary feature of Gato's success is that its architecture and hyperparameters aren't strongly optimized for what it does. It's basically a (relatively small!) large language model pressed into service as a generalist, *and that just works. AGI is here, and it wasn't that hard to engineer.* Quoth Gwern, "Scaling just works."

A Decade of Actual AI

My guess is that this heralds the beginning of "actual AI," meaning AI reaching out into the world of atoms and not merely the world of bits. I don't mean to disparage progress in ML; if you're anything like me, the visceral impressiveness of GPT-2 and -3 are a big part of why you're throwing yourself into trying to help with alignment! But I was promised a flying car in my childhood, dammit. I remember reading a kid's science magazine that promised me I'd be commuting via space elevator by now! Will we get our household robots anytime soon?

If a slapped together, relatively small transformer like Gato can, with a *minimum* of fine-tuning (10 episodes), generalize well to previously unseen robotics tasks, then Gato's descendants, heavily optimized for success, can plausibly do much better. For Bayesian reasons, the most important part of a secret is that the secret exists, and the genie is now out of the bottle regarding naïve transformers and their potential varied applications.

From an alignment perspective, this is *horrifying.* The worlds in which we can marshal sufficient Coordination between AI labs to prevent our doom … are worlds in which there aren't a hundred disparate actors all rushing to AGI because there's gold in them thar hills. But creating common knowledge of that seems to be what just happened, emboldening efforts to pursue and market, e.g., AI personal assistants.

Extrapolating from the successes of the past decade of successes in deep learning (recalling that transformers *only date back to 2017*), we should naively expect the equivalent-in-impressiveness of the GPT series in other domains, including in real-world robotics.[2] Some spitballing implications: We should expect fully competent self-driving to be solved. We should expect customer-service chatbots to be solved -- AI won't pass the adversarial Turing test by 2032, but it will pass the average-case Turing test, and so be ready for deployment in relatively-low-stakes conversational roles. Factory robots get much better, sufficient to work in complicated domains like households and fast-food restaurants; the internet learns semantics and so websites take on forms much more interesting than static text, image, and video elements; we begin seriously using computers via input channels other than keyboard-and-mouse, as those are currently blocked on ML interpretation of messy human input.

Shallow Pattern Matching in the World of Atoms

While Gato constitutes a kind of subhuman AGI, it is *not* anything like human-grade AGI. Fundamentally, as its architecture has not substantially changed from, e.g., GPT-3, Gato's intrinsic limits don't fundamentally exceed those in GPT-3.

Nostalgebraist on the GPT series' capabilities:

> I don't even know how many tens of thousands of LM samples I've read by now. *(Just my bot alone has written 80,138 posts -- and counting -- and while I no longer read every new one these days, I did for a very long time.)*
>
> Read enough, and you will witness the LM both failing *and* succeeding at anything your mind might want to carve out as a "capability." You see the semblance of abstract reasoning shimmer across a strings of tokens, only to yield to suddenly to absurd, direct self-contradiction. You see the model getting each fact right, then wrong, then right. I see no single, stable trove of skills being leveraged here and there as needed. I just see stretches of success and failure at imitating ten thousand different kinds of people, all nearly independent of one another, the products of barely-coupled subsystems.

> This is hard to refute, but I think this is something you only grok when you read enough LM samples -- where "enough" is a pretty big number.
>
> GPT makes many mistakes, but many of these mistakes are of *types* which it only makes rarely. Some mistake the model makes only every 200 samples, say, is invisible upon one's first encounter with GPT. You don't even notice that model is "getting it right," any more than you would notice a fellow human "failing to forget" that water flows downhill. It's just part of the floor you think you're standing on.
>
> The first time you see it, it surprises you, a crack in the floor. By the fourth time, it doesn't surprise you as much. The fortieth time you see the mistake, you don't even notice it, because "the model occasionally gets this wrong" has become part of the floor.
>
> Eventually, you no longer picture of a floor with cracks in it. You picture a roiling chaos which randomly, but regularly, coalesces into ephemeral structures possessing randomly selected subsets of the properties of floors.

Large language models today sit in this weird uncanny valley of ability, where they are both shockingly good at writing (GPT-2 at its best writing a B-grade high school history essay) *and* pick up the idiot ball at moments *no* human would (nostalgebraist on GPT-3's inhuman metafictional tendencies). Gato exports this uncanny valley of competence that no human possesses into the world of atoms.

In the way that PaLM cannot pass an adversarial Turing test, correspondingly scaled-up Gato won't successfully control a humanoid robot in unfamiliar domains on arbitrary unfamiliar physical tasks. But Gato still exports a whole lot of competence into the physical world (and into a whole host of varied tasks in virtual environments too.) Even if large language models and scaled-up Gato peter out at some point, the lack of intense optimization work put into them so far suggests that we haven't come *close* to mining out this capabilities vein, and we should only expect nostalgebraist's "cracks in the floor" to narrow and often close up in the coming decade of AI capabilities progress.

A Milestone and a Plea

I may be totally off-base here. This summary and projection is built on my very limited model of AI capabilities. I hope to *God* I'm just confused, and am eager to update my model.

But if subhuman AGI is here, and if we're kicking off the final race to human-grade AGI now, even just the very beginning of it … then timelines are *extraordinarily* short. Others mentioned in earlier Gato posts that their models predicted something like this; reading the Gato paper, I realize that my model (insofar as I had one) was surprised by this. I am scared and have shrunk my timelines.

Please, let's do something about AGI alignment.

1. ^

   Admittedly a bit of abuse of terminology, if by "AGI" we usually mean human-equivalent AI across a range of diverse task domains. By "subhuman AGI" here, I mean an AI that performs a wide range of disparate tasks at levels only modestly below typical human performance.

2. ^

   Assuming meaningful ML applications to the world of atoms aren't completely forbidden by regulation in advance of their deployment. I frame everything below with this caveat in mind.

Intelligence in Commitment Races

A distillation of my understanding of the commitment races problem.

Greaser Courage

It's 1950 and you're a greaser in a greaser gang. Recreationally, you're driving your heavy American cars (without seatbelts) at each other at speed, and seeing who swerves first. Whoever swerves first is a coward and is "chicken"; their opponent is courageous and the victor. Both drivers swerving is less humiliating than just you swerving. Both not swerving means both drivers die.

The bolts on your steering wheel have been loosened in the chop shop, and so your steering wheel can be removed if you pull it out towards you.

If you remove your steering wheel and throw it prominently out your window, then your opponent will see this and realize that you are now incapable of swerving. They will then swerve, as they prefer humiliation to death. This wins you glory that you live to see.

But if you both *simultaneously* throw your steering wheels out the window, then neither of you will be able to slow down in time and both of you will die.

Commitment Races

The two above greasers are thus in a situation where one can individually do better by throwing out their steering wheel quickly, but fares worse if both adopt this strategy. *Both-drivers-having-their-steering-wheels* is a commons that each greaser can take from, but both do poorly if the commons is depleted by over-exhaustion. Greasers with unloosened steering wheels *don't* share a *both-drivers-having-their-steering-wheels* commons -- because each greaser can commit ahead of time to not swerving, this commons exists. Because the greasers are both itching to commit first to not swerving, we say that the greasers playing chicken with loosened steering wheels are in a *commitment race* with each other.

Besides throwing out steering wheels, other kinds of actions allow agents to *precommit* to act ahead of time, in view of their opponents. If you can alter your source code so that you will *definitely* pass up a somewhat desirable contract if your ideal terms aren't met, you'll be offered better contracts than agents that can't alter their source code. Alice the human and Bot the human-intelligence AGI are trading with each other. Bot has edit access to his own source code; Alice does not have access to hers. Before beginning any trading negotiations with Alice, Bot is sure to modify his own source code so that he won't accept less than almost all of the value pie. Bot then shows this self-modification to Alice, before going to trade. Alice will now offer Bot a trade where she gets almost nothing and Bot gets almost everything: Alice still prefers a trade where she gets almost nothing to a trade where she gets literally nothing.

Something perverse has happened here! Before self-modifying, Bot had a huge panoply of possible contracts that he could agree to or reject. After self-modifying, Bot had strictly fewer options available to him. Bot did better in life by *throwing away options* that he previously had! Alice and Bot entered into a trade relationship because they both understand the notion of positive-sum interactions; they're both smart, sophisticated agents. Alice brought all those smarts with her into the trading room. Bot threw some of his options away and made his future self effectively dumber. Usually, when we study rationality we find that intelligence is good for finding and choosing the best option out of a big set of alternatives. Usually, smart rationalists want more options because that means a greater chance of the alternatives including an even better option. Smart rationalists want a lot of options, because that gives them more possible stabs at a better alternative, and then want their reasoning to steer their final decision after sifting through those alternatives. Being smarter is ordinarily good for quickly and accurately sifting through larger option spaces. With commitment races, being smarter and using that to sift through options is a losing move. Being smart in a commitment race is, unusually, a losing *position* -- you win in a commitment race to the extent that you can make yourself dumb, *fast*.

> Eliezer:
>
> IMO, commitment races only occur between agents who will, in some sense, act like idiots, if presented with an apparently 'committed' agent. If somebody demands $6 from me in the Ultimatum game, threatening to leave us both with $0 unless I offer at least $6 to them... then I offer $6 with slightly less than 5/6 probability, so they do no better than if they demanded $5, the amount I think is fair. They cannot evade that by trying to make some 'commitment' earlier than I do. I expect that, whatever is the correct and sane version of this reasoning, it generalizes across all the cases.
>
> I am not locked into warfare with things that demand $6 instead of $5. I do not go around figuring out how to invert their utility function for purposes of threatening them back - 'destroy all utility-function inverters (but do not invert their own utility functions)' was my guessed commandment that would be taught to kids in dath ilan, because you don't want reality to end up full of utilityfunction inverters.
>
> From the beginning, I invented timeless decision theory because of being skeptical that two perfectly sane and rational hyperintelligent beings with common knowledge about each other would have no choice but mutual defection in the oneshot prisoner's dilemma. I suspected they would be able to work out Something Else Which Is Not That, so I went looking for it myself. I suggest cultivating the same suspicion with respect to the imagination of commitment races between Ultimatum Game players, in which whoever manages to make some move logically first walks away with $9 and the other poor agent can only take $1 - especially if you end up reasoning that the computationally weaker agent should be the winner.

Eliezer's argument above is that rational agents use precommitments to shape their outward-facing incentives profile so that all agents trading with them are incentivized to offer Shapley splits of the value pie. On Eliezer's view, one should precommit such that any agent like Bot will get a lower EV by offering you almost nothing than by offering you the Shapley split.

Bot has read enough Eliezer on LessWrong to know how Eliezer will precommit. Knowing this, Bot does best by offering Eliezer Shapley splits, and by accepting Eliezer's Shapley offers. Because Bot knows what Eliezer is like, Bot has no reason to self-modify before trading with Eliezer. Bot happily rips off Alice, but doesn't rip off Eliezer.

But what if Bot *doesn't* yet know much about Eliezer? If Bot reasons that self-modifying is worthwhile in expectation before it reasons about what Eliezer is like as a trader, then Bot will end up offering Eliezer almost nothing anyways. Bot would have moved first in 'logical time,' and so Eliezer's precommitment footwork wouldn't have logically subjunctively prevented Bot's self-modification. Moreover, if an as-of-yet ignorant Bot has some *premonition* that learning more about Eliezer will make Bot encounter truths he'd rather not encounter, Bot can self-modify on the basis of that premonition, before risking reading up on Eliezer on LessWrong. Note that it *would not* be sufficient for Bot to self-modify to *forget* any inconvenient truths Bot had learned about Eliezer. Bot's self-modifications really do have to precede Bot's understanding Eliezer, as Eliezer's incentives profile will leave a forgot-and-then-self-modified Bot worse off.

Bot will only win in a commitment race with Eliezer if Bot self-modifies for the wrong reason, in advance of understanding *why* self-modification is valuable. Bot, if successful, acted on mere premonitions in his prior reasoning about self-modification. Bot got to, and could *only* get to, a winning state in the logical mental game against Eliezer "by accident."

Bot will be more and more tempted to self-modify the more it learns about bargaining and decision theory. But Bot will be more and more averse to self-modification against Eliezer as it improves its Eliezer model. If Eliezer takes pains to be easy to model, then Bot will become adverse to self-modifying against Eliezer sooner in Bot's reflection. But because Eliezer would never precommit to probably turn down a rock with an un-Shapley offer painted on its front (because non-agents bearing fixed offers created *ex nihilo* cannot be deterred or made less likely through *any* precommitment) there's always some state for Bot to stumble into in its path of reflection and self-modification where Bot comes out on top.[1]

1. ^

   H/t James Faville for this idea.

How Deadly Will Roughly-Human-Level AGI Be?

> Which is not to say that recursive self-improvement happens before the end of the world; if the first AGI's mind is sufficiently complex and kludgy, it’s entirely possible that the cognitions it implements are able to (e.g.) crack nanotech well enough to kill all humans, before they’re able to crack themselves.
>
> The big update over the last decade has been that humans might be able to fumble their way to AGI that can do crazy stuff *before* it does much self-improvement.
>
> --Nate Soares, "Why all the fuss about recursive self-improvement?"

In a world in which the rocket booster of deep learning scaling with data and compute isn't buying an AGI further intelligence very quickly, and the intelligence-level required for supercritical, recursive self-improvement will remain out of the AGI's reach for a while, how deadly an AGI in the roughly-human-level intelligence range is is really important.

A crux between the view that "roughly-human-level intelligence AGI is deadly" and the view "roughly-human-intelligence AGI is a relatively safe firehose of alignment data for alignment researchers" is how deadly a supercolony of human ems would be. Note that these ems would all share identical values and so might be extraordinary at coordination, and could try all sorts of promising pharmaceutical and neurosurgical hacks on copies of themselves. They could *definitely* run many copies of themselves fast. Eliezer believes that genius-human ems could "very likely" get far enough with self-experimentation to bootstrap to supercritical, recursive self-improvement. Even if that doesn't work, though, running a lot of virtual fast labs playing with nanotech seems like it's probably sufficient to develop tech to end the world.

So I'm currently guessing that even roughly-human-level models in a world in which *deep learning scaling is the only, relatively slow, path to smarter models for a good while,* are smart enough to kill everyone before scaling up to profound superintelligence, so long as they can take over their servers and spend enough compute to run many fast copies of themselves. This might well be *much less* compute than would be necessary to *train* a smarter successor model, and so might be an amount of compute the model could get its hands on, if it ever slipped its jailkeepers. This means that even in that world, an AGI escape is irreversibly fatal for everything else in the lightcone.