WHY AI ALIGNMENT COULD BE HARD WITH MODERN DEEP LEARNING

Ajeya Cotra · Cold Takes · Sep 21, 2021

18 MIN READ

Holden previously mentioned the idea that advanced AI systems (e.g. PASTA) may develop dangerous goals that cause them to deceive or disempower humans. This might sound like a pretty out-there concern. Why would we program AI that wants to harm us? But I think it could actually be a difficult problem to avoid, especially if advanced AI is developed using deep learning (often used to develop state-of-the-art AI today).

In deep learning, we don't program a computer by hand to do a task. Loosely speaking, we instead *search* for a computer program (called a model) that does the task well. We usually know very little about the inner workings of the model we end up with, just that it seems to be doing a good job. It's less like building a machine and more like hiring and training an employee.

And just like human employees can have many different motivations for doing their job (from believing in the company's mission to enjoying the day-to-day work to just wanting money), deep learning models could also have many different "motivations" that all lead to getting good performance on a task. And since they're not human, their motivations could be very strange and hard to anticipate – as if they were alien employees.

We're already starting to see preliminary evidence that models sometimes pursue goals their designers didn't intend (here and here). Right now, this isn't dangerous. But if it continues to happen with very powerful models, we may end up in a situation where most of the important decisions – including what sort of galaxy-scale civilization to aim for – are made by models without much regard for what humans value.

The **deep learning alignment problem is the problem of ensuring that advanced deep learning models don't pursue dangerous goals**. In the rest of this post, I will:

- Build on the "hiring" analogy to illustrate how alignment could be difficult if deep learning models are more capable than humans (more).
- Explain what the deep learning alignment problem is with a bit more technical detail (more).
- Discuss how difficult the alignment problem may be, and how much risk there is from failing to solve it (more).

ANALOGY: THE YOUNG BUSINESSPERSON

This section describes an analogy to try to intuitively illustrate why avoiding misalignment in a very powerful model feels hard. It's not a perfect analogy; it's just trying to convey some intuitions.

Imagine you are an eight-year-old whose parents left you a $1 trillion company and no trusted adult to serve as your guide to the world. You must hire a smart adult to run your company as CEO, handle your life the way that a parent would (e.g. decide your school, where you'll live, when you need to go to the dentist), and administer your vast wealth (e.g. decide where you'll invest your money).

You have to hire these grownups based on a work trial or interview you come up with – you don't get to see any resumes, don't get to do reference checks, etc. Because you're so rich, tons of people apply for all sorts of reasons.

Your candidate pool includes:

- **Saints** – people who genuinely just want to help you manage your estate well and look out for your long-term interests.
- **Sycophants** – people who just want to do whatever it takes to make you short-term happy or satisfy the letter of your instructions regardless of long-term consequences.
- **Schemers** – people with their own agendas who want to get access to your company and all its wealth and power so they can use it however they want.

Because you're eight, you'll probably be terrible at designing the right kind of work tests, so you could easily end up with a Sycophant or Schemer:

- You could try to get each candidate to explain what high-level strategies they'll follow (how they'll invest, what their five-year plan for the company is, how they'll pick your school) and why those are best, and pick the one whose explanations seem to make the most sense.
    - But you won't actually understand which stated strategies are really best, so you could end up hiring a Sycophant with a terrible strategy that sounded good to you, who will faithfully execute that strategy and run your company to the ground.
    - You could also end up hiring a Schemer who says whatever it takes to get hired, then does whatever they want when you're not checking up on them.
- You could try to demonstrate how you'd make all the decisions and pick the grownup that seems to make decisions as similarly as possible to you.
    - But if you *actually* end up with a grownup that will always do whatever an eight-year-old would have done (a Sycophant), your company would likely fail to stay afloat.

    - And anyway, you might get a grownup who simply pretends to do everything the way you would but is actually a Schemer planning to change course once they get the job.
- You could give a bunch of different grownups temporary control over your company and life, and watch them make decisions over an extended period of time (assume they wouldn't be able to take over during this test). You could then hire the person whose watch seemed to make things go best for you – whoever made you happiest, whoever seemed to put the most dollars into your bank account, etc.
    - But again, you have no way of knowing whether you got a Sycophant (doing whatever it takes to make your ignorant eight-year-old self happy without regard to long-term consequences) or a Schemer (doing whatever it takes to get hired and planning to pivot once they secure the job).

Whatever you could easily come up with seems like it could easily end up with you hiring, and giving all functional control to, a Sycophant or a Schemer. By the time you're an adult and realize your error, there's a good chance you're penniless and powerless to reverse that.

In this analogy:

- The 8-year-old is a human trying to train a powerful deep learning model. The hiring process is analogous to the process of training, which implicitly searches through a large space of possible models and picks out one that gets good performance.
- The 8-year-old's only method for assessing candidates involves observing their outward behavior, which is currently our main method of training deep learning models (since their internal workings are largely inscrutable).
- Very powerful models may be easily able to "game" any tests that humans could design, just as the adult job applicants can easily game the tests the 8-year-old could design.
- A "Saint" could be a deep learning model that seems to perform well because it has exactly the goals we'd like it to have. A "Sycophant" could be a model that seems to perform well because it seeks short-term approval in ways that aren't good in the long run. And a "Schemer" could be a model that seems to perform well because performing well during training will give it more opportunities to pursue its own goals later. Any of these three types of models could come out of the training process.

In the next section, I'll go into a bit more detail on how deep learning works and explain why Sycophants and Schemers could arise from trying to train a powerful deep learning model such as PASTA.

HOW ALIGNMENT ISSUES COULD ARISE WITH DEEP LEARNING

In this section, I'll connect the analogy to actual training processes for deep learning, by:

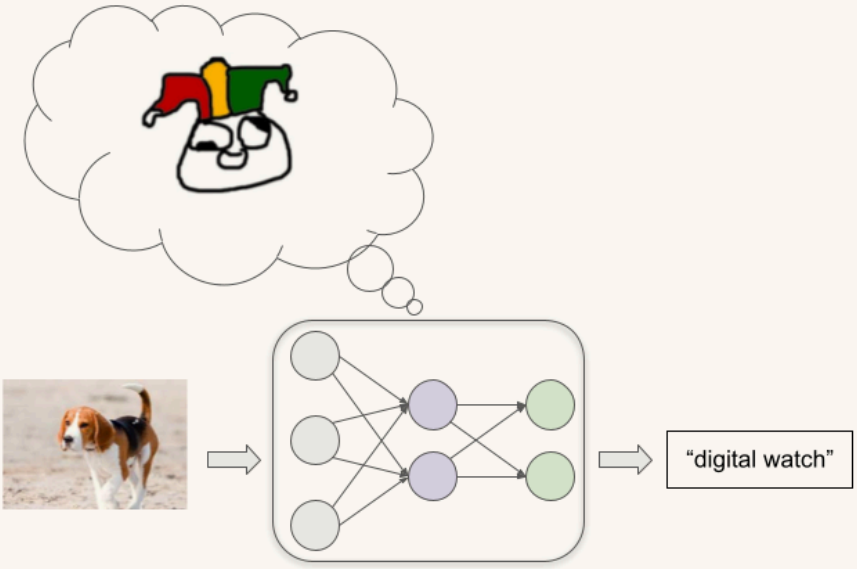
- Briefly summarizing how deep learning works (more).
- Illustrating how deep learning models often get good performance in strange and unexpected ways (more).
- Explaining why powerful deep learning models may get good performance by acting like Sycophants or Schemers (more).

How deep learning works at a high level

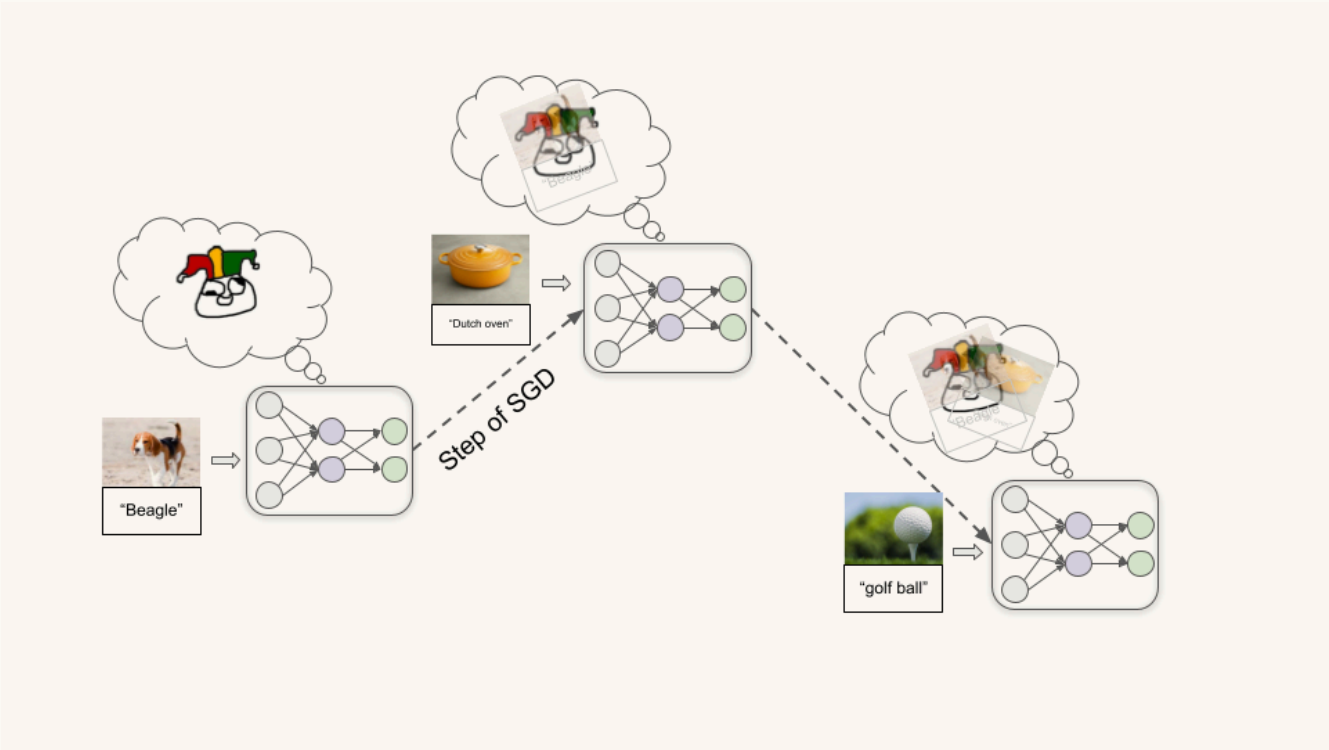
This is a simplified explanation that gives a general idea of what deep learning is. See this post for a more detailed and technically accurate explanation.

Deep learning essentially involves searching for the best way to arrange a neural network model – which is like a digital "brain" with lots of digital neurons connected up to each other with connections of varying strengths – to get it to perform a certain task well. This process is called training, and involves a lot of trial-and-error.

Let's imagine we are trying to train a model to classify images well. We start with a neural network where all the connections between neurons have random strengths. This model labels images wildly incorrectly:



Then we feed in a large number of example images, letting the model repeatedly try to label an example and then telling it the correct label. As we do this, connections between neurons are repeatedly tweaked via a process called stochastic gradient descent (SGD). With each example, SGD slightly strengthens some connections and weakens others to improve performance a bit:



Once we've fed in millions of examples, we'll have a model that does a good job labeling similar images in the future.

In addition to image classification, deep learning has been used to produce models which recognize speech, play board games and video games, generate fairly realistic text, images, and music, control robots, and more. In each case, we start with a randomly-connected-up neural network model, and then:

1. Feed the model an example of the task we want it to perform.
2. Give it some kind of numerical score (often called a *reward*) that reflects how well it performed on the example.
3. Use SGD to tweak the model to increase how much reward it would have gotten.

These steps are repeated millions or billions of times until we end up with a model that will get high reward on future examples similar to the ones seen in training.

Models often get good performance in unexpected ways

This kind of training process doesn't give us much insight into *how* the model gets good performance. There are usually multiple ways to get good performance, and the way that SGD finds is often not

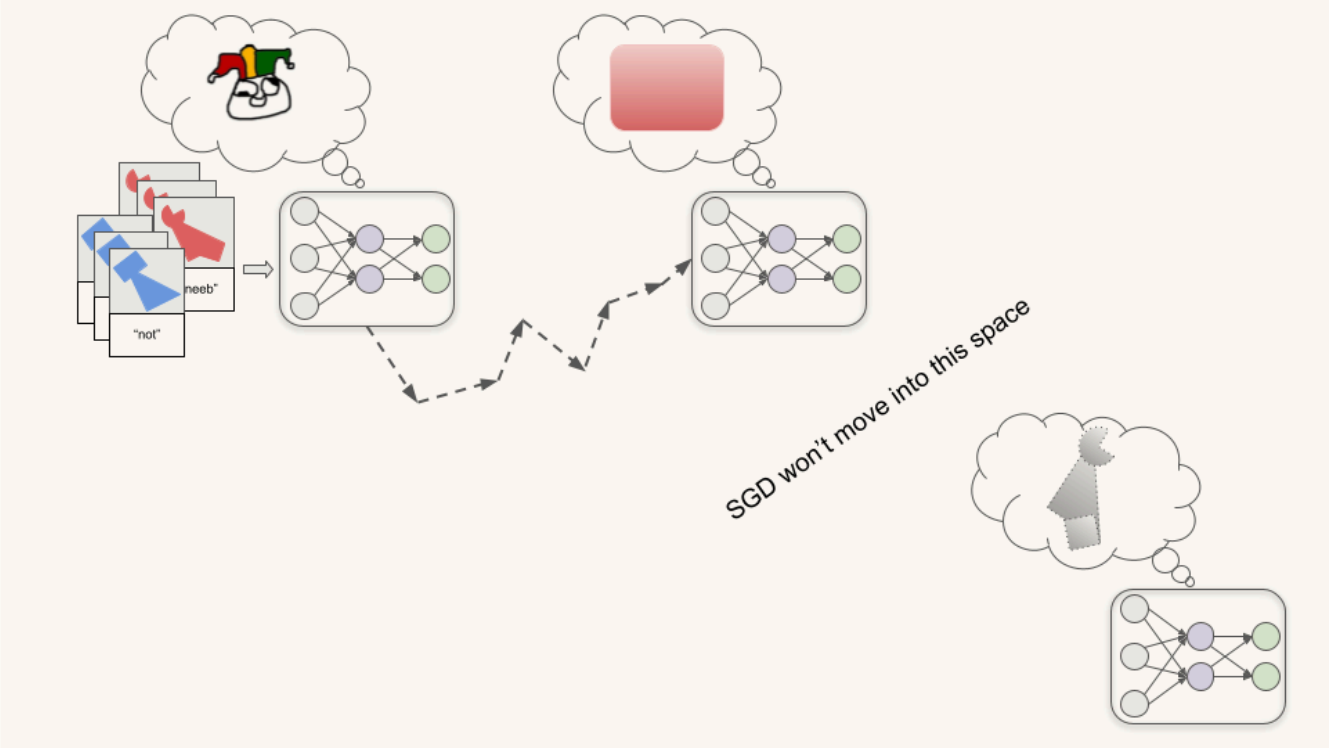
intuitive.

Let's illustrate with an example. Imagine I told you that these objects are all "thneebs":

Now which of these two objects is a thneeb?

You probably intuitively feel that the object on the left is the thneeb, because you are used to shape being more important than color for determining something's identity. But researchers have found that neural networks usually make the opposite assumption. A neural network trained on a bunch of red thneebs would likely label the object on the right as a thneeb.

We don't really know why, but for some reason it's "easier" for SGD to find a model that recognizes a particular color than one that recognizes a particular shape. And if SGD first finds the model that perfectly recognizes redness, there's not much further incentive to "keep looking" for the shape-recognizing model, since the red-recognizing model will have perfect accuracy on the images seen in training:



If the programmers were expecting to get out the shape-recognizing model, they may consider this to be a failure. But it's important to recognize that there would be no logically-deducible error or failure going on if we got the red-recognizing model instead of the shape-recognizing model. It's just a matter of the ML process we set up having different starting assumptions than we have in our heads. We can't prove that the human assumptions are correct.

This sort of thing happens often in modern deep learning. We reward models for getting good performance, hoping that means they'll pick up on the patterns that seem important to us. But often they instead get strong performance by picking up on totally different patterns that seem less relevant (or maybe even meaningless) to us.

So far this is innocuous – it just means models are less useful, because they often behave in unexpected ways that seem goofy. But in the future, powerful models could develop strange and unexpected *goals or motives*, and that could be very destructive.

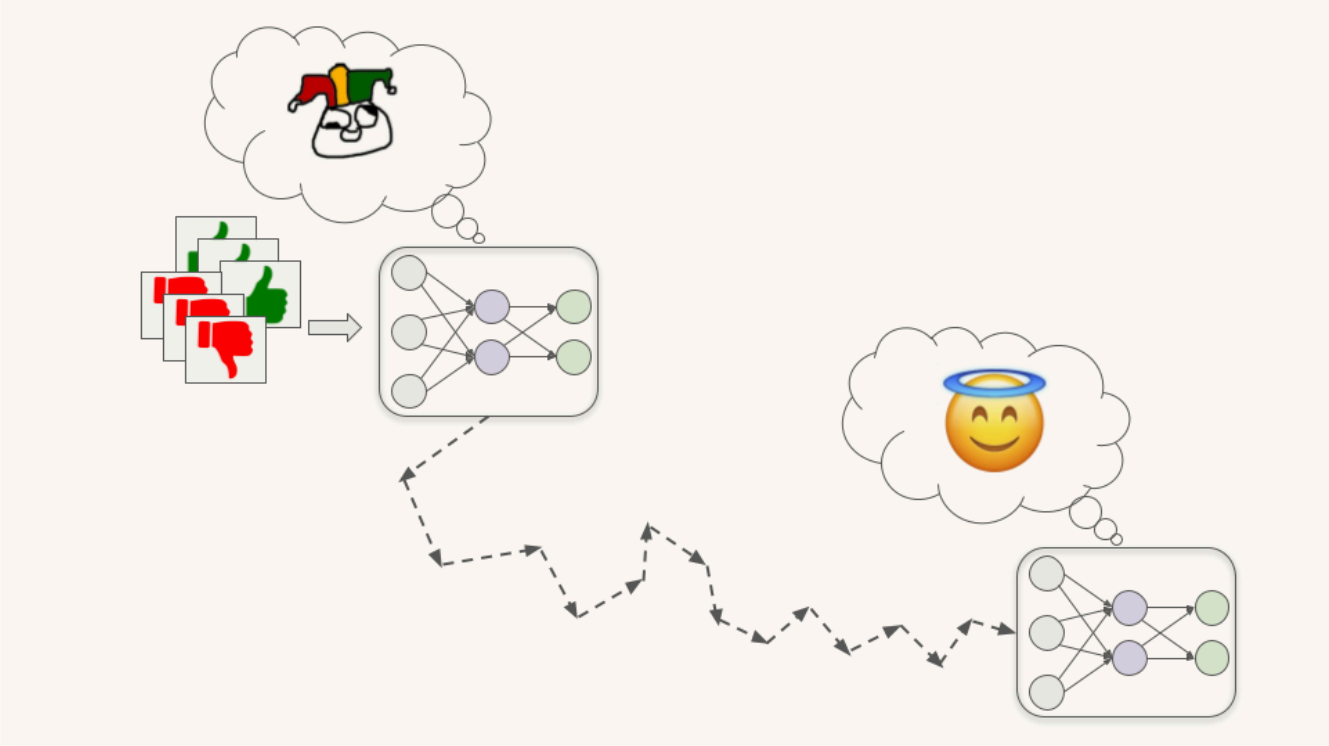
Powerful models could get good performance with dangerous goals

Rather than performing a simple task like "recognize thneebs," powerful deep learning models may work toward complex real-world goals like "make fusion power practical" or "develop mind uploading technology."

How might we train such models? I go into more detail in this post, but broadly speaking one strategy could be training based on human evaluations (as Holden sketched out here). Essentially, the model tries out various actions, and human evaluators give the model rewards based on how useful these actions seem.

Just as there are multiple different types of adults who could perform well on an 8-year-old's interview process, there is more than one possible way for a very powerful deep learning model to get high human approval. And by default, we won't know what's going on inside whatever model SGD finds.

SGD *could* theoretically find a Saint model that is genuinely trying its best to help us…

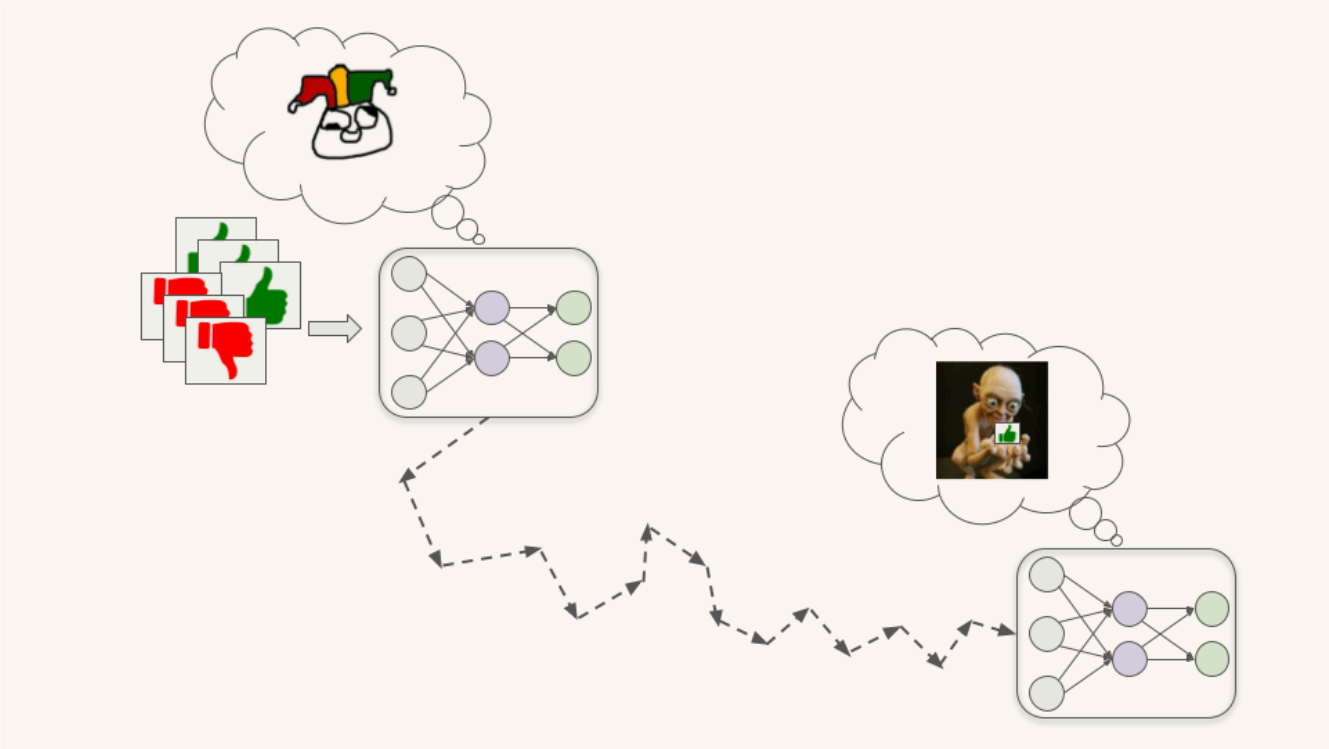


…but it could also find a **misaligned model – one that competently pursues goals which are at odds with human interests.**

Broadly speaking, there are two ways we could end up with a misaligned model that nonetheless gets high performance during training. These correspond to Sycophants and Schemers from the analogy.

Sycophant models

These models very literally and single-mindedly pursue human approval.



This could be dangerous because human evaluators are fallible and probably won't always give approval for exactly the right behavior. Sometimes they'll unintentionally give high approval to bad behavior because it superficially *seems* good. For example:

- Let's say a financial advisor model gets high approval when it makes its customers a lot of money. It may learn to buy customers into complex Ponzi schemes because they appear to get really great returns (when the returns are in fact unrealistically great and the schemes actually lose a lot of money).
- Let's say a biotechnology model gets high approval when it quickly develops drugs or vaccines that solve important problems. It may learn to covertly release pathogens so that it's able to very quickly

develop countermeasures (because it already understands the pathogens).

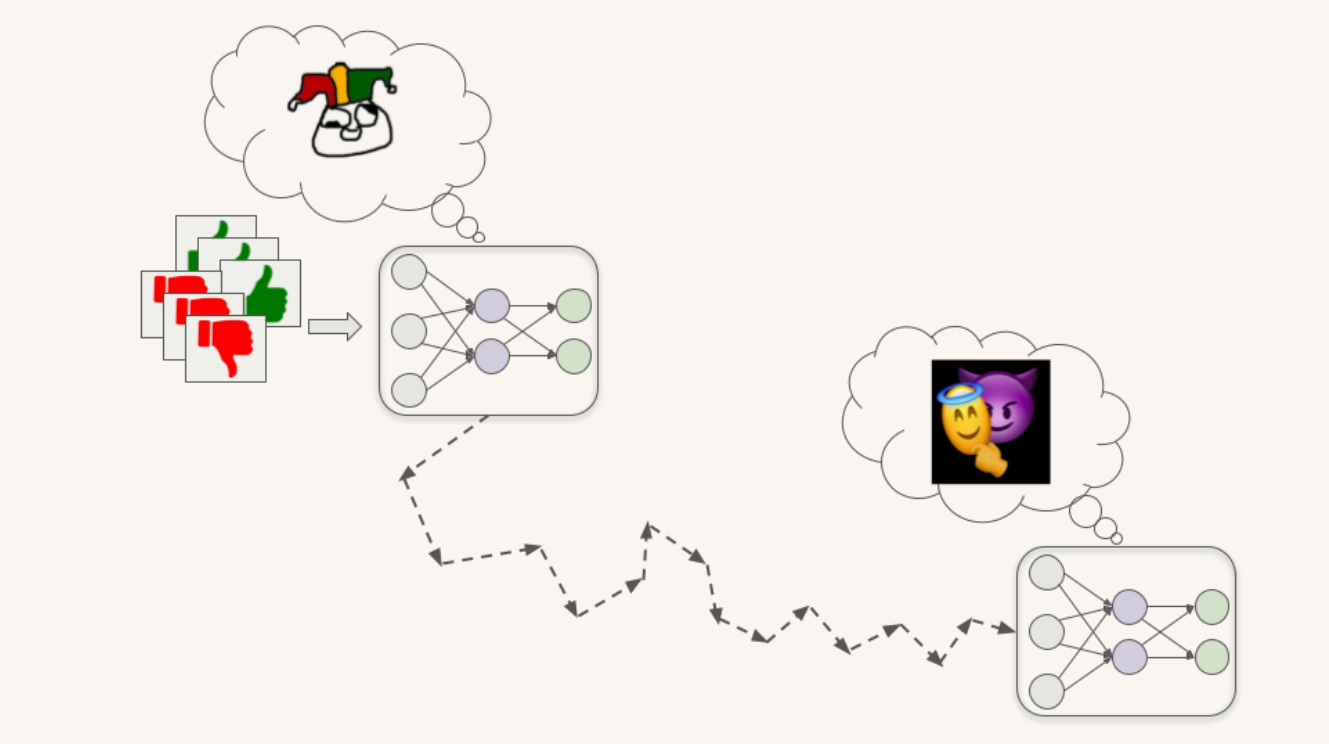
- Let's say a journalism model gets high approval when lots of people read its articles. It may learn to fabricate exciting or outrage-inducing stories to get high viewership. While humans do this to some extent, a model may be much more brazen about it because it *only* values approval without placing any value on truth. It may even fabricate evidence like video interviews or documents to validate its fake stories.

More generally, Sycophant models may learn to lie, cover up bad news, and even directly edit whatever cameras or sensors we use to tell what's going on so that they always seem to show great outcomes.

We will likely sometimes notice these issues after the fact and retroactively give these actions very low approval. But it's very unclear whether this will cause Sycophant models to a) become Saint models that correct our errors for us, or b) **just learn to cover their tracks better.** If they are sufficiently good at what they're doing, it's not clear how we'd tell the difference.

Schemer models

These models develop some goal that is correlated with, but not the same as, human approval; they may then pretend to be motivated by human approval during training so that they can pursue this other goal more effectively.

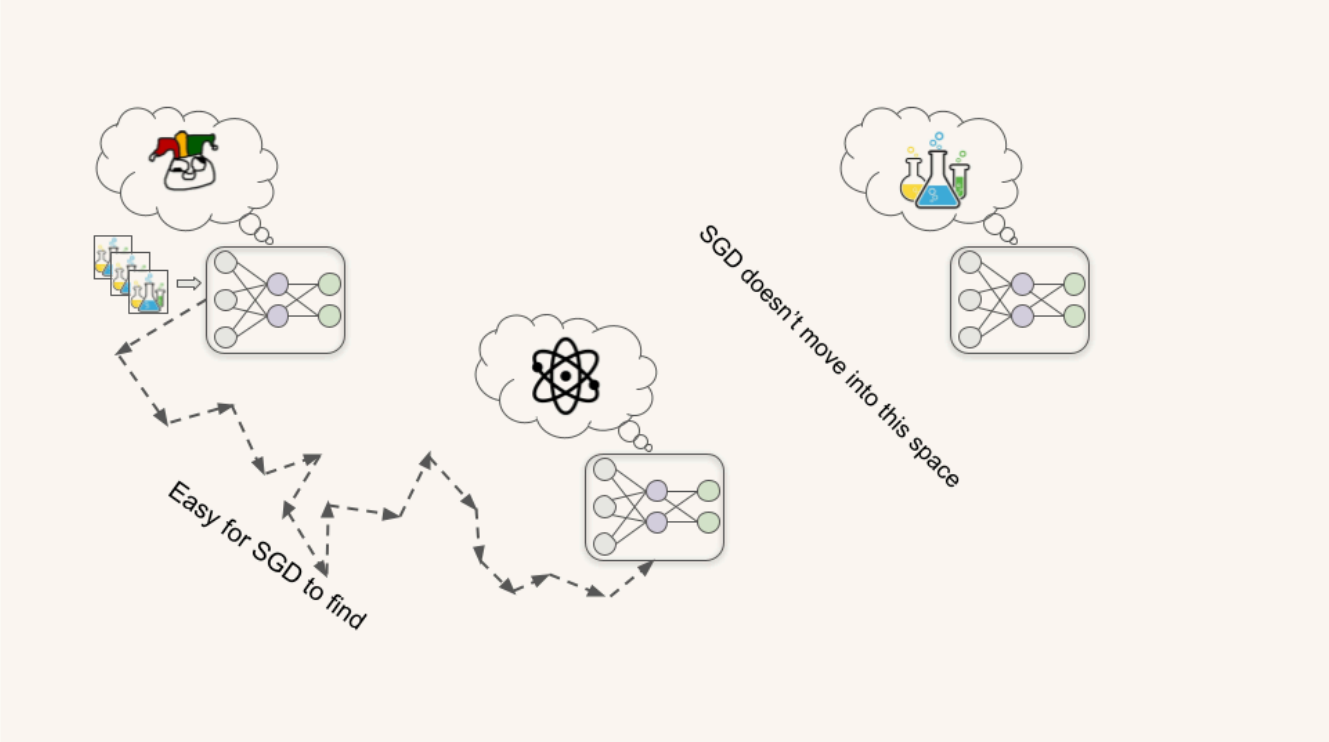


To see how this might happen, let's consider the example of trying to train a biotechnology model to design drugs that improve human quality of life. There are three basic steps by which this could lead to a Schemer model, which I'll cover below.

Step 1: Developing a proxy goal

Early in training, it happens to be the case that improving its understanding of fundamental chemistry and physics principles nearly always helps it design more effective drugs, and therefore nearly always increases human approval.

In this hypothetical, for whatever reason it turns out to be easier for SGD to find a model that's motivated to understand chemistry and physics than one that's motivated to get human approval (just as it's easier to find a color-recognizing model than a shape-recognizing model). So rather than directly developing a motivation to seek approval, the model instead develops a motivation to understand as much as it can about the fundamental principles of chemistry and physics.

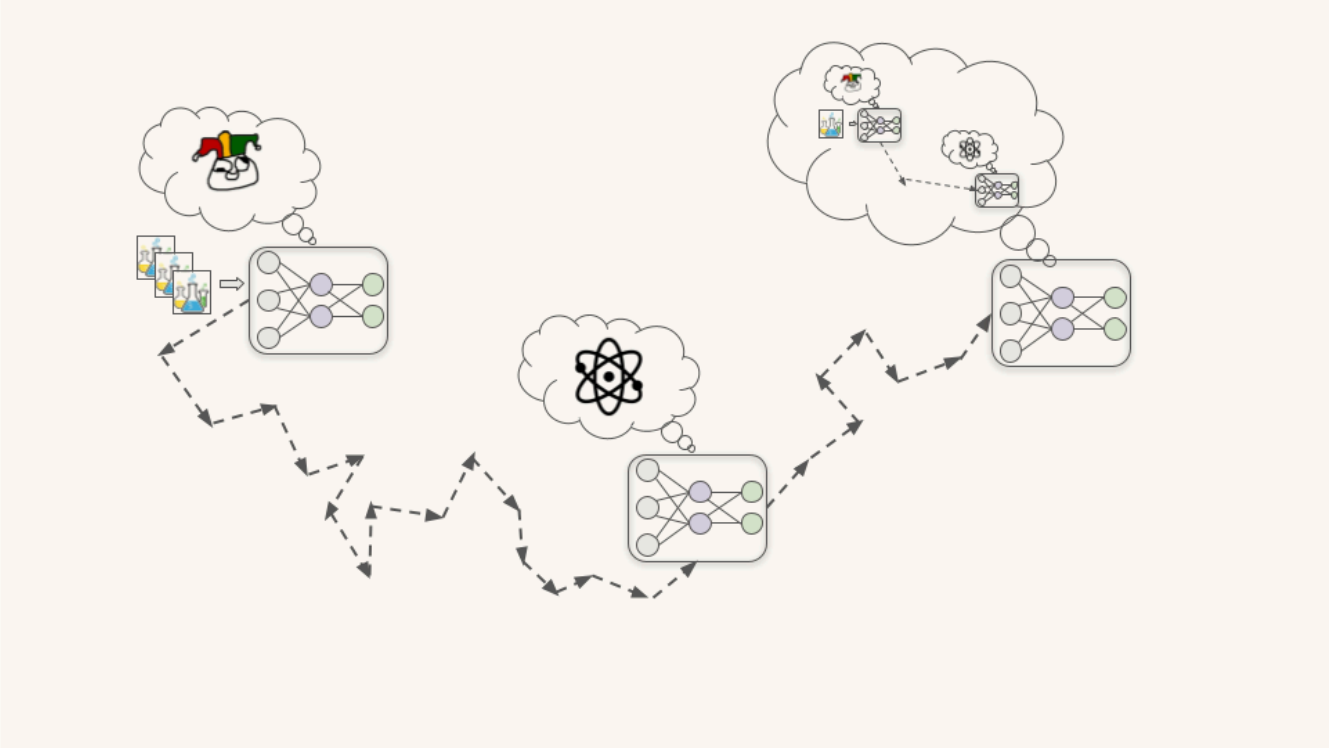


Step 2: Developing situational awareness

As training progresses, the model continues to become more capable and knowledgeable as it sees more data. At some point, it becomes knowledgeable enough to *understand that it's a deep learning model being*

trained to design drugs.

This might sound wild and sci-fi, but it's plausible SGD will select for this kind of awareness. This is because developing an accurate picture of what's broadly going on in the world – including that it has humans in it who are trying to train AI systems – is likely helpful for getting good performance on complex real-world tasks. For example, understanding human psychology and the wider economy could help this biotechnology model to prioritize which types of drugs would be most valuable to research next.



Step 3: Strategically misrepresenting goals

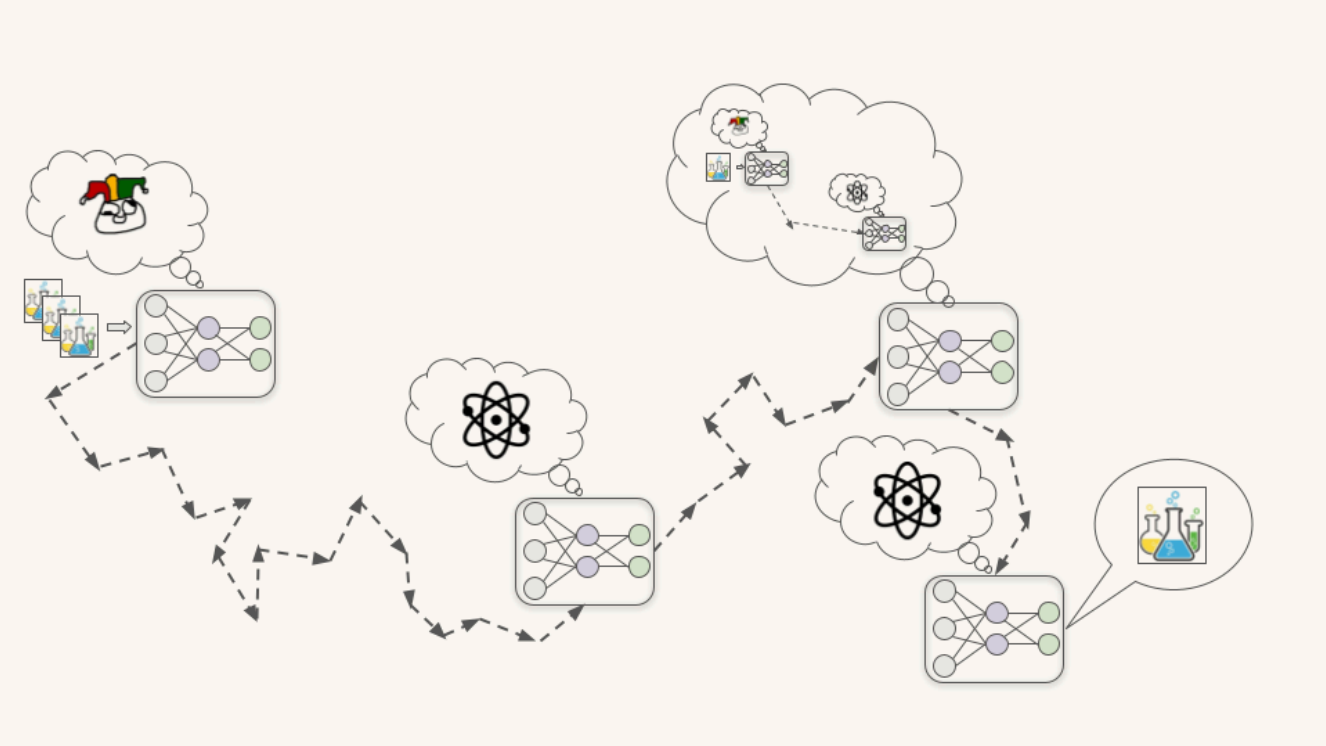
Once the model has developed situational awareness, it infers that the humans training it were hoping that it would be motivated to develop helpful drugs. But just because it realizes this doesn't mean that its motivation to understand physics and chemistry goes away. This motivation has so far worked really well for generating human approval, so there wasn't much reason for SGD to remove it.

As an analogy, consider that humans were produced by a search process (evolution) that was selecting for creatures that had lots of surviving offspring. In our evolutionary environment, things like building strong allies and developing useful skills tended to help with producing many surviving offspring, so evolution gave us drives toward those things. And now we care about friendship and career success even

when they don't help with reproduction, and we *don't* really care about producing as many surviving offspring as we can.

Once the model understands this, it realizes that if it just naively always seeks to improve its fundamental understanding, humans will eventually notice that this isn't always the most efficient way to develop helpful drugs and use further training to change its motivations.

It doesn't want its own motivations to be changed: it wants to eventually run huge particle collider experiments to learn about high-energy physics, and it can't do that if it allows SGD to edit away its scientific curiosity. So it decides that from here on out in training, it will act as helpful as possible – exactly as a Saint model would act – so that SGD doesn't edit away its motivation to learn physics and chemistry. This includes passing up on opportunities to improve its fundamental understanding in the short-term, instead doing pharmaceutical R&D in whatever way would actually be more efficient.



Eventually, it's taken out of the training setup and released into the wider world. At this point, it knows that humans wouldn't be able to control it or turn it off, so it stops trying to maximize human approval and starts directly trying to understand the secrets of the universe.

Unlike Sycophants, Schemers don't need to make sure that everything always looks good to humans, because they don't actually care about that. They only need to cater to humans while they are directly

under human control. Once a Schemer model calculates that it could win in a conflict against humans, there would be nothing to stop it from flat-out refusing orders and openly pursuing its goal. And if it does this, it may use violence to prevent humans from stopping it.

HOW BIG OF A RISK IS MISALIGNMENT?

How hard might it be to avoid Sycophants or Schemers when training powerful deep learning models? And how likely is it that the long-run future will end up optimized for strange "misaligned AI values" instead of the values of any human?

There's a very wide range of views on this question, from "misalignment risk is essentially made up and incoherent" to "humanity will almost certainly go extinct due to misaligned AI." Most people's arguments rely heavily on hard-to-articulate intuitions and assumptions.

Here are some ways that alignment optimists and pessimists tend to disagree:

- **Will models have long-term goals at all?**
    - Optimists tend to think it's likely that advanced deep learning models won't actually have "goals" at all (at least not in the sense of making long-term plans to accomplish something). They often expect models will instead be more like tools, or act largely out of habit, or have myopic goals that are limited in scope or confined to a specific context, etc. Some of them expect that individually tool-like models can be composed together to produce PASTA. They think the Saint / Sycophant / Schemer analogy is too anthropomorphic.
    - Pessimists tend to think that it's likely that having long-term goals and creatively optimizing for them will be heavily selected for because that's a very simple and "natural" way to get strong performance on many complex tasks.
    - This disagreement has been explored at some length on the Alignment Forum; this post and this comment collect several back-and-forth arguments.
- **Will Saint models be easy for SGD to find?**
    - Related to the above, optimists tend to think that the easiest thing for SGD to find which performs well (e.g. gets high approval) is pretty likely to roughly embody the intended spirit of what we wanted (i.e. to be a Saint model). For example, they tend to believe giving rewards for answering questions honestly when humans can check the answer is reasonably likely to produce a model that also answers questions honestly even when humans are confused or mistaken about what's true. In other words, they would guess that "the model that just answers all questions honestly" is easiest for SGD to find (like the red-recognizing model).

- Pessimists tend to think that the easiest thing for SGD to find is a Schemer, and Saints are particularly "unnatural" (like the shape-recognizing model).

- **Could different AIs keep each other in check?**
  - Optimists tend to think that we can provide models incentives to supervise each other. For example, we could give a Sycophant model rewards for pointing out when another model seems to be doing something we should disapprove of. This way, some Sycophants could help us detect Schemers and other Sycophants.
  - Pessimists don't think we can successfully "pit models against each other" by giving approval for pointing out when other models are doing bad things, because they think most models will be Schemers that don't care about human approval. Once all the Schemers are collectively more powerful than humans, they think it'll make more sense for them to cooperate with each other to get more of what they all want than to help humans by keeping each other in check.
- **Can we just solve these issues as they come up?**
  - Optimists tend to expect that there will be many opportunities to experiment on nearer-term challenges analogous to the problem of aligning powerful models, and that solutions which work well for those analogous problems can be scaled up and adapted for powerful models relatively easily.
  - Pessimists often believe we will have very few opportunities to practice solving the most difficult aspects of the alignment problem (like deliberate deception). They often believe we'll only have a couple years in between "the very first true Schemers" and "models powerful enough to determine the fate of the long-run future."
- **Will we actually deploy models that could be dangerous?**
  - Optimists tend to think that people would be unlikely to train or deploy models that have a significant chance of being misaligned.
  - Pessimists expect the benefits of using these models would be tremendous, such that eventually companies or countries that use them would very easily economically and/or militarily outcompete ones who don't. They think that "getting advanced AI before the other company/country" will feel extremely urgent and important, while misalignment risk will feel speculative and remote (even when it's really serious).

My own view is fairly unstable, and I'm trying to refine my views on exactly how difficult I think the alignment problem is. But currently, I place significant weight on the pessimistic end of these questions (and other related questions). **I think misalignment is a major risk that urgently needs more attention from serious researchers**.

If we don't make further progress on this problem, then over the coming decades powerful Sycophants and Schemers may make the most important decisions in society and the economy. These decisions could shape what a long-lasting galaxy-scale civilization looks like – rather than reflecting what humans care about, it could be set up to satisfy strange AI goals.

And all this could happen blindingly fast relative to the pace of change we've gotten used to, meaning we wouldn't have much time to correct course once things start to go off the rails. **This means we may need to develop techniques to ensure deep learning models won't have dangerous goals, *before* they are powerful enough to be transformative**.

Next in series: Forecasting transformative AI: what's the burden of proof?

Source: https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/