# The Urgency of Interpretability

In the decade that I have been working on AI, I've watched it grow from a tiny academic field to arguably the most important economic and geopolitical issue in the world. In all that time, perhaps the most important lesson I've learned is this: the progress of the underlying technology is inexorable, driven by forces too powerful to stop, but the *way* in which it happens—the order in which things are built, the applications we choose, and the details of how it is rolled out to society—are eminently possible to change, and it's possible to have great positive impact by doing so. We can't *stop* the bus, but we can *steer* it. In the past I've written about the importance of deploying AI in a way that is positive for the world, and of ensuring that democracies build and wield the technology before autocracies do. **Over the last few months, I have become increasingly focused on an additional opportunity for steering the bus: the tantalizing possibility, opened up by some recent advances, that we could succeed at *interpretability*—that is, in understanding the inner workings of AI systems—*before* models reach an overwhelming level of power.**

People outside the field are often surprised and alarmed to learn that we do not understand how our own AI creations work. They are right to be concerned: this lack of understanding is essentially unprecedented in the history of technology. For several years, we (both Anthropic and the field at large) have been trying to solve this problem, to create the analogue of a highly precise and accurate MRI that would fully reveal the inner workings of an AI model. This goal has often felt very distant, but multiple recent breakthroughs have convinced me that we are now on the right track and have a real chance of success.

At the same time, the field of AI as a whole is further ahead than our efforts

at interpretability, and is itself advancing very quickly. We therefore must move fast if we want interpretability to mature in time to matter. This post makes the case for interpretability: what it is, why AI will go better if we have it, and what all of us can do to help it win the race.

## The Dangers of Ignorance

Modern generative AI systems are opaque in a way that fundamentally differs from traditional software. If an ordinary software program does something—for example, a character in a video game says a line of dialogue, or my food delivery app allows me to tip my driver—it does those things because a human specifically programmed them in. Generative AI is *not like that at all*. When a generative AI system does something, like summarize a financial document, we have no idea, at a specific or precise level, why it makes the choices it does—why it chooses certain words over others, or why it occasionally makes a mistake despite usually being accurate. As my friend and co-founder Chris Olah is fond of saying, generative AI systems are *grown* more than they are *built*—their internal mechanisms are "emergent" rather than directly designed. It's a bit like growing a plant or a bacterial colony: we set the high-level conditions that direct and shape growth[1], but the exact structure which emerges is unpredictable and difficult to understand or explain. Looking inside these systems, what we see are vast matrices of billions of numbers. These are *somehow* computing important cognitive tasks, but exactly how they do so isn't obvious.

**Many of the risks and worries associated with generative AI are ultimately consequences of this opacity, and would be much easier to address if the models were interpretable.** For example, AI researchers often worry about misaligned systems that could take harmful actions not intended by their creators. Our inability to understand models' internal mechanisms means that we cannot meaningfully predict such behaviors,

and therefore struggle to rule them out; indeed, models *do* exhibit unexpected emergent behaviors, though none that have yet risen to major levels of concern. More subtly, the same opacity makes it hard to find definitive evidence *supporting* the existence of these risks at a large scale, making it hard to rally support for addressing them—and indeed, hard to know for sure how dangerous they are.

To address the severity of these alignment risks, we will have to see inside AI models much more clearly than we can today. For example, one major concern is AI deception or power-seeking. The nature of AI training makes it possible that AI systems will develop, on their own, an ability to deceive humans and an inclination to seek power in a way that ordinary deterministic software never will; this emergent nature also makes it difficult to detect and mitigate such developments[2]. But by the same token, we've never seen any solid evidence in truly real-world scenarios of deception and power-seeking[3] because we can't "catch the models red-handed" thinking power-hungry, deceitful thoughts. What we're left with is vague theoretical arguments that deceit or power-seeking might have the incentive to emerge during the training process, which some people find thoroughly compelling and others laughably unconvincing. Honestly I can sympathize with both reactions, and this might be a clue as to why the debate over this risk has become so polarized.

Similarly, worries about misuse of AI models—for example, that they might help malicious users to produce biological or cyber weapons, in ways that go beyond the information that can be found on today's internet—are based[4] on the idea that it is very difficult to reliably prevent the models from knowing dangerous information or from divulging what they know. We can put filters on the models, but there are a huge number of possible ways to "jailbreak" or trick the model, and the only way to discover the existence of a jailbreak is to find it empirically. If instead it were possible to look inside models, we

might be able to systematically block all jailbreaks, and also to characterize what dangerous knowledge the models have.

AI systems' opacity also means that they are simply not used in many applications, such as high-stakes financial or safety-critical settings, because we can't fully set the limits on their behavior, and a small number of mistakes could be very harmful. Better interpretability could greatly improve our ability to set bounds on the range of possible errors. In fact, for some applications, the fact that we can't see inside the models is literally a legal blocker to their adoption—for example in mortgage assessments where decisions are legally required to be explainable. Similarly, AI has made great strides in science, including improving the prediction of DNA and protein sequence data, but the patterns and structures predicted in this way are often difficult for humans to understand, and don't impart biological insight. Some research papers from the last few months have made it clear that interpretability can help us understand these patterns.

There are other more exotic consequences of opacity, such as that it inhibits our ability to judge whether AI systems are (or may someday be) sentient and may be deserving of important rights. This is a complex enough topic that I won't get into it in detail, but I suspect it will be important in the future.[5]

## A Brief History of Mechanistic Interpretability

For all of the reasons described above, figuring out what the models are thinking and how they operate seems like a task of overriding importance. The conventional wisdom for decades was that this was impossible, and that the models were inscrutable "black boxes". I'm not going to be able to do justice[6] to the full story of how that changed, and my views are inevitably colored by what I saw personally at Google, OpenAI, and Anthropic. But

Chris Olah was one of the first to attempt a truly systematic research program to open the black box and understand all its pieces, a field that has come to be known as *mechanistic interpretability*. Chris worked on mechanistic interpretability first at Google, and then at OpenAI. When we founded Anthropic, we decided to make it a central part of the new company's direction and, crucially, focused it on LLM's. Over time the field has grown and now includes teams at several of the major AI companies as well as a few interpretability-focused companies, nonprofits, academics, and independent researchers. It's helpful to give a brief summary of what the field has accomplished so far, and what remains to be done if we want to apply mechanistic interpretability to address some of the key risks above.

The early era of mechanistic interpretability (2014-2020) focused on vision models, and was able to identify some neurons inside the models that represented human-understandable concepts, such as a "car detector" or a "wheel detector", similar to early neuroscience hypotheses and studies suggesting that the human brain has neurons corresponding to specific people or concepts, often popularized as the "Jennifer Aniston" neuron (and in fact, we found neurons much like those in AI models). We were even able to discover how these neurons are connected—for example, the car detector looks for wheel detectors firing below the car, and combines that with other visual signals to decide if the object it's looking at is indeed a car.

When Chris and I left to start Anthropic, we decided to apply interpretability to the emerging area of language, and in 2021 developed some of the basic mathematical foundations and software infrastructure necessary to do so. We immediately found some basic mechanisms in the model that did the kind of things that are essential to interpret language: copying and sequential pattern-matching. We also found some interpretable single neurons, similar to what we found in vision models, which represented various words and concepts. However, we quickly discovered that while

*some* neurons were immediately interpretable, the vast majority were an incoherent pastiche of many different words and concepts. We referred to this phenomenon as *superposition*,[7] and we quickly realized that the models likely contained billions of concepts, but in a hopelessly mixed-up fashion that we couldn't make any sense of. The model uses superposition because this allows it to express more concepts than it has neurons, enabling it to learn more. If superposition seems tangled and difficult to understand, that's because, as ever, the learning and operation of AI models are not optimized in the slightest to be legible to humans.

The difficulty of interpreting superpositions blocked progress for a while, but eventually we discovered (in parallel with others) that an existing technique from signal processing called *sparse autoencoders* could be used to find *combinations* of neurons that *did* correspond to cleaner, more human-understandable concepts. The concepts that these combinations of neurons could express were far more subtle than those of the single-layer neural network: they included the concept of "literally or figuratively hedging or hesitating", and the concept of "genres of music that express discontent". We called these concepts *features,* and used the sparse autoencoder method to map them in models of all sizes, including modern state-of-the-art models. For example, we were able to find over 30 million features in a medium-sized commercial model (Claude 3 Sonnet). Additionally, we employed a method called *autointerpretability*—which uses an AI system itself to analyze interpretability features—to scale the process of not just finding the features, but listing and identifying what they mean in human terms.

Finding and identifying 30 million features is a significant step forward, but we believe there may actually be a *billion* or more concepts in even a small model, so we've found only a small fraction of what is probably there, and work in this direction is ongoing. Bigger models, like those used in

Anthropic's most capable products, are more complicated still.

Once a feature is found, we can do more than just observe it in action—we can increase or decrease its importance in the neural network's processing. The MRI of interpretability can help us develop and refine interventions—almost like zapping a precise part of someone's brain. Most memorably, we used this method to create "Golden Gate Claude", a version of one of Anthropic's models where the "Golden Gate Bridge" feature was artificially amplified, causing the model to become obsessed with the bridge, bringing it up even in unrelated conversations.

Recently, we've moved onward from tracking and manipulating features to tracking and manipulating groups of features that we call "circuits". These circuits show the steps in a model's thinking: how concepts emerge from input words, how those concepts interact to form new concepts, and how those work within the model to generate actions. With circuits, we can "trace" the model's thinking. For example, if you ask the model "What is the capital of the state containing Dallas?", there is a "located within" circuit that causes the "Dallas" feature to trigger the firing of a "Texas" feature, and then a circuit that causes "Austin" to fire after "Texas" and "capital". Even though we've only found a small number of circuits through a manual process, we can already use them to see how a model reasons through problems—for example how it plans ahead for rhymes when writing poetry, and how it shares concepts across languages. We are working on ways to automate the finding of circuits, as we expect there are millions within a model that interact in complex ways.

## The Utility of Interpretability

All of this progress, while scientifically impressive, doesn't directly answer the question of how we can use interpretability to reduce the risks I listed

earlier. Suppose we have identified a bunch of concepts and circuits—suppose, even, that we know all of them, and we can understand and organize them much better than we can today. So what? How do we *use* all of it? There's still a gap from abstract theory to practical value.

To help close that gap, we've begun experimenting with using our interpretability methods to find and diagnose problems in models. Recently, we did an experiment where we had a "red team" deliberately introduce an alignment issue into a model (say, a tendency for the model to exploit a loophole in a task) and gave various "blue teams" the task of figuring out what was wrong with it. Multiple blue teams succeeded; of particular relevance here, some of them productively applied interpretability tools during the investigation. We still need to scale these methods, but the exercise helped us gain some practical experience using interpretability techniques to find and address flaws in our models.

Our long-run aspiration is to be able to look at a state-of-the-art model and essentially do a "brain scan": a checkup that has a high probability of identifying a wide range of issues including tendencies to lie or deceive, power-seeking, flaws in jailbreaks, cognitive strengths and weaknesses of the model as a whole, and much more. This would then be used in tandem with the various techniques for training and aligning models, a bit like how a doctor might do an MRI to diagnose a disease, then prescribe a drug to treat it, then do another MRI to see how the treatment is progressing, and so on[8]. It is likely that a key part of how we will test and deploy the most capable models (for example, those at AI Safety Level 4 in our Responsible Scaling Policy framework) is by performing and formalizing such tests.

## What We Can Do

On one hand, recent progress—especially the results on circuits and on

interpretability-based testing of models—has made me feel that we are on the verge of cracking interpretability in a big way. Although the task ahead of us is Herculean, I can see a realistic path towards interpretability being a sophisticated and reliable way to diagnose problems in even very advanced AI—a true "MRI for AI". In fact, on its current trajectory I would bet strongly in favor of interpretability reaching this point within 5-10 years.

On the other hand, I worry that AI itself is advancing so quickly that we might not have even this much time. As I've written elsewhere, we could have AI systems equivalent to a "country of geniuses in a datacenter" as soon as 2026 or 2027. I am very concerned about deploying such systems without a better handle on interpretability. These systems will be absolutely central to the economy, technology, and national security, and will be capable of so much autonomy that **I consider it basically unacceptable for humanity to be totally ignorant of how they work.**

We are thus in a race between interpretability and model intelligence. It is not an all-or-nothing matter: as we've seen, every advance in interpretability quantitatively increases our ability to look inside models and diagnose their problems. The more such advances we have, the greater the likelihood that the "country of geniuses in a datacenter" goes well. There are several things that AI companies, researchers, governments, and society can do to tip the scales:

First, AI researchers in companies, academia, or nonprofits can **accelerate interpretability by directly working on it**. Interpretability gets less attention than the constant deluge of model releases, but it is arguably more important. It also feels to me like it is an ideal time to join the field: the recent "circuits" results have opened up many directions in parallel. Anthropic is doubling down on interpretability, and we have a goal of getting to "interpretability can reliably detect most model problems" by 2027. We are

also investing in interpretability startups.

But the chances of succeeding at this are greater if it is an effort that spans the whole scientific community. Other companies, such as Google DeepMind and OpenAI, have some interpretability efforts, but I strongly encourage them to allocate more resources. If it helps, Anthropic will be trying to apply interpretability commercially to create a unique advantage, especially in industries where the ability to provide an explanation for decisions is at a premium. If you are a competitor and you don't want this to happen, you too should invest more in interpretability!

Interpretability is also a natural fit for academic and independent researchers: it has the flavor of basic science, and many parts of it can be studied without needing huge computational resources. To be clear, some independent researchers and academics do work on interpretability, but we need many more[9]. Finally, if you are in another scientific field and are looking for new opportunities, interpretability may be a promising bet, as it offers rich data, exciting burgeoning methods, and enormous real-world value. Neuroscientists especially should consider this, as it's much easier to collect data on artificial neural networks than biological ones, and some of the conclusions can be applied back to neuroscience. If you're interested in joining Anthropic's Interpretability team, we have open Research Scientist and Research Engineer roles.

Second, governments can **use light-touch rules to encourage the development of interpretability research** and its application to addressing problems with frontier AI models. Given how nascent and undeveloped the practice of "AI MRI" is, it should be clear why it doesn't make sense to regulate or mandate that companies conduct them, at least at this stage: it's not even clear *what* a prospective law should ask companies to do. But a requirement for companies to transparently disclose their safety and security

practices (their Responsible Scaling Policy, or RSP, and its execution), including how they're using interpretability to test models before release, would allow companies to learn from each other while also making clear who is behaving more responsibly, fostering a "race to the top". We've suggested safety/security/RSP transparency as a possible direction for California law in our response to the California frontier model task force (which itself mentions some of the same ideas). This concept could also be exported federally, or to other countries.

Third, governments can use **export controls to create a "security buffer" that might give interpretability more time** to advance before we reach the most powerful AI. I've long been a proponent of export controls on chips to China because I believe that democratic countries must remain ahead of autocracies in AI. But these policies also have an additional benefit. If the US and other democracies have a clear lead in AI as they approach the "country of geniuses in a datacenter", we may be able to "spend" a portion of that lead to ensure interpretability[10] is on a more solid footing before proceeding to truly powerful AI, while still defeating our authoritarian adversaries[11]. Even a 1- or 2-year lead, which I believe effective and well-enforced export controls can give us, could mean the difference between an "AI MRI" that essentially works when we reach transformative capability levels, and one that does not. One year ago we couldn't trace the thoughts of a neural network and couldn't identify millions of concepts inside them; today we can. By contrast, if the US and China reach powerful AI simultaneously (which is what I expect to happen without export controls), the geopolitical incentives will make any slowdown at all essentially impossible.

All of these—accelerating interpretability, light-touch transparency legislation, and export controls on chips to China—have the virtue of being good ideas in their own right, with few meaningful downsides. We should do

all of them anyway. But they become even more important when we realize that they might make the difference between interpretability being solved before powerful AI or after it.

Powerful AI will shape humanity's destiny, and we deserve to understand our own creations *before* they radically transform our economy, our lives, and our future.

*Thanks to Tom McGrath, Martin Wattenberg, Chris Olah, Ben Buchanan, and many people within Anthropic for feedback on drafts of this article*.