ANTHROP\C

Unlocking AI understanding: Advancing interpretability at Anthropic

A fundamental problem for AI safety is that nobody understands how large language models work. Think of it like the human brain—we know it's capable of incredible feats, but neuroscientists are nowhere near to fully cracking its code.

Anthropic's Interpretability team pioneered the use of a method called "Dictionary Learning" that throws light on the inner workings of AI models. The method uncovers the way that the model represents different concepts—ideas like, say, "friendship", "screwdrivers", or "Paris"—within its neural network.

Knowing how AIs organize concepts helps to make them more interpretable: we can, to some limited degree, work out what they're "thinking", which has big implications for how we use them for work and elsewhere. And as we'll detail below, it might also help make them safer.

Overcoming the Challenge of Superposition

One key obstacle to understanding AI models is the phenomenon of "superposition." Unlike traditional computer programs where each component has a clear, singular purpose, the neurons inside AI models don't correspond to individual concepts. Instead, information is distributed across the network in complex, overlapping patterns. In this respect it's similar to the English alphabet: outside of exceptions like "I", a single character doesn't mean anything on its own; only in combination with other characters does it take on meaning. And with AI models, we don't know how that alphabet fits together: even when we look inside the "black box" of an AI model, we don't immediately understand what we're seeing.

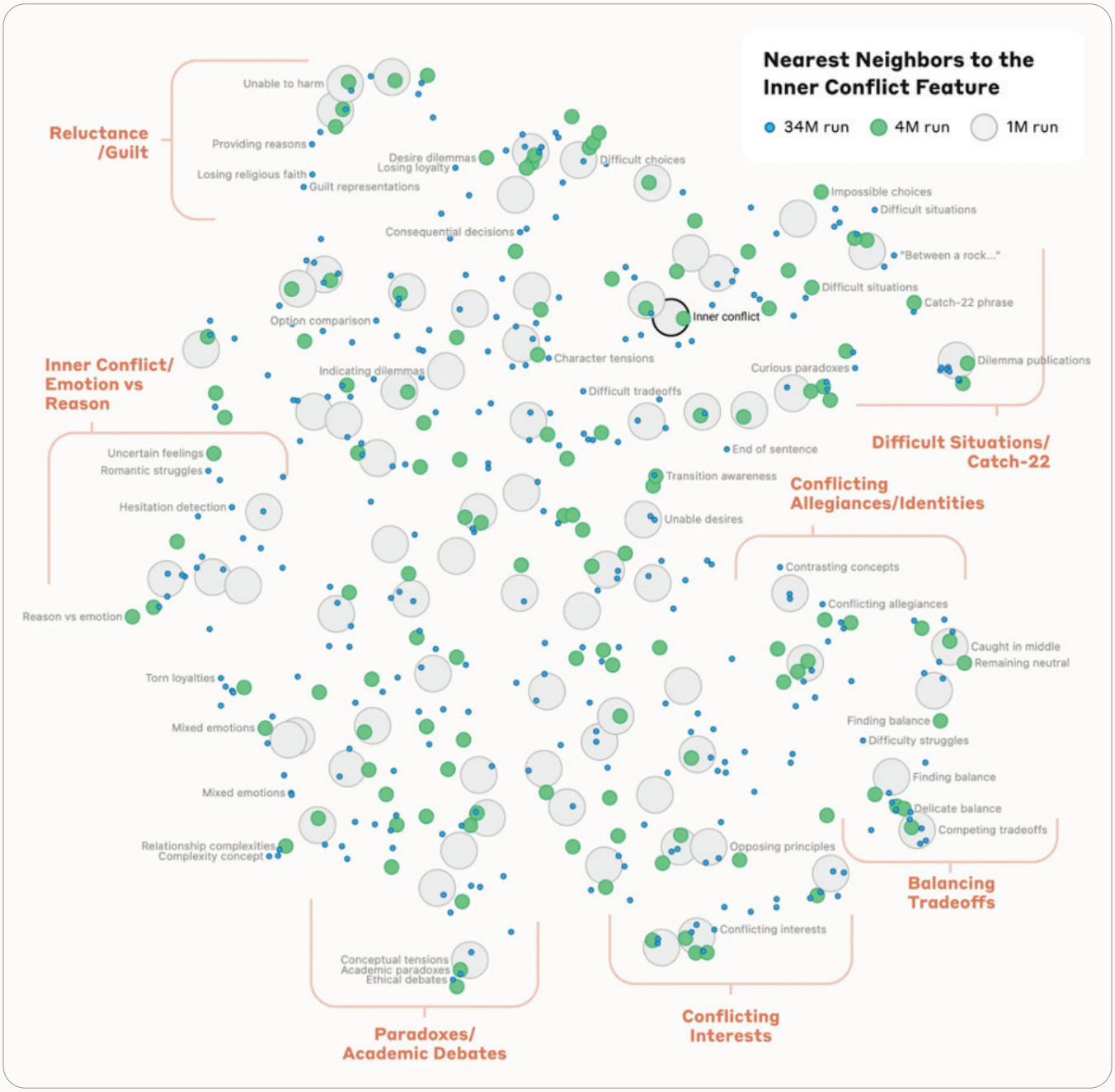
To truly understand AI models, we need specialized methods to break down this superposition, much like neuroscientists use various techniques (MRI scans, EEG, and so on) to understand the human brain. This is where our Dictionary Learning technique comes in: it allows us to decipher the features that are represented inside a model. In future, we might be able to manipulate these features—amplifying them or dampening them down—to change, in a very precise way, the way the model behaves.

Mapping the mind of a large language model

Our research uncovered many millions of features that are represented inside Claude 3 Sonnet, from concrete objects to abstract concepts. For instance, in the figure below you can see a map of features that relate to the abstract idea of "inner conflict": you can see how features that are more closely related in their meaning can be grouped together. You can also see how specific these features are: for example, the model understands the concepts of "hesitation detection" and "competing tradeoffs".

But just like in the human brain, the concepts can also be much more concrete. We found, for example, that there was a specific feature for the Golden Gate Bridge—it activated when users asked Claude about famous bridges in San Francisco, or red suspension bridges, or many similar prompts.

As a demonstration we made available a model where we'd amplified the Golden Gate Bridge feature. That meant that Claude became completely fixated on the bridge, bringing it up in response to almost any query. This "Golden Gate Claude" (which only existed for 24 hours) showed how changes to very specific features could have a big impact on a model's behavior: potentially pointing the way to a future where users have much stronger ability to steer AI models.



The Future

We're still in the early stages of this research. There's a lot to do to make these Interpretability techniques practically viable. But there's significant potential for enterprises who use AI models:

Enhanced Control: As we better understand how models represent information, we might be able to develop methods to steer and control their outputs more precisely (improving current prompt engineering methods), giving businesses the control they need to securely deploy AI systems while ensuring they are ethically aligned.

Improved Safety: Just as neuroscientists use their findings to address neurological disorders, our Interpretability efforts could help us identify and mitigate potential risks or biases in AI models. We can picture using Interpretability-derived techniques to warn about models that are being deceptive or considering potentially dangerous outputs. This could also lead to more reliable AI systems for business applications, improving safety in areas like hallucination and delivering more reliable, predictable outputs.

Regulatory Readiness: As AI regulation evolves, the ability to explain and justify AI decisions—especially in heavily-regulated industries such as Financial Services—will be crucial. Interpretability techniques promise an audit trail for how a model came to produce a specific output—a major contrast from the "black box" decisions they produce at present.

The field of AI Interpretability is complex and rapidly evolving. As AI becomes ever more deeply integrated into business and life, we're committed to ongoing research and development to build upon these initial findings and translate them into practical benefits. We invite you to explore how our advances in Interpretability can help drive your organization's AI strategy.