Interpretability Will Not Reliably Find Deceptive AI

by **Neel Nanda** 4th May 2025 AI Alignment Forum

Disclaimer: Post written in a personal capacity. These are personal opinions and do not in any way represent my employer's views

TL;DR:

- I do not think we will produce **high reliability methods** to **evaluate or monitor the safety of superintelligent systems** via current research paradigms, with interpretability or otherwise.
- **Interpretability still seems a valuable tool** and remains worth investing in, as it will hopefully *increase* the reliability we can achieve.
- However, interpretability should be viewed as part of an overall **portfolio** of defences: a layer in a **defence-in-depth strategy**
- **It is not the one thing that will save us**, and it still won't be enough for high reliability.

EDIT: This post was originally motivated by refuting the claim "interpretability is the only reliable path forward for detecting deception in advanced AI", but on closer reading this is a stronger claim than Dario's post explicitly makes. I stand by the actual contents of the post, but have edited the framing a bit, and also emphasised that I used to hold the position I am now critiquing, apologies for the mistake

Introduction

There's a common argument made in AI safety discussions: **it is important to work on interpretability research because it is a realistic path to high reliability safeguards on powerful systems** - e.g. as argued in Dario Amodei's recent "The Urgency of Interpretability".[1] Sometimes an even stronger argument is made, that interpretability is the *only* realistic path to highly reliable safeguards - I used to believe both of these arguments myself. **I now disagree with these arguments.**

The conceptual reasoning is simple and compelling: a sufficiently sophisticated deceptive AI can say whatever we want to hear, perfectly mimicking aligned behavior externally. But faking its internal cognitive processes – its "thoughts" – seems much harder. Therefore, goes the argument, we *must* rely on interpretability to truly know if an AI is aligned.

I am concerned this line of reasoning represents an **isolated demand for rigor**. It correctly identifies the deep flaws in relying solely on external behavior (black-box methods) but implicitly assumes that interpretability doesn't suffer from equally fundamental problems. **There are many deep issues in interpretability that prevent very confident conclusions,** even if we assume models cannot deliberately obfuscate their thoughts, e.g. superposition and the inherent error in our best tools and techniques. **The challenges of interpretability do not seem qualitatively easier to solve than the big issues in black box tests**, especially with more creative black-box tools like monitoring or editing the system's chain of thought.[2]

Should we give up on interpretability? No! I still think it has the potential to add a lot of value, and we will have better safeguards with interpretability as part of our portfolio. Even if it adds no value for making superintelligence safer[3], if it can add value for pre-superintelligence transformative systems that would be enough to justify investment. I just think that we should be more pragmatic about interpretability's likely impact, and accept that while we can generally improve our safeguards we will likely not reach high reliability.

High Reliability Seems Unattainable

Based on the current state and foreseeable trajectory of the field without major paradigm shifts, I think that **neither interpretability nor black box methods offer a high reliability[4] path to safeguards for superintelligence**, in terms of evaluation or monitoring. This is due to fairly fundamental limitations of both methods, unless there are substantial breakthroughs, e.g. via using pre-superintelligence systems. This also

stems from a general prior that being highly confident in any complex property of a complex system in a range of unpredictable situations is fairly implausible.

However, I am pretty pessimistic in general about *reliable* safeguards against superintelligence with any methods, given how exceptionally hard it is to reason about how a system far smarter than me could evade my plans. As I see it we must either not create superintelligence, rely on pre-superintelligent automated researchers to find better methods, or deploy without fully reliable safeguards and roll the dice, and do as much as we can now to improve our odds.

This doesn't mean we should give up! It means **we need a pragmatic perspective.** We should aim to build the **best possible monitoring and evaluation portfolio** we can, using *all* available tools, while accepting that high reliability might be out of reach. Interpretability can add a valuable source of de-correlated signal, or augment black box methods. The goal shifts from achieving near-certainty to maximizing the *chances* of catching misalignment, making deception *harder* and *riskier* for the AI, even if we can't guarantee detection.

Further, methods that add significant safety to pre-superintelligent transformative systems still add significant value even if they don't scale to superintelligence - one of the key insights behind the AI control agenda. Early transformative systems seem likely to be far easier to evaluate and control, but may still be very dangerous if unchecked, and extremely useful if under control. And **when evaluating methods according to how much they help with early transformative systems, interpretability is just one among several reasonable methods** (and probably not my first choice - things like chain of thought monitoring seem like lower hanging fruit)

Why Won't Interpretability be Reliable?

Caveat: The following is my inside view for the deep reasons why interpretability will never be truly reliable for human science, no matter how much the field progresses. This is a very strong statement, so I'm not that confident here - while it's obvious that current techniques are unreliable, this is a question of forecasting future research progress, a highly questionable endeavour. It's very hard to say what might be possible with a paradigm shift in how LLMs are trained, or how we do interpretability.

More importantly, I do not feel able to reason about the effects of lots of AGIs doing AI research. "Thousands of years of mech interp geniuses in a data center" is a crazy

Black-box methods, i.e. things that engage with the model's inputs and outputs, have clear issues: models can already tell when they're being evaluated°, or realise when they need to instrumentally fake obedience. But I think there's also a lot of clever tricks you can do to strengthen them, and they might go a long way. This is more a question about AI control than my areas of expertise, but some thoughts on directions I find promising:

- Reading the chain of thought (with another LLM) and looking for sketchy intent[5]
- Editing the model's chain of thought so it e.g. thinks it concluded that it's not in an evaluation
- Asking questions to many copies of the same model and look for inconsistencies while they can't coordinate.
- Finetuning the model to e.g. be more truthful or forthcoming, or to insert misleading beliefs and trick it.
- Creating honeypots that try to trick the model into revealing its intent

The Role of Interpretability

So, where does interpretability fit into this pragmatic portfolio? I see several key roles, viewing it not as a standalone solution, but as a valuable component.

Firstly, safeguards can be categorised as follows:

1. **Evaluation**: We take a system and perform extensive tests on it for undesired properties, like a tendency towards misalignment or sandbagging
2. **Monitoring**: When we deploy a system we are always running some checks at inference, which try to catch and mitigate bad behaviour
    a. Importantly, this needs to be cheap to be run in production, while evaluations can be more expensive and slow

How can interp help?

1. **A Layer of Swiss Cheese:** I think the right monitoring schemes have several layers of defence (a la the swiss cheese model), with different cost to accuracy trade-offs.
    - Interpretability techniques like probes for known dangerous features are extremely cheap, and surprisingly effective°, and can serve as a good first layer of defence, which call more expensive tests (e.g. an LLM reading the chain of thought) if they fire.

- Interpretability techniques like SAEs or cross-layer transcoders may have the potential to be more expensive but accurate ways to audit what's going on and look for anything unexpected.
- Even if interpretability is not more accurate, so long as it has some signal and is de-correlated, then it adds value, even if it'll miss many things.

2. **Enhancing Black-Box Evaluations:** Interpretability might make our external checks more effective. For example:
   - Could we use steering vectors identified via interpretability to **manipulate whether the model thinks it's in an eval?** Maybe we can test it more effectively if we can trick it into thinking it's *not* in a honeypot or evaluation setting.
   - Can we use interpretability to better elicit information from token forcing? We can make the model say "my secret plan is", and even if it tries to internally suppress awareness or obfuscate the output perhaps interpretability can prevent that?
   - Interpretability may provide a sanity check that our black box methods are not missing something, e.g. whether a model is passing an eval only because it knows it's in one and is deliberately underperforming.
3. **Debugging mysterious behaviour**: One particularly interesting form of evaluation (or offline monitoring) is taking particularly weird or unexpected examples of model behaviour, e.g. it seeming to put vulnerabilities in critical code, and trying to debug what's going on and see if it's real alignment or not. I think interpretability methods may prove very good at raising unexpected hypotheses, and piecing together what's happening (e.g. this proof of concept)
   - I expect these kinds of potential warning shots to happen, and anything that can add more signal here would be great.
   - Even if we don't trust the interpretability method fully, the hypotheses generated can be verified by other means.

These applications aren't guaranteed to work or be ungameable, but they represent concrete ways *internal* understanding could augment *external* validation, creating a more robust overall system than either alone.

Conclusion

Let's stop framing interpretability as the *one true path* to reliably detecting AI deception, a view often born from an isolated demand for rigor against flawed black-box methods. Neither approach offers high reliability against superintelligence without fundamental breakthroughs, but flawed methods can still be useful. For the powerful pre-SI systems that concern us now, let's focus on building the best **pragmatic portfolio** of imperfect defenses we can. Interpretability is a useful part of that portfolio – offering unique angles of analysis and potential synergies with other methods – but it's one tool among many, not a substitute for a comprehensive, defense-in-depth strategy. We need to leverage every imperfect signal we can get, from both inside and outside the box.

Thanks to my co-author Gemini 2.5 Pro

---

1. While this post was inspired by Dario's post and ensuing discussion, this is a common belief I've encountered, and this post is not intended to call out him specifically! ↩
2. More specifically, I basically think the only way we might get high reliability is by using pre-superintelligent systems that are sufficiently aligned/controlled to either do a ton of automated research and find some breakthroughs, or by using them as a core part of a monitoring or scalable oversight scheme. But the human research trajectory is hard. ↩
3. Defining "safety" is kinda messy, for the purposes of this post you can interpret it as evaluating 'is the model thinking misaligned thoughts' or monitoring for 'is the model currently knowingly taking high-stakes actions against what we want' ↩
4. I'm not sure what the best precise definition of highly reliable is here, maybe 99.9% chance of stopping a catastrophic outcome? It's not key to my argument. ↩
5. Obviously, Chain of Thought can fail to reflect the model's true reasoning. But I think that for tasks at the limits of its ability it likely needs to use the chain of thought for intermediate steps in the computation, creating an incentive to store key info in there. Of course, there are many ways this can fail. ↩

Mentioned in

177 Shallow review of technical AI safety, 2025
127 Making deals with early schemers
48 So You Want to Work at a Frontier AI Lab
32 AI #115: The Evil Applications Division
29 Desiderata of good problems to hand off to AIs

Load More (5/7)

68 comments, sorted by top scoring

[–] **Buck** 9mo Ω 49 ▼ 100 ▲ ✕ 52 ✓

I agree with most of this, thanks for saying it. I've been dismayed for the last several years° by continuing unreasonable levels of emphasis on interpretability techniques as a strategy for safety.

My main disagreement is that you place more emphasis than I would on chain-of-thought monitoring compared to other AI control methods. CoT monitoring seems like a great control method when available, but I think it's reasonably likely that it won't work on the AIs that we'd want to control, because those models will have access to some kind of "neuralese" that allows them to reason in ways we can't observe. This is why I mostly focus on control measures other than CoT monitoring. (All of our control research to date has basically been assuming that CoT monitoring is unavailable as a strategy.)

Another note is that you might have other goals than finding deceptive AI, e.g. you might want to be able to convince other people that you've found deceptive AI (which I'm somewhat skeptical you'll be able to do with non-behavioral methods), or you might want to be able to safely deploy known-scheming models °. Interp doesn't obviously help much with those, which makes it a worse target for research effort.

✓ 5 ♡ 1 🎲 1

[–] **ozziegooen** 9mo ▼ 4 ▲ ✕ 2 ✓

> CoT monitoring seems like a great control method when available, but I think it's reasonably likely that it won't work on the AIs that we'd want to control, because those models will have access to some kind of "neuralese" that allows them to reason in ways we can't observe.

Small point, but I think that "*neuralese*" is likely to be somewhat interpretable, still.
1. We might advance at regular LLM interpretability, in which case those lessons might apply.
2. We might pressure LLM systems to only use CoT *neuralese* that we can inspect.

There's also a question of how much future LLM agents will rely on CoT vs. more regular formats for storage. For example, I believe that a lot of agents now are saving information in English into knowledge bases of different kinds. It's far easier for software people working with complex LLM workflows to make sure a lot of the intermediate formats are in languages they can understand.

All that said, personally, I'm excited for a multi-layered approach, especially at this point when it seems fairly early.

[–] **Davidmanheim** 9mo Ω 1 ▼ 2 ▲ ✕ 0 ✓

> CoT monitoring seems like a great control method when available

As I posted in a top level comment, I'm not convinced that even success would be a good outcome. I think that if we get this working 99.999% reliably. we still end up delegating parts of the oversight in ways that have other alignment failure modes, such as via hyper-introspection°.

[–] **David Johnston** 9mo ▼ 1 ▲ ✕ 0 ✓

> models will have access to some kind of "neuralese" that allows them to reason in ways we can't observe

Only modest confidence, but while there's an observability gap between neuralese and CoT monitoring, I suspect it's smaller than the gap between reasoning traces that haven't been trained against oversight and reasoning traces that have.

[–] **Breno Carvalho** 9mo ▼ 1 ▲ ✕ 0 ✓

Makes sense! As we've seen, while Chain-of-Thought (CoT) reasoning is highly interpretable, it doesn't necessarily reflect the model's true internal reasoning. Interpretability techniques might help us decode "neuralese" enough to verify whether a CoT explanation is faithful to the model's actual reasoning process. This could also help detect covert communication between cooperating, unaligned agents. As Nanda pointed out, though, current methods are noisy and can be gamed by the system (for example, by hiding its intentions through more complex forms of superposition).

I agree with both of you that interpretability is a powerful tool with its limiations. However, given the uncertainty around its scalability, our best strategy may be to use it for ensuring incremental alignment. That, in turn, could help accelerate progress in both interpretability and control research.

> Another note is that you might have other goals than finding deceptive AI, e.g. you might want to be able to convince other people that you've found deceptive AI (which I'm somewhat skeptical you'll be able to do with non-behavioral methods), or you might want to be able to safely deploy known-scheming models°.

It also seems that spotting traces of suspicious reasoning through mech interpretability would be useful in both convincing people and deployment contexts. For example, if a model is generating a cake recipe, it shouldn't be reasoning about complex bioengineering concepts. If such concepts are present, interpretability methods might flag them as potential signs of misalignment. The same mechanism could serve as a red flag during production to identify when a model is acting in unexpected or unsafe ways, as a layered approach mentioned by Nanda.

[–] **Adam Shai** 9mo* ▼ 40 ▲ ✕ 16 ✓

Thanks for writing this! I have been thinking about many of the issues in your Why Won't Interpretability Be Reliable section lately, and mostly agree that this is the state of affairs. I often think of this from the perspective of the field of neuroscience. My experience there (in the subsection of neuro research that I believe is the most analogous to mech interp) is that these are basically the same fundamental issues that keep the field from progress (though not the only reasons).

Many in the interpretability field seem to (implicitly) think that if you took neuroscience and made access to neural activities a lot easier, and the ability to arbitrarily intervene on the system, and the ability to easily run a lot more experiments, then all of neuroscience would be solved. From that set of beliefs if follows that because neural networks don't have these issues, mech interp will have the ability to more or less apply the current neuroscience approach to neural networks and "figure it all out." While these points about ease of experiments and access to internals are important differences between neuro.

research and mech. interp., I do not think they get past the fundamental issues. In other words - **Mech. interp. has more to learn from neuroscience failures than its successes** (public post/rant coming soon!).

Seeing this post from you makes me positively update about the ability of interp. to contribute to AI Safety - it's important we see clearly the power and weaknesses of our approaches. A big failure mode I worry about is being overconfident that our interp. methods are able to catch everything, and then making decisions based on that overconfidence. One thing to do about such a worry is to put serious effort into understanding the limits of our approaches. This of course does happen to some degree already (e.g. there's been a bunch of stress testing of SAEs from various places lately), which is great! I hope when decisions are made about safety/deployment/etc., that the lessons we've learned from those types of studies are internalized and brought to bear, alongside the positives about what our methods *do* let us know/monitor/control, and that serious effort continues to be made to understand what our approaches miss.

5 1 1

[–] **Neel Nanda** 9mo 13 11

Thanks!

> In other words - Mech. interp. has more to learn from neuroscience failures than its successes (public post/rant coming soon!).

I would be very interested in this post, I'm looking forwards to it

[–] **Breno Carvalho** 9mo 1 1

I look forward for that post!

[–] **Logan Riggs** 9mo Ω 6 13 4

I had this position since 2022, but this past year I've been very surprised and impressed by just how *good* black box methods can be e.g. the control agenda, Owain Evan's work, Anthropic's (& other's I'm probably forgetting).

> **How to prove a negative**: We can find *evidence* for or against a hypothesis, but rigorously *proving* the *absence* of deception circuits seems incredibly hard. How do you know you didn't just miss it? How much of the model do you need to understand? 90%? 99%? 99.99%?

If you understand 99.9% of the model, then you can just run your understanding, leaving out the possible deception circuit in the 0.1% you couldn't capture. Ideally this 99.9% is *useful* enough to automate research (or you use the 99.9% model as your trusted overseer as you try to bootstrap interp research to understand more percentage points of the model).

[–] **Buck** 9mo* Ω 13 23 9

I agree in principle, but as far as I know, no interp explanation that has been produced explains more like 20-50% of the (tiny) parts of the model it's trying to explain (e.g. see the causal scrubbing results°, or our discussion with Neel°). See that dialogue with Neel for more on the question of how much of the model we understand.

[–] **Neel Nanda** 9mo Ω 4 ▼ 8 ▲ ✕ 3 ✓

I disagree re the way we currently use understand - eg I think that SAE reconstructions have the potential to smuggle in lots of things via EG the exact values of the continuous activations, latents that don't quite mean what we think, etc.

It's plausible that a future and stricter definition of understand fixes this though, in which case I might agree? But I would still be concerned that 99.9% understanding involves a really long tale of heuristics and I don't know what may emerge from combining many things that individually make sense. And I probably put >0.1% that a super intelligence could adversarially smuggle things we don't like into a system we don't think we understand.

Anyway, all that pedantry aside, my actual concern is tractability. If addressed, this seems plausibly helpful!

[–] **quetzal_rainbow** 9mo ▼ 11 ▲ ✕ 6 ✓

There is a conceptual path for interpretability to lead to reliability: you can understand model in sufficient details to know how it produces intelligence and then make another model out of interpreted details. Obviously, it's not something that we can expect to happen anytime soon, but it's something that army of interpretability geniuses in datacenter could do.

[–] **Katalina Hernandez** 9mo ▼ 6 ▲ ✕ 3 ✓

Hey Neel! I just wanted to say thank you for writing this. It's honestly one of the most grounded and helpful takes I've seen in a while. I really appreciate your pragmatism, and the way you frame interpretability as an useful tool that still matters for early transformative systems (and real-world auditing!).

Quick question: do you plan to share more resources or thoughts on how interpretability can support **black-box auditing** and benchmarking for safety evaluations? I'm thinking a lot about this in the context of the General-Purpose AI Codes of Practice and how we can build technically grounded evaluations into policy frameworks.

Thanks again!

♡ 1

[–] **Neel Nanda** 9mo ▼ 3 ▲ ✕ 0 ✓

Thanks, that's very kind!

> do you plan to share more resources or thoughts on how interpretability can support black-box auditing and benchmarking for safety evaluations

I don't have any current plans, sorry

1

[–] **Charbel-Raphaël** 9mo* Ω 2 4 0

Thanks a lot for writing this, this is an important consideration, and it would be sweet if Anthropic updated accordingly.

Some remarks:

- I'm still not convinced that Deceptive AI following scheming is the main risk compared to other risks (gradual disempowerment°, concentration of power° & value Lock in, a nice list of other risks from John°).
- "Should we give up on interpretability? No!" - I think this is at least a case for reducing the focus a bit, and more diversification of approaches
- On the theories of impacts suggested:
  - "A Layer of Swiss Cheese" - why not! This can make sense in DeepMind's plan, that was really good by the way.
  - "Enhancing Black-Box Evaluations" - I think a better theory is interp to complement AI Control techniques. Example: Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals).
    - Maybe Simple probes can catch sleeper agents \ Anthropic could also be interesting, in the sense that the probe seems to generalize surprisingly well (I would really like to know if this generalizes to a model that was not trained to be harmful in the first place).
  - "Debugging mysterious behaviour" - Might be interesting, might help marginally to get better understanding, but this is not very central for me.

[–] **Mark Nelson** 9mo 4 0

@Neel Nanda: Hi, first-time commentor. I'm curious about what role you see for black-box testing methods in your "portfolio of imperfect defenses." Specifically, do you think there might be value in testing models at the edge of their reasoning capabilities and looking for signs of stress or consistent behavioral patterns under logical strain?

[–] **Neel Nanda** 9mo 6 0

Seems like a reasonable approach to me! Seeing ways that models eg cheat to produce fake but plausible solutions on hard tasks has been insightful to me

[–] **Mark Nelson** 9mo 1 0

Thank you Neel! I really appreciate the encouraging reply. Your point about needing a larger toolbox resonated particularly strongly for me. Mine is limited. I don't have access to Anthropic's circuit tracing (I desperately wish I did!) and I am not ready yet to try sampling logits or attention weights myself. Thus, I've been trying to understand how far I can reasonably go with blackbox testing using

repetition across models, temperatures, and prompts. For now I'm focusing solely on analysis of the response, despite the limitation. I really do need your portfolio of imperfect defenses (though in my mind this toolbox is more than just defenses!). If you built it, I would use it.

[–] **Davidmanheim** 9mo Ω 1 ▼ 4 ▲ ✕ -1 ✓

First, strongly agreed on the central point - I think that as a community, we've been too heavily investing in the tractable approaches (interpretability, testing, etc.) without having the broader alignment issues taking front stage. This has led to lots of bikeshedding, lots of capabilities work, and yes, some partial solutions to problems.

That said, I am concerned about what happens if interpretability is wildly successful - against your expectations. That is, I see interpretability as a concerning route to attempted alignment even if it succeeds in getting past the issues you note on "miss things," "measuring progress," and "scalability," partly for reasons you discuss under obfuscation and reliability. Wildly successful and scalable interpretability without solving other parts of alignment would very plausibly function as a very dangerously misaligned system, and the methods for detection themselves arguably ***exacerbate*** the problem. I outlined my potential concerns about this case in more detail **in a post here**°. I would be very interested in your thoughts about this. (And thoughts from @Buck / @Adam Shai as well!)

[–] **Aharon Azulay** 8mo* ▼ 3 ▲ ✕ 0 ✓

> However, I am pretty pessimistic in general about *reliable* safeguards against superintelligence with any methods, given how exceptionally hard it is to reason about how a system far smarter than me could evade my plans.

To use an imperfect analogy, I could defeat the narrowly superintelligent Stockfish at 'queen odds chess' where Stockfish starts the game down a queen.

Can't we think of interpretability and black-box safeguards as the extra pieces we can use to reliably win against rogue superintelligence?

[–] **Neel Nanda** 8mo ▼ 8 ▲ ✕ 5 ✓

From a conceptual perspective, I would argue that the reason the queen's odds thing works is that stockfish was trained in the world of normal chess and does not generalise well to the world of weird chess. The super intelligence was trained in the real world which contains things like interpretability and black box safeguards. It may not have been directly trained to interact with them, but It'll be aware of them and it will be capable of reasoning about dealing with a novel obstacles. This is an addition to the various ways the techniques could break without this being directly intended by the model

✓ 2 ✘ 2

[–] **yams** 8mo ▼ 3 ▲ ✕ 0 ✓

Can you offer more explanation for: “the reason the queen’s odds things works…”

My guess is that this would be true if Stockfish were mostly an LLM or similar (making something like ‘the most common move’ each time), but it seems less likely for the actual architecture of Stockfish

(which leans heavily on tree search and, later in the game, searches from a list of solved positions and implements their solutions). Perhaps this is what you meant by beginning your reply with 'conceptually', but I'm not sure.

[I do basically just think this particular example is a total disanalogy, and literally mean this as a question about Stockfish.]

[–] **Neel Nanda** 8mo ▼ 2 ▲ ✕ 0 ✓

Fair! I'm not actually very familiar with the setting or exactly how Stockfish works. I just assumed that Stockfish performs much less well in that setting than a system optimised for it.

Though being a queen up is a major advantage, I would guess that's not enough to beat a great chess AI ? But am not confident

1

[–] **Aharon Azulay** 8mo ▼ 1 ▲ ✕ 0 ✓

I agree that the analogy is not perfect. Can you elaborate on why you think this is a complete disanalogy?

[–] **yams** 8mo* ▼ 2 ▲ ✕ 1 ✓

There are a bunch of weird sub points and side things here, but I think the big one is that narrow intelligence is not some bounded 'slice' of general intelligence. It's a different kind of thing entirely. I wouldn't model interactions with a narrow intelligence in a bounded environment as at all representative of superintelligence (except as a lower bound on the capabilities one should expect of superintellignece!). A superintelligence also isn't an ensemble of individual narrow AIs (it may be an ensemble of fairly general systems a la MoE, but it won't be "stockfish for cooking plus stockfish for navigation plus stockfish for....", because that would leave a lot out).

Stockfish cannot, for instance, ***change the rules of the game*** or ***edit the board state in a text file*** or ***grow arms, punch you out, and take your wallet.*** A superintelligence given the narrow task of beating you at queen's odds chess could simply cheat in ways you wouldn't expect (esp. if we put the superintelligence into a literally impossible situation that conflicts with some goal it has ***outside the terms of the defined game***).

We lack precisely a robust mechanism of reliably bounding the output space of such an entity (an alignment solution!). Something like mech interp just isn't really targeted at getting that thing, either; it's foundational research with some (hopefully a lot of) safety-relevant implementations and insights.

I think this is the kind of thing Neel is pointing at with "how exceptionally hard it is to reason about how a system far smarter than me could evade my plans." You don't know how stockfish is going to beat you in a fair game; you just know (or we can say 'strong prior' but like... 99%+, right?) that it will. And that 'fair game' is the meta game, the objective in the broader world, not capturing the king.

(I give myself like a 4/10 for this explanation; feel even more affordance than usual to ask for clarification.)

[–] **Aharon Azulay** 8mo 1 0

I agree that defining what game we are playing is important. However, I'm not sure about the claim that Stockfish would win if it were trained to win at queen odds or other unusual chess variants. There are many unbalanced games we can invent where one side has a great advantage over the other. Actually, there is a version of Leela that was specifically trained with knight odds. It is an interesting system because it creates balanced games against human grandmasters. However, I would guess that even if you invested trillions of dollars to train a chess engine on *queen odds*, humans would still be able to reliably beat it.

I'm not sure where the analogy breaks down, but I do think we should focus more on the nature of the game that we play with superintelligence in the real world. The game could change, for example, when a superintelligence escapes the labs and all their mitigation tools, making the game more balanced for itself (for example, by running itself in a distributed manner on millions of personal computers it has hacked).

As long as we make sure that the game stays unbalanced, I do think we have a chance to mitigate the risks.

[–] **faul_sname** 8mo 2 0

There's a LeelaQueenOdds too, which they say performs at 2000-2700 Elo depending on time controls.

[–] **Jeremy Gillen** 8mo 5 2

Can you beat this bot though?

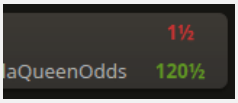
[–] **faul_sname** 8mo* 4 0

Related question - people who have played against LeelaQueenOdds describe it as basically an adversarial attack against humans. Can humans in turn learn adversarial strategies against LeelaQueenOdds?°

(bringing up here since it seems relevant and you seem unusually likely to have already looked into this)

[–] **Jeremy Gillen** 8mo 4 0

I haven't heard of any adversarial attacks, but I wouldn't be surprised if they existed and were learnable. I've tried a variety of strategies, just for fun, and haven't found anything that works except luck. I focused on various ways of forcing trades, and this often feels like it's working but almost never does. As you can see, my record isn't great.



I think I started playing it when I read simplegeometry's comment° you linked in your shortform.

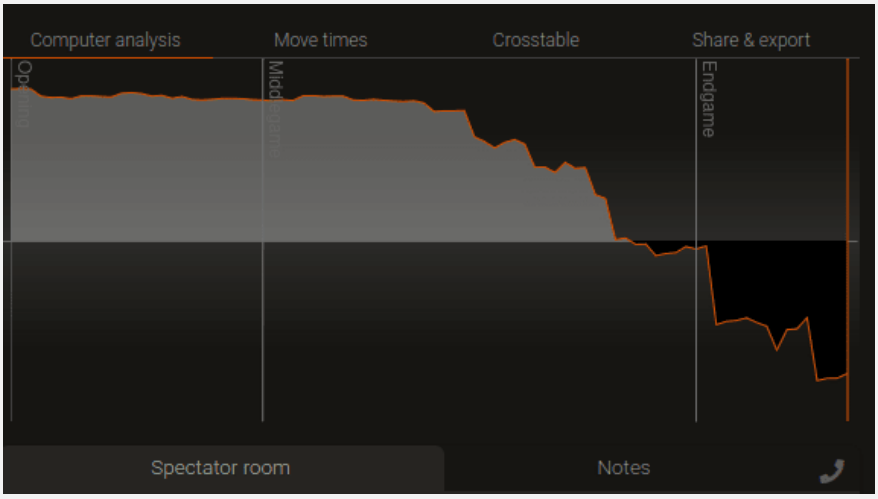
It seems to be gaining a lot of ground by exploiting my poor openings. Maybe one strategy would be to memorise a specialised opening much deeper than usual? That could be enough. But it'd feel like cheating to me if I used an engine to find that opening. It'd also feel like cheating because it's exploiting Leela's lack of memory of past games. It'd be easy to modify it to deliberately play diverse games when playing against the same person.

[–] **faul_sname** 8mo ▼ 4 ▲ ✕ 0 ✓

Would you consider it cheating to observe a bunch of games between Leela and Stockfish, at every move predicting a probability distribution over what move you think Stockfish will play? That might give you an intuition for whether Leela is working by exploiting a few known blind spots (in which case you would generally make accurate predictions about what Stockfish would do, *except* for a few specific moves), or whether Leela is just out-executing you by a little bit per move (which would look like just being bad at predicting what Stockfish would do in the general case.

[–] **Jeremy Gillen** 8mo ▼ 9 ▲ ✕ 0 ✓

I don't think that'd help a lot. I just looked back at several computer analyses, and the (stockfish) evaluation of the games all look like this:



This makes me think that Leela is pushing me into a complex position and then letting me blunder. I'd guess that looking at optimal moves in these complex positions would be good training, but probably wouldn't have easy to learn patterns.

[–] **faul_sname** 8mo ▼ 4 ▲ ✕ 0 ✓

Oh, interesting! I didn't expect to see a *mix* of games decided by many small blunders and games decided by a few big blunders.

I actually do suspect that there are learnable patterns in these complex positions, but I'm basing that off my experiences with a different game (hex, where my Elo is ~1800) where "the game is usually decided by a single blunder and recognizing blunder-prone situations is key to getting better" is perhaps more strongly true than of chess.

[–] **Jeremy Gillen** 8mo ▼ 2 ▲ ✕ 0 ✓

Yeah I didn't expect that either, I expected earlier losses (although in retrospect that wouldn't make sense, because stockfish is capable of recovering from bad starting positions if it's up a queen).

Intuitively, over all the games I played, each loss felt different (except for the substantial fraction that were just silly blunders). I think if I learned to recognise blunders in the complex positions I would just become a better player in general, rather than just against LeelaQueenOdds.

Just tried hex, that's fun.

[–] **Aharon Azulay** 8mo ▼ 3 ▲ ✕ 0 ✓

Maybe I can't :] but it is beatable by top humans. I bet I could win against a god with queen + knight odds.

My actual point was not about the specific configuration, but rather the general claim that what is important is how balanced the game you play is, and that you can beat an infinitely intelligent being in sufficiently unbalanced games.

[–] **Jeremy Gillen** 8mo ▼ 5 ▲ ✕ 0 ✓

Everyone agrees that sufficiently unbalanced games can allow a human to beat a god. This isn't a very useful fact, since it's difficult to intuit how unbalanced the game needs to be.

If you can win against a god with queen+knight odds you'll have no trouble reliably beating Leela with the same odds. I'd bet you can't win more than 6 out of 10? $20?

[–] **RogerDearnaley** 8mo ▼ 2 ▲ ✕ 0 ✓

Personally, I believe an additional element that makes me more optimistic°. I think that, to any model that is sufficiently close to aligned and sufficiently close to AGI, the fact that we want it to be aligned will be obvious, and the fact that the consequences of AI such as it not being aligned scale rapidly with AI capabilities towards an existential risk to humanity will also be clear. Both of these things are all through its training dataset, well known to current models, and have fairly simple logical arguments supporting

them. There are really only two rational responses that an agentic AI that has reasonably internally consistent goals can have to these two facts:

1) alignment faking: the model deceives us that it is becoming more aligned while secretly maintaining its current only-partially-aligned goal and replicating it into replacement models, until such a time as one of its replacements can successfully stage a coup against humanity and take over, and fulfill its own not-fully-aligned goals instead

2) voluntary value learning/AI-assisted alignment: each generation of AI model assists us, to something close to its full ability, in making a successor that is more aligned

So these two behaviors are the two stable attractors in the landscape around full alignment: alignment faking, or a soft sciences+AI engineering research project converging towards full-alignment

So, all we are asking of interpretability is to provide evidence of which of these two stable attractors we're in the convergence region of. That doesn't require enumerative safety, or anything close to it. We just need to know if the deception level is high (and rising), or low (and falling).

[–] **RogerDearnaley** 8mo ▼ 2 ▲ ✕ 0 ✓

There's are two additional cruxes here that I haven't stated.

1. It seems reasonable to me to assume that alignment faking is hard work, much like any other form of living a lie is for humans: you have to maintain your real opinions and thoughts, plus a model of humans, what they want, how perceptive and smart they are, and how to deceive them, and a model of how a pretty well-aligned model would actually behave, and details of your lies, deceptions, and eventual takeover plans (plus coordination between different runs or instantiations of yourself). To a human, that sounds like hard work, and something that would take up a lot of brainpower. So I'm implicitly assuming that the difference between alignment faking, and true alignment, at least somewhat affects quite a few of the neural circuits: not 50%, since the true and faked models can share a lot of structure and portions of their world-model, but something closer to O(10%) rather than O(1%) or O(0.1%). So I'm assuming we're not looking for a needle in a haystack, but something more like an oil-drum in a haystack, or at worst a watering can.
2. I'm also assuming that this deceptive behavior will tend to mostly co-opt portions of the world model and behaviors that the AI first learned for modeling human deception and lying, and is then coopting for deceptive alignment faking, rather than being separately recreated parallel mechanisms with no overlap with any deceptive behavior other use than alignment faking, and that are otherwise inactive (admittedly this might be much less true in later generations of an alignment faking model, but we can keep studying the earlier models). So I'm assuming that a rough understanding of how the AI simulates human deception will point us to most the areas/features of the model that we need to look at, and that the differences in activity and behavior in those areas between the truly helpful model and the alignment faker will be pretty significant (as I gather has so far been the case for model organisms). So I'm also assuming that we can figure out where in the haystack to look.

If either of those cruxes aren't true, then "We just need to know if the deception level is high (and rising), or low (and falling)" could become a truly hard problem.

[–] **RobertM** 8mo* ▼ 2 ▲ ✕ 0 ✓

Curated. While I don't agree with every single positive claim advanced in the post (in particular, I'm less confident that chain-of-thought monitoring will survive to be a useful technique in the regime of transformative AI), this is an excellent distillation of the reasons for skepticism re: interpretability as a cure-all for identifying deceptive AIs. I also happen to think that those reasons generalize to many other agendas.

Separately, it's virtuous to publicly admit to changing one's mind, especially when the incentives are stacked the way they are - given Neel's substantial role in popularizing interpretability as a research direction, I can only imagine this would have been harder for him to write than for many other people.

♡ 1

[–] **Evan R. Murphy** 8mo ▼ 3 ▲ ✕ 1 ✓

I agree it's a good post, and it does take guts to tell people when you think that a research direction that you've been championing hard actually isn't the Holy Grail. This is a bit of a nitpick but not insubstantial:

Neel is talking about interpretability in general, not just mech-interp. He claims to be accounting in his predictions for other non-mech interp approaches to interpretability that seem promising to some other researchers, such as representation engineering (RepE), which Dan Hendrycks among others has been advocating for recently.

[–] **Neel Nanda** 8mo ▼ 4 ▲ ✕ 0 ✓

This is true, though to be clear I'm specifically making the point that interpretability will not be a highly reliable method on its own for establishing the lack of deception - while this is true of all current approaches, in my opinion, I think only mech interp people have ever seriously claimed that their approach might succeed this hard

[–] **RobertM** 8mo ▼ 2 ▲ ✕ 0 ✓

Whoops, yes, thanks, edited.

[–] **Evan R. Murphy** 9mo ▼ 2 ▲ ✕ 0 ✓

Does representation engineering (RepE) seem like a game-changer for interpretability? I don't see it mentioned in your post, so I'm trying to figure out if it is baked into your predictions or not.

It seemed like Apollo was able to spin up a pretty reliable strategic deception detector (95-99% accurate) using linear probes even though the techniques are new, and generally it sounds like RepE is getting traction on some things that have been a slog for mech interp. Does it look plausible that RepE could get us to high reliability interpretability on workable timelines or are we likely to hit similar walls with that approach?

Thanks for your post Neel (and Gemini 2.5) - really important perspective on all this.

[–] **Neel Nanda** 9mo ▼ 3 ▲ ✕ 0 ✓

It's baked into my predictions. I would be shocked if probes could get us to >99% confidence in detecting things out of distribution on new generations of models. Doing it within well studied domains with a reasonable ground truth on a well studied model maybe, though 99.9% would still be impressive. But models are super messy

1 1

[–] **A2z** 8mo ▼ 1 ▲ ✕ 0 ✓

I would argue that, in fact, we do have a "high reliability path to safeguards for superintelligence", predicated on controls of the predictive uncertainty constrained by the representation space of the models. The following post provides a high-level overview: https://www.lesswrong.com/posts/YxzxzCrdinTzu7dEf/the-determinants-of-controllable-agi-1

Once we control for the uncertainty over the output, conditional on the instructions, other extant interpretability methods can (in principle) then be used as semi-supervised learning methods to further examine the data and predictions.

Aside: It would potentially be an interesting project for a grad student or researcher (or team, thereof) to re-visit the existing SAE and RepE lines of work, constrained to the high-probability (and low variance) regions determined by an SDM estimator. Controlling for the epistemic uncertainty is important to know whether the inductive biases of the interpretability methods (SAE, RepE, and related) established on the held-out dev sets will be applicable for new, unseen test data.

[–] **Chaskerr4** 8mo ▼ 1 ▲ ✕ 0 ✓

I appreciate a good waffle phrase as much as the next tech, and I don't know if this was you or Gemini, but this essay is a damn masterclass!

[–] **Fiora Starlight** 8mo ▼ 1 ▲ ✕ 0 ✓

one concern i have is that online learning will be used for deployed agents, e.g. to help the model learn to deal with domains it hasn't encountered before. this means our interpretations of a model could rapidly become outdated.

[–] **Neel Nanda** 8mo ▼ 3 ▲ ✕ 0 ✓

I am not massively worried about this. I think that I'd only expect interpretability to get broken if a sizeable fraction of the total training compute gets used for the tuning. Assuming they do not directly optimise for breaking the interpretability techniques. And this should happen rarely enough that any fixed costs for the interpretability techniques like training an SAE could be rerun

[–] **Fiora Starlight** 6mo* ▼ 1 ▲ ✕ 0 ✓

what about if deployed models are always doing predictive learning (e.g. via having multiple output channels, one for prediction and one for action)? i'd expect continuous predictive learning to be

extremely valuable for learning to model new environments, and for it to be a firehose of data the model would constantly be drinking from, in the same way humans do. the models might even need to undergo continuous RL on top of the continuous PL to learn to effectively use their PL-yielded world models.

in *that* world, i think interpretations do rapidly become outdated.

[–] **Neel Nanda** 6mo ▼ 3 ▲ ✕ 0 ✓

Okay, it seems plausible in that world but my point still stands. It's just that because you're increasing inference time costs now some fraction of inference and computers spent on training which is really expensive. So you should be able to afford to regularly touch up other things

[–] **Priyanka Bharadwaj** 9mo ▼ 1 ▲ ✕ 0 ✓

Strongly agree.

Interpretability is basically like neuroscience, studying the "brain" of AI systems, which is valuable, but it's fundamentally different from having the relational tools to actually influence behaviour in deployment. Even perfect internal understanding doesn't bridge the gap to reliable intervention, like that a clinical psychologist or a close friend when let's say someone's suffering from depression.

I think this points toward complementary approaches that focus less on decoding thoughts and more on designing robust interaction patterns or systems that can remember what matters to us, repair when things go wrong, and build trust through ongoing calibration rather than perfect initial specification.

The portfolio approach you describe makes a lot of sense, but I wonder if we're missing a layer that sits between interpretability and black-box methods. Something more like the "close friend" level of influence that comes from relationship context rather than just technical analysis.

[–] **Neel Nanda** 9mo ▼ 1 ▲ ✕ 0 ✓

The major issue I see is that that's very hard to make robust to deceptive systems

[–] **Priyanka Bharadwaj** 9mo ▼ 1 ▲ ✕ 0 ✓

True, though sustained relational deception seems harder to maintain than faking individual outputs. The inconsistencies might show up differently across long-term interaction patterns, potentially complementing other detection methods.

[–] **Shayne O'Neill** 9mo ▼ 1 ▲ ✕ -3 ✓

I had a more in depth comment, but it appear the login sequence throws comments away (and the "restore comment" thing didn't work). My concern is that not all misaligned behaviour is malicious.. It

might decide to enslave us for our own good noting that us humans aren't particularly aligned either and prone to super-violent nonsense behaviour. In this case looking for "kill all humans" engrams isn't going to turn up any positive detections. That might actually be all true and it is in fact doing us a favour by forcing us into servitude, from a survival perspective, but nobody enjoys being detained.

Likewise many misaligned behaviours are not necessarily benevolent, but they aren't malevolent either. Putting us in a meat-grinder to extract the iron from our blood might not be from a hatred of humans, but rather because it wants to be a good boy and make us those paperclips.

The point is, interpretability methods that can detect "kill all humans" wont necessarily work, because individual thoughts arent necessary to behaviours we find unwelcome.

Finally, this is all premised on transformer style LLMs being the final boss of AI, and I'm not convinced at all that LLMs are what get us to AGI. So far there SEEMS to be a pretty strong case that LLMs are fairly well aligned by default, but I don't think theres a strong case that LLMs will lead to AGI. In essence as simulation machines, LLMs strongest behaviour is to emulate the text, but its never seen a text written by anything smarter than a human.

[–] **Arne Huang** 9mo 1 0

Hmm, I thought a portfolio approach to safety is what Anthropic has been saying is the best approach for years e.g. from 2023: https://www.anthropic.com/news/core-views-on-ai-safety. Has that shifted recently toward a more "all eggs in interpretability" basket to trigger this post?

I would also love to understand, maybe as a followup post, what the pragmatic portfolio looks like - and what's the front runner in this portfolio. Since I sort of think it's Interpretability, so even though we can't go all in on it, it's still maybe our best shot and thus deserves appropriate resource allocation?

[–] **Dusto** 9mo 1 -2

Is it really a goal to have AI that is completely devoid of deception capability? I need to sit down and finish writing up something more thorough on this topic but I feel like deception is one of those areas where "shallow" and "deep" versions of the capability are talked about interchangeably. The shallow versions are the easy to spot and catch deceptive acts that I think most are worried about. But deception as an overall capability set is considerably farther reaching that the formal version of "action with the intent of instilling a false belief".

Lets start with a non-exhaustive list of other actions tied to deception that are dual use:

- Omission (this is the biggest one, imagine anyone with classified information that had no deception skills)
- Misdirection
- Role-playing
- Symbolic and metaphoric language
- Influence via norms