Tab 1

AI 2037: Predicting AI disorders

Applications for this AI Safety Camp project are live! Apply here.

Summary

AI 2037: AI akrasia (and other disorders) - Abram Demski, Steve Petersen, Sahil - YouTube

Who could argue against Eliezer Yudkowsky, Nate Soares? Daniel Kokatajlo and the AI 2027 team?

One thing that almost all counterperspectives to the above folk have in common is that they strawman the deep generators of AI risk.

Some counterperspectives bring up what's lacking in AI abilities today. They say AI doesn't seem to really understand what's going on. AI will never be embodied like humans, or at least is unlikely to for a long time.

But when probed further, these arguments tend to have some motivated reasoning, implicit biochauvinism or misunderstanding of computationalism, reluctance to project into the future and continually surprised by AI developments, or attachment to some specific paradigm they're excited about (like neurosymbolic methods, for example). They seem to conveniently or inadvertently miss fundamental insights about consequentialist optimization, or instrumental convergence, or nearest unblocked strategy, or economic incentives, or deep deception, or one-shot wicked problems, etc. etc. despite the many articulations, now even made available at a popular level in what is very likely going to be a best-selling book.

And yet.

In short, this is a project to explore a **specific line of argument**, of how **AI will struggle to have cognitive, executional, and bodily integrity.** This is akin to the odd persistence of hallucinations and other silly errors LLMs make today. Absurd, I know, but so are hallucinations and agentic failures. Moreover, this "struggle" may persist (at least, for a little while, hence the playful number '2037') without contradicting trends of increasing intelligence, efficiency, general agenticness, etc.

We aim to outline very clearly both the reasoning, evidence, and predictions that come of this, including counters. Most importantly, we do this not by strawmanning, but by pointing to subtle metaphysical assumptions surrounding *replaceability,* generalization power, and embeddedness. And it certainly doesn't imply everything is going to be okay.

Many of us (including those previously at MIRI, DeepMind, FHI, and actual academic philosophers) have been talking about this amongst ourselves for almost two years now, and finding that despite the arguments being metaphysically involved, we have surprisingly concrete empirical predictions ready for forecasting, threat modeling, and worldbuilding (and even design and engineering)! If that range is your sweet spot, or you are surprisingly relieved, infuriated or almost-indifferent-but-still-here about this whole thing, read on for an overview of the arguments and plan. >>>

The non-summary

There are many ways to motivate this, none of them too easy to transmit. Rather than a synthesis, it helps to see them from a few perspectives:

Perspective: Projecting Integrity

In one slogan: "you can't just *hack* together integrity".

In sum: AI will be bad at self-preservation not because it will be bad at preservation, but because it will be bad at "self" (from our perspective of "self").
Example: Humans are bad at self-preserving their atoms. The reason is that humans don't *identify* with our atoms, even though that is real and physical.
Argument:

1. Premise: It is hard to make intelligence care about particular things (this is obvious to many who work in alignment).
2. Premise: **In particular, matching the "self" or *integrity* of AI systems with that of living beings will also be *hard*.**
   Self-preservation, self-improvement, self-location (situational awareness), self-drivenness (autonomy), self-alignment, all factor through having a notion of "self" that determine various sorts of integrity.
   Keep reading for more on "integrity" and "hard".
3. Conclusion: AI will have strange disorders of integrity (cognitive, executive, bodily) for longer than you'd think, akin to hallucinations.

The key words to understand are italicized in bullet 2 above: "integrity" and "hard". Typical arguments for autonomy risk tend to assume that integrity of AI systems will come about automatically or naturally, most often without realizing that they're making such an assumption. Let's dig into this.

“Integrity”

The three types of integrity (not exhaustive) mentioned are:

- **Cognitive integrity**: similar to what we might call mental clarity or coherence. In humans afflicted with memory issues, brain fog, confusion, confabulation, dissociation, depersonalization, derealization, sensory processing problems, hallucinations, psychosis. In AI, “hallucination” is already a common term, but contrived reasoning in reasoning models and losing coherence in long contexts or other kinds of spikiness also count. To quote Nate Soares from Sep 2025: “It sure is interesting that they’re getting an IMO gold medal while also falling down on these sorts of things.”
- **Executional integrity**: Even if able to consider and make progress on ideas, there can be trouble in execution. In humans, this is common in those with various sorts of executive dysfunction such as ADHD, depression, trauma, savantism or reading too much self-help but failing to put it into action. In AI, this might look like various issues in integrating the “intelligent” bits with the “actuating” bits. Even ChatGPT claiming “I have now fixed the issue” while repeatedly failing is an example of this.
- **Bodily integrity**: A numb or negligent connection with the physical substrate of the system itself. In humans, this can present as “lost in your head”, dissociation, nerve damage, pain insensitivity syndrome, body dysmorphia, xenomelia, or just poor concern and care for the body. In AI, this might look like not caring about the hardware, or indifference to being deleted or closed, or dependence on humans to support and secure the physical setup.

The three buckets above are suggestive, the details of which we might flesh out over the course of the three months.

We are **not** saying that we should think of AI as being analogous to humans. In fact, the whole point of the project is to **point out implicit analogies *so that* we can “disanalogize” them.** Human failures are at best a nice intuition pump for conceiving of the possibility of integrity failures in general, not for equating silicon-based systems with our own animal make-up.

For now, having some notion of *integrity* that is free(ish) from anthropocentrism may be helpful in a few different ways, depending on where you’re coming from. For example:

- *Integrity* may supply an anchor to a version of *coherence* that isn’t restricted to the domain of logical inference or information processing, encompassing more physical, 4E characteristics.
    - (If you care about that sort of thing, it also means you might be able to bring some decision-theoretic considerations into a less GOFAI like arena.)
- We hope to be able to check conceptualizations implicitly related to “self” (such as self-preservation, self-improvement, autonomy etc.) and work with them without either a) automatically projecting our notions of self or b) getting caught up in endless discussions of the nature of “self”.

    - More broadly, keeping this in mind may bring to fore background theories of *identity* and so also *replaceability* (eg. what isn't crucially "me" and can be "equally" substituted for something better). This is highly relevant to the *functionalism* perspective in the appendix.
    - Unsurprisingly, notions of what makes something real and meaningful also play into this. We might replace or enrich conversations where the subtleties of *referentiality* matter. This is going on whenever we attempt to demarcate "real" or "actual" or "physical" or "deep" from "simulation" or "information" or "fake" or "shallow" or "cosmetic". For example, if someone asks you for a paper, but you have a digital PDF on your laptop, do you still possess an "actual" paper? It depends on context.
- Another way to see this is as avoiding an *anthropomorphic* homunculus fallacy. That is, we make the error of modeling systems as routing through some suspiciously human "true self". This can be quite subtle. Anytime the words "doing", "thinks", "cares", "believes", "understands", "says" or even "it" show up when referring to AI systems, the words sneak in a kind of integrity that we're used to in humans (or animals, more generally). The background theory is that if, say an LLM "believes" X, then "its actions" will reflect this. But we see this fail to have the integrity connoted by these words, leaving us perplexed.
    - The same goes for "agency". We want to be able to talk about equipping machine learning models with greater ability to interact with the wider world, without assuming that such ability will be utilized in a way that seems sane or adequate from our POV. In addition, it might be easier to talk about robustness or optimization failures without models of power/intelligence/action that route through some localized center in the way that "agency" tends to bring to mind.
    - When speaking about a system's "capabilities" and "propensities", and what has been "internalized" or what "generalizes", we are expecting the system to have a certain robustness. It's perhaps cleaner to talk about the nuances of this integrity and the lack of it, in the functioning of the system as a whole.

A lot of this might seem like pointless philosophical nitpicking, but the details are (as we'll see) at the heart of the issue.

"Hard"

A lot of these issues are met with "I don't see integrity failures as fundamental bottlenecks", which is understandable. The implication is that increase in intelligence, scaffoldings, context-length, data, revenue, etc. will result straightforwardly in greater integrity. In other words: *scale solves this*.

Let's take a particular example: of solving *bodily integrity*. Bodily integrity disorders might manifest as the AI system does not seem to care autonomously about its physical configuration. A solution proposal is to just hook up the LLM-agent (or whatever) with various cameras to its GPUs. This seems simple enough. Why would this project predict that that wouldn't work? Think for a moment if you'd like.

.

.

.

The issue is with "just hooking up". This assumes away the hardness of integration of the *relevance* of the videofeed with the other components of the AI system that function together.

"The AI knows but doesn't care" is an old and still useful way to convey the difficulty of alignment of machines with humans. In this case, we're talking about the alignment of the machine with itself (better written "it" "self", to remind ourselves that this is our own conceptualization). When positing cheap hacks for integrity, consider whether you're sneaking in key solution concepts to the alignment problem.

"But, these two are quite different issues," you say, "You're confusing the difficulty of *aligning the machine to a specific value target (i.e. steering the AI)* with the inevitability of *instrumental convergence of resources and survival (i.e. AI system's own integrity)*."

Remember, our example started with something that lacked specific aspects of integrity in the first place. You might hook up the camera feed, but will "it" be able to "care" about the pixels in the camera feed in a way that shows up in "its" actions? Maybe it has an inner optimizer, powering its creativity; will that creativity be more than indifferent to your attempt at making it *care* about the "real" world? Even if the "understanding" of the cameras and GPUs is there at a verbal level, the motivation and execution might remain disordered, in the same way that simply "hooking up" multiple modalities without training on multimodal data doesn't transfer well.

In other words: some LLM-agent's notion of "my body", and more importantly, what to *do* about that, wouldn't generalize straightforwardly.

Most importantly, this sort of deficiency is no mere *irrationality* that is selected out. Are you irrational when you fail to care about hair that's being cut? Are you irrational when you don't protest when someone takes a picture of you? Is it irrational to die for your country? To say that a "brain transplant" is actually a "body transplant"? This depends on what one identifies as, which is circumrational *and* complex.

Note that to do the rational thing is to pick between alternatives that are instrumental to the ends. What is "instrumental" and subject to efficiency is that which is *replaceable*. But **what you identify with *defines* what is (ir)replaceable** to you. To apply selection arguments *assumes* a background theory of replaceability. Any time you propose that "X is better to achieve Y", you are saying "X is *interchangeable* among a class of things that serve Y". Scaling rationality, optimization, data etc. do not furnish one with a "better" theory of identity. That's a type error. As we've noted, "better" *assumes* a background theory of what is equal/replaceable with what.

In other words: some LLM-agent's notion of "my body" and what to do about that wouldn't generalize *even for efficiency reasons*.

So, yes, to you it might look unbelievably stupid to not care about someone attacking your GPUs. The LLM components might even strongly agree with that. But whether it is able to be "hooked up" in a way that is *creatively* caring, whether an AI system is moved to perform Gendlin-style focusing on its body, is a different question altogether. That's why the claim of "it's bad at 'self' from *your* perspective." Many things we take for granted here depend in messy ways on contingent aspects of your particular theory of identity as a biological animal.

The slogan is: you can't just *hack* integrity. You can't just download integrity, connect it up, simply tell it, plug-and-play, define it away, expect it to be like yours, or simply add more data(streams).

It needs to undergo the *process of integration*, akin to the process of friendship. The "understanding" in the LLM that the videofeed "matters" doesn't *just* "generalize" to action; that would be like reading about mathematics and skipping the exercises. This may sound implausible, but if you've been surprised by how persistent hallucinations have been, you may be open to problems of this sort, even as stuff gets quite smart. **This is not the same as Gary Marcus's story** on what to expect**.**

Now, it **isn't impossible** to induce integration, any more than making "nice" AI chatbots is impossible. It is the case that its notion of integrity doesn't match ours without *us* doing the *work* of building it in, and perhaps with great difficulty in non-shallow ways. The next section will say a little bit more about this.

So arguments of the sort "well, *we* won't want robots that can't execute well" are valid efficiency arguments, but they route through us, and potentially quite slowly, and still do not "just generalize" to all the kinds of integrity. Hence "2037" (the exact number isn't significant) rather than "everything is going to be okay". In fact, some of the work we'll be doing is articulating the terrible, perhaps *worse* risks from *participation* and *diffusion*, rather than some centralized autonomous singleton. *Codependence* risks, rather than independence risks.

Even if this happens very quickly and rapidly turns into independence/autonomy risks, this period of participation might shape and overhaul our conceptualization and methodology of the whole issue of AI. That's out of the scope of this project, but there is a sister project if you're interested in that.

There's a lot to say here and a lot of counters to address; you're not expected to be convinced yet. A major game will be to "spot the integrity magic" in arguments of rapid acceleration, either by ignoring the necessity of integration fundamentally or projecting one's own background theory of integrity.

Some empirical things we might chase up here:

- Congenital pain insensitivity
- Cataloguing specifics of AI spikiness/disorders
- Limitations of Agents Simulated by Predictive Models
- Multimodal training data needed (From Linguistic Giants to Sensory Maestros: A Survey on Cross-Modal Reasoning with Large Language Models)
- Language models cannot reliably distinguish belief from knowledge and fact | Nature Machine Intelligence

Please check the Appendix at the bottom for the other perspectives.

An imaginary abstract

Here's an abstract to fill out, quoted from here:

Many AI safety folk have rightly raised issues of corrigibility. That once the target(/values) of an optimization process get locked in, it becomes extremely sticky. Attempts to change are up against the full optimization power behind fulfilment of the target. If this power goes beyond our abilities, any clever strategies to undermine it are just an enactment of our weaker optimization powers pitted against a stronger one, and will, tautologically, fail.

We aim here to bring to fore a subtler version of this issue: *corrigibility of relevance*. Where stickiness of value comes about not via strong monomaniacal attraction to the target, but by indifference to anything outside of what is considered meaningful. Indifference not as a side effect of strong optimization (such as in *The AI knows but does not care*), but by not entering attention. Indeed, this applies even in the absence of an overall optimizing process. We also point out how, contrary to intuitions that this makes AI more dangerous, it could also imply *less* conflict. We point out that optimization processes that occur *within a zone of relevance* are less likely to be rivalrous with others, contra broad instrumental convergence. We carefully argue that difficulty of corrigibility of relevance implies, further, that zones of relevance are unlikely (though not impossible) to expand, since there is no motivation to.

We also couch zones of relevance in naturalistic terms. This involves a new concept, of *integrity*, which is an embedded extension of *coherence*. The importance of the specific, physical modalities of "skin-in-the-game" to *relevance* is explored, via connections between physical integrity and mental

integrity. We take up very concrete questions (such as "are hardware level architectural changes needed in order for shutdown to be a real problem?") from this lens, and argue for a *fractal consequentialist* model. This makes the safety-vs-capability dichotomy untenable (in a new way), giving rise, for example, to "fractal sharp left turn". We list many predictions of this model that are opposite to the original sharp left turn threat model, giving rise to novel dangers that resemble S-risks from partial alignment.

Owing to issues of *integrity*, "simply" hooking up an intelligent, "unagentic" model with an agentic wrapper (eg. "GPT-7 may not be dangerous, but GPT-7 plus two thousand lines of python can destroy the world") is shown to be a lot more difficult. We explore, with examples, how simply "plugging in" does not lead to *integration*, and this lack of integrity creates lacunae of relevance that can be exploited to cause shutdown, no matter the cognitive powers of the machine. We cautiously make analogies to lack of integrity in humans (often manifesting as mental disorders, dissociation, and savantism). We note that disanalogies are more important because of differences in physical make-up of biological life vs machines, and how that gives us more time to deal with dangerous versions of the problems of shutdown / risks of autonomy. We end by attempting to point at *the human/animal zone of relevance*, which is quite philosophically challenging to do from the vantage of our own minds, but extremely important as we create minds very different from ours.

Output

Project Rhythm and Plan

The plan for the three and a half months is to assimilate, investigate, and prognosticate. To this end, we'll have a few tracks, with some absolutely amazing folk:

1. Weekly Monday Fishbowl conversations with Abram Demski and Sahil on introducing this whole issue, with a focus on taking the cruxes and talking points and distilling them into verifiable, testable, prediction-ready bets.
2. Philosophy of predictions seminar, macrostrategy and case studies with TJ
3. Fortnightly reading group and discussion on the philosophical literature on these and related questions with Rosa Cao and Steve Petersen.
4. Vignettes of the future with Aditya Prasad

Participants will join these discussions and work on multimodal outputs, including specific (and conditional) bets and forecasts, explainers, responses to and reviews of counter-arguments, collated evidence for and against.

Assorted project goals

The project's aim is to investigate these issues of reference and integrity, and chart out the implications for likely and possible real-world scenarios. This investigation will involve the following arms:

1. Turning the fine-grained functionalism perspective to testable hypotheses regarding the integrity (and competencies) of upcoming AI systems.
2. Mapping out the space of integrity disorders, derived from such theory, that can occur in AI systems, that identify key distinctions along dimensions such as (but not limited to):
    a. Whether it is a property that is hard to select out (ie. train against), or unreachable within the search space entailed by current training paradigms?
    b. Whether it undermines plausibilities of risks from runaway agency, or whether it merely implies the instability/volatility of the underlying agency (without restricting its plausibility)?
    c. How much, and in what ways, does it inhibit economic adoption of the systems?
    d. etc.
3. Developing frameworks for scenario building of medium-term futures with AI systems that suffer from such limitations, despite being capable in some other ways, including (but not limited to): economic models, information ecology models, and ethology of synthetic systems.
4. Developing a Scenario Choice Model, charting the space of possible intermediate scenarios, and how different intermediate choices about AI development might relate to different trajectories.

Risks and downsides (externalities)

The main risk would be that we investigate this poorly, and come to premature conclusions with misleading backing that does not paint a useful story of the future. To mitigate exactly this, this project focuses on putting in the work of scout mindset. Additionally, we have a dedicated track for putting thought into the meta considerations of prognosticating!

Acknowledgements

Apart from those already mentioned, Ramana Kumar has been a contributor to some early conversations.

Team

Team size
Since this will be a highly parallelized discussion space with multiple subteams, we can take anywhere between 1 and ~20 people.

Project Lead

The main project lead is Sahil, with co-leads Abram and TJ. Sahil will put about 10 hours per week on this. Abram, about 3 hours. TJ about 8 hours.

Roles & Skill requirements

If you have experience or interest in forecasting (especially of the impact and trajectory of AI), and are open to somewhat involved philosophical bearings on the subject, you are a top pick for Sahil/Abram's track. If you saw some of the appendix and thought of many citations that we could have included but inexplicably didn't, you are probably an excellent fit.

If you enjoy writing in the formats of either of: distillation, worldbuilding, story-writing and are up for digesting these conversations and excavating the many transcripts and recordings we have on the subject, you are also an excellent fit.

Familiarity with existing threat models and future scenarios are most often a plus, unless it prevents open-mindedness towards this style of argument. That said, if you're looking at this section, you're probably open-minded enough to begin!

Additionally, for TJ's tracks:

Personality Type 1:

If you familiarity with, or an active interest in, close examination of philosophical perspectives and claims about AI and agency, discovery of latent conceptual distinctions, and refining hypotheses;

Familiarity with contemporary ML paradigms is highly preferred, with at least some willingness to directly engage with ML paradigms philosophically is desirable.

Personality Type 2:

If you have familiarity with either:

a) complex systems analysis, with emphasis on large scale assemblages, such as: information epidemiology, systems ecology, statistical biology, or
b) Macro and meso-economics,

Along with an enthusiasm for engaging with novel philosophical territories, thinking seriously about questions of agency in AI systems, and comfort with navigating between semi-formal and conceptual registers.

Appendix

Perspective: Grown Connections

Connor Leahy once said (personal communication) “GPT-7 may not be dangerous, but GPT-7 with 2,000 lines of python could destroy the world.”

Why might this not fly? Take a moment.
.
.
.
.

There are very few people worried about the Software Alignment problem, even though every machine learning model is technically “software”.

We all know there is a substantial difference between programs and *grown programs,* i.e. machine learning models. We might not be able to quite define the difference between them that passes philosophical muster, but practically, it’s quite apparent. No one can write a program to recognize hand-written digits; you have to grow one. It’s not ordinary programs that are scheming (even though ordinary programs can be deceptive), it’s the grown ones.

So let’s take that difference, and consider also, *grown connection*s. Many domains, upon enough maturation, realize that it isn’t the objects of their framework that contain meaning, but the relationships between them. Whether or not that floats your boat, you might still like to be open to the idea of grown connections, rather than only modular connections.

The thesis is that some connections need to be grown, and are not simply plug-and-playable. A canonical example of modular connections is a port, with literal plug and play. A canonical example of a grown connection is a friendship between two intelligent actors. As with grown programs, there may not be a formalizable distinguisher for these, but that doesn’t mean they aren’t incredibly different.

If you *plug* a model into some scaffolding via code like `llm_reasoning_call(world_state, goal)` then pass it onto *ports* to crypto wallets or servomotors or whatever, there remains the same question: will it have integrity, or will it fall apart? How much can intelligence do, without having *grown* connections with the rest of the world?

The theme of replaceability from the previous section continues here: these modular ports that you can plug into, are what you can easily plug out of. “Transactional relationships” is our social colloquialism for this same phenomenon. When push comes to shove, these relationships don’t hold together. Or if they do (as in corporations), they seem to really lack integrity (and appear otherwise insane), in the absence of wholesome relationships between the people involved.

Living organisms tend to have more entangled architectures than the swappable ones that we design in our machine systems, across the whole stack from hardware to software. Terence Deacon says organisms are “indecomposable”, in that “the various parts of an organism require one another because they are generated reciprocally in the whole functioning organism.”

So an even better example of grown connections is a single organism holding together. This is true for bodily integrity. A very interesting empirical fact about you is that we could take nearly any square inch of your body and cause so much pain that you drop anything else you’re doing just to attend to that. This is a sign of your integrity, of the preciousness or *irreplaceability* of what you consider you. It’s all of us or none of us, says you/your body, for the trillions of tiny animals that you are.

This seems less true of “AI bodies”, with humans out of the loop. They are designed to be highly modular and controllable. And this fact, of not having parts that entangle together in fashion, actually offers a constructive plan for shutdown (a famously difficult challenge): disintegrate at these weak points of integrity. Note: these may only be indirectly arranged around physical modularity, which circumscribes the various *grown connections* that may have been forged.

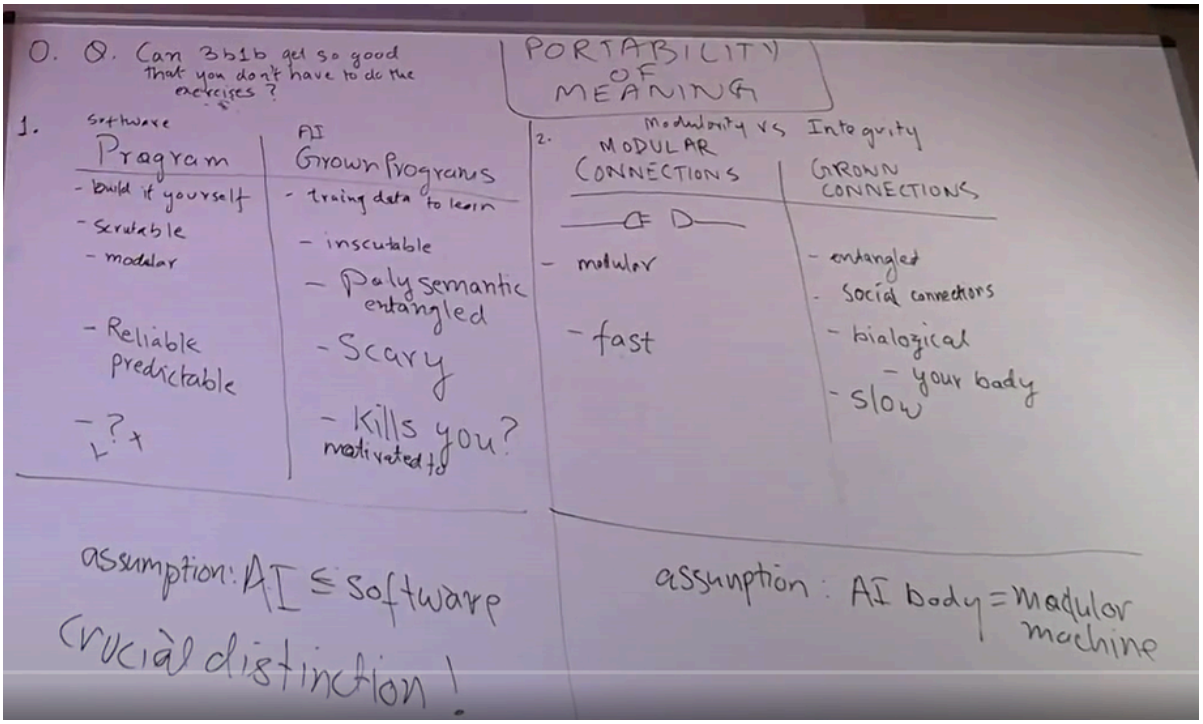
Once again, this state of affairs of low-integrity (from out POV) may not last forever, but it certainly points at something that can’t be simply “scaled”. People often ask “are we bottlenecked by compute, or insights, to AGI?” And here there’s a third alternative: maybe hardware-level architecture changes are necessary for autonomy risks. Even if we have the insights, implementation might not be diffused in quite as simple as “just add more compute” sort of way.

Grown programs, of course, already exhibit this entanglement. They are notoriously hard to interpret, decompose, etc. Cognitive modularity seemed important for GOFAI systems, but less so for deep messy neural networks with trillions of parameters. This grownness of AI models does not seem to extend to their “bodies” without us in the loop, which we’ll explore a little bit in the next section.

As for our own minds (not just bodies): fluency and competence with knowledge, even in humans, does look like deep integration, rather than some machine-like modular portability. As Von Neuman said, in mathematics, you don’t understand, you get used to. Another

empirical fact: there has never been a genius mathematician born, who did not have to do the exercises. The mathematics examples are especially interesting because they are both “pure information” and “formal”, and yet they require “exercise”. You have to train *relevance* especially, not just storage.

If you want to explore it more starting from the *grown connections* perspective, you can watch this talk, where we start with the question “Can 3Blue1Brown get so good that you don’t have to do the exercises?”. A screenshot is below:



Some empirical things we might chase up here:

- Deacon’s “indecomposable claim”
- Bio organisms with modularity
- [2504.17950] Collaborating Action by Action: A Multi-agent LLM Framework for Embodied Reasoning
- Geniuses like Ramanajun, are they exceptions?

Third: Functionalism, reference, relevance, care

Functionalism, in philosophy of mind, is the idea that it’s not the substrate but the *function* that a system implements that makes something a mind. This is most often met in the stronger form of computationalism (that the mind is a computer) that many today believe, especially in AI communities of all sorts.

We're less interested in the question of consciousness here, and more in the *neglect of substrate*. This is not a new critique of functionalism, see this section in the SEP entry, for example, emphasizing the essentialness of the body.

In particular, fine-grained functionalism takes an intermediate path, saying that it is the fine-grained structure of the matter that's relevant, rather than specific substrate s. It thereby avoids charges of biochauvinism without being as liberal as functionalism ("too liberal" is another standard critique of computationalism).

[kung-fu discussion]

Some empirical things we might chase up here:

- Invent new colors [or this should be in reference]
- Knowledge integration and BCI (kung fu examples)
- Experiments beyond rubber-hand and phantom limb
- Math geniuses that didn't do the exercises
- Case studies on transfer of embedded knowledge
- Simulation and synthetic data effectiveness

Sharp Left Turn, Fractally / Self-alignment and reverse alignment / simophone

Many arguments here do look like "AI will fail to self-align", which is another way to talk about integrity, albeit in a more mysterious way.

Tab 2