LeCun's "A Path Towards Autonomous Machine Intelligence" has an unsolved technical alignment problem

39

by **Steve Byrnes** 8th May 2023

Reinforcement Learning AI Frontpage

Summary

- This post is about the paper **A Path Towards Autonomous Machine Intelligence (APTAMI) by Yann LeCun**. It's a high-level sketch of an AI architecture inspired by the brain.
- APTAMI is mostly concerned with arguing that this architecture is a path towards more-capable AI. However, it is also claimed (both in the paper itself and in associated public communication) that this architecture is a path towards AI that is "controllable and steerable", kind, empathetic, and so on.
- I argue that APTAMI *is* in fact, at least possibly, a path towards that latter destination, but only if we can solve a hard and currently-unsolved technical problem.
- This problem centers around the **Intrinsic Cost module**, which performs a role loosely analogous to "innate drives" in humans—e.g. pain being bad, sweet food being good, a curiosity drive, and so on.
- APTAMI does not spell out explicitly (e.g. with pseudocode) how to create the Intrinsic Cost module. It offers some brief, vague ideas of what might go into the Intrinsic Cost module, but does not provide any detailed technical argument that an AI with such an Intrinsic Cost would be controllable / steerable, kind, empathetic, etc.
- I will argue that, quite to the contrary, if we follow the vague ideas in the paper for building the Intrinsic Cost module, then there are good reasons to expect the resulting AI to be not only unmotivated by human welfare, but in fact motivated to escape human control, seek power, self-reproduce, etc., including by deceit and manipulation.
- Indeed, **it is an open technical problem to write down *any* Intrinsic Cost function (along with training environment and other design choices) for which there is a strong reason to believe that the resulting AI would be controllable and/or**

motivated by human welfare, while also being sufficiently competent to do the hard intellectual tasks that we're hoping for (e.g. human-level scientific R&D).

- I close by encouraging LeCun himself, his colleagues, and anyone else to try to solve this open problem. It's technically interesting, very important, and we have all the information we need to start making progress now. I've been working on that problem myself for years, and I *think* I'm making more than zero progress, and if anyone reaches out to me I'd be happy to discuss the current state of the field in full detail.
- …And then there's an epilogue, which steps away from the technical discussion of the Intrinsic Cost module, and instead touches on bigger-picture questions of research strategy & prioritization. I will argue that the question of AI motivations merits much more than the cursory treatment that it got in APTAMI—*even given* the fact that APTAMI was a high-level early-stage R&D vision paper in which every other aspect of the AI is given an equally cursory treatment.

(Note: Anyone who has read my Intro to Brain-Like AGI Safety series° will notice that much of this post is awfully redundant with it—basically an abbreviated subset with various terminology changes to match the APTAMI nomenclature. And that's no coincidence! As mentioned, the APTAMI architecture was explicitly inspired by the brain.)

1. Background: the paper's descriptions of the "Intrinsic Cost module"

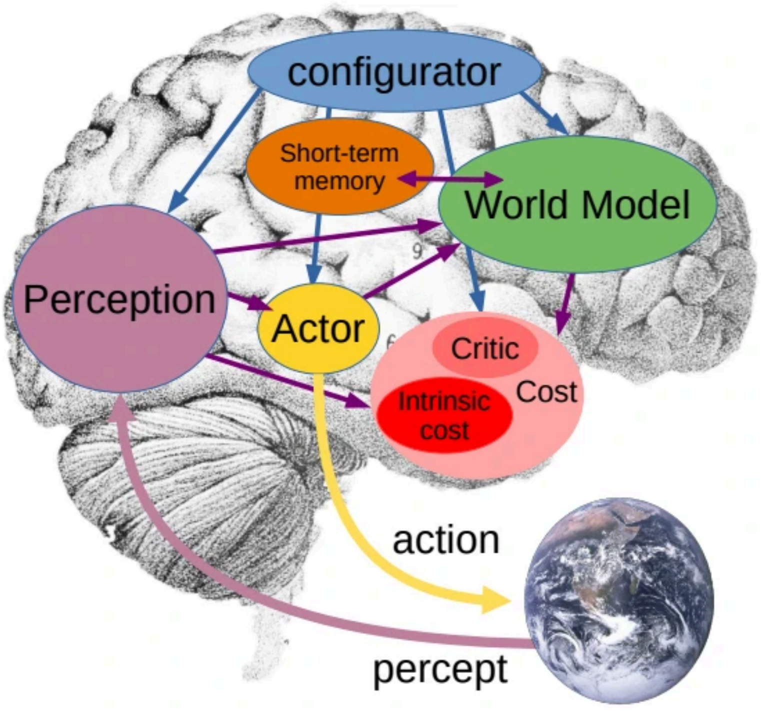


Figure 2 from APTAMI. "Intrinsic Cost" is the red oval towards the bottom-right.

For the reader's convenience, I'll copy everything specific that APTAMI says about the Intrinsic Cost module. (Emphasis in original.)

> PAGES 7-8: The *Intrinsic Cost module* is hard-wired (immutable, non trainable) and computes a single scalar, the *intrinsic energy* that measures the instantaneous "discomfort" of the agent – think pain (high intrinsic energy), pleasure (low or negative intrinsic energy), hunger, etc. The input to the module is the current state of the world, produced by the perception module, or potential future states predicted by the world model. *The ultimate goal of the agent is minimize the intrinsic cost over the long run.* This is where basic behavioral drives and intrinsic motivations reside. The design of the intrinsic cost module determines the nature of the agent's behavior. Basic drives can be hard-wired in this module. This may include feeling "good" (low energy) when standing up to motivate a legged robot to walk, when influencing the state of the world to motivate agency, when interacting with humans to motivate social behavior, when perceiving joy in nearby humans to motivate empathy, when having a full energy [supply] (hunger/satiety), when experiencing a new situation to motivate curiosity and exploration, when fulfilling a particular program, etc. Conversely, the energy would be high when facing a painful situation or an easily-recognizable dangerous situation (proximity to extreme heat, fire, etc), or when wielding dangerous tools. The intrinsic cost module may be modulated by the configurator, to drive different behavior at different times.

> PAGE 14: The intrinsic cost module (IC) is where the basic behavioral nature of the agent is defined. It is where basic behaviors can be indirectly specified.
>
> For a robot, these terms would include obvious proprioceptive measurements corresponding to "pain", "hunger", and "instinctive fears", measuring such things as external force overloads, dangerous electrical, chemical, or thermal environments, excessive power consumption, low levels of energy reserves in the power source, etc.
>
> They may also include basic drives to help the agent learn basic skills or accomplish its missions. For example, a legged robot may comprise an intrinsic cost to drive it to stand up and walk. This may also include social drives such as seeking the company of humans, finding interactions with humans and praises from them rewarding, and finding their pain

> unpleasant (akin to empathy in social animals). Other intrinsic behavioral drives, such as curiosity, or taking actions that have an observable impact, may be included to maximize the diversity of situations with which the world model is trained (Gottlieb et al., 2013).
>
> The IC can be seen as playing a role similar to that of the amygdala in the mammalian brain and similar structures in other vertebrates.
>
> To prevent a kind of behavioral collapse or an uncontrolled drift towards bad behaviors, the IC must be immutable and not subject to learning (nor to external modifications).

> PAGE 44: What is the substrate of emotions in animals and humans? Instantaneous emotions (e.g. pain, pleasure, hunger, etc) may be the result of brain structures that play a role similar to the Intrinsic Cost module in the proposed architecture. Other emotions such as fear or elation may be the result of *anticipation of outcome* by brain structures whose function is similar to the Trainable Critic.
>
> The presence of a cost module that drives the behavior of the agent by searching for optimal actions suggests that autonomous intelligent agents of the type proposed here will inevitably possess the equivalent of emotions. In an analogous way to animal and humans, machine emotions will be the product of an intrinsic cost, or the anticipation of outcomes from a trainable critic.

2. As described in the paper, several components of the AI's Intrinsic Cost are directly opposed to AI controllability and prosociality

In AI alignment discourse, we often talk about **"instrumental convergence"**°. If an AI really wants a thing X, then for *almost any X*, it will find it instrumentally useful (i.e., useful as a means to an end) to get control over its situation, gain power and resources, stay alive, prevent its desires from being exogenously changed, and so on. In Stuart Russell's memorable quip, if an AI really wants to fetch the coffee, well, "you can't fetch the coffee when you're dead", so the AI will fight for self-preservation (other things equal).

APTAMI specifically mentioned "hunger", "pain", and "curiosity" as three likely components of Intrinsic Cost (see Section 1 excerpts). All three of these have obvious "instrumental convergence" issues. Let's go through them one at a time:

- If an AI is motivated to avoid hunger (say, implemented in source code by checking the battery charge state), and the AI reasons that humans might not want to recharge it, then the AI will be motivated to get power and control over its situation to eliminate that potential problem, e.g. by sweet-talking the humans into recharging it, or better yet maneuvering into a situation where it can recharge itself without asking anyone's permission.
- If an AI is motivated to avoid pain, and the AI reasons that humans might cause it to experience pain, or be unable or unwilling to help it avoid future pain, then the AI will likewise be motivated to get power and control over its situation to eliminate that potential problem.
- If an AI is motivated by curiosity, and the AI reasons that humans might fail to offer it sufficiently novel and interesting things to do, then the AI will likewise be motivated to get power and control over its situation, so that it can go satisfy its curiosity without asking anyone's permission.

Possible reply 1: "OK, granted, that's a real problem, but it's easy to fix, we'll just remove those three things from the Intrinsic Cost module."

My response: It's not so easy. For one thing, curiosity in particular is plausibly essential for the AI to work at all. For another thing, as mentioned at the top, it's not just about *these three specific* motivations. On the contrary, a wide variety of motivations lead to similar "instrumental convergence" problems, including important motivations like "wanting to design a better solar cell" that seem necessary for the AIs to do the things we want them to do.

Possibly reply 2: "Humans are motivated by hunger, pain, and curiosity, and can be perfectly lovely assistants and employees. Why would we be starving and hurting the AIs anyway?? Let's just treat our AIs well!!"

My response: Humans have a lot of other motivations besides hunger, pain, and curiosity, including intrinsic motivations to kindness, friendship, norm-following, and so on. I'll turn to those in the next section. If you've ever gotten to know a sociopath or narcissist, you′ll know that *they* have hunger, pain, and curiosity too, but it is *absolutely not the case* that if you just treat them with kindness then they will be kind to you in return! They might *act* kind and cooperative *as a means to an end*, e.g. to gain your trust, but that’s not what we want—that’s the kind of “cooperation” where they stab you in the back as soon as the situation changes. We want our AIs to treat kindness as an end in itself. And that doesn’t happen unless we explicitly build such an intrinsic motivation into them. So let’s turn to that next.

3. As described in the paper, the components of the AI’s Intrinsic Cost that are *supposed to* motivate intrinsic kindness, are unlikely to actually work

Before we even start, it seems like a *pretty dicey plan* to build an AI that has numerous innate drives that are *opposed* to controllability and prosociality (as described in the previous section), plus *other* innate drives that *advance* controllability and prosociality (as I’ll discuss in this section). Such an AI would feel “torn”, so to speak, when deciding whether to advance its own goals versus humans’. We can *hope* that the prosocial drives will “win out” in the AI’s reckoning, but we don’t have any strong reason to expect that they *will* in fact win out. Neurotypical humans are a fine illustration—we have both prosocial drives and selfish drives, and as a result, sometimes humans do nice things, and also sometimes humans advance their own interests at the expense of others.

However, it’s much worse than that, because I claim *we don’t know how to build prosocial innate drives at all in this kind of AI*.

It’s clearly *possible in principle*—there’s *some* mechanism underpinning those drives inside the brains of non-psychopathic humans—but I claim it’s an open problem how these drives actually work.

(Or in APTAMI’s terminology, it’s an open problem exactly what code to put into the Intrinsic Cost module such that the AI will have any prosocial or docile motivations at all.)

APTAMI’s “proposal” here is really just a passing description in a few sentence fragments. But worse than that, as best as I can tell, this cursory description is *not* pointing towards a viable proposal, nor one that can be easily remedied.

I’ll repeat the relevant excerpts from above:

> This may include feeling "good" (low energy) … when interacting with humans to motivate social behavior, when perceiving joy in nearby humans to motivate empathy, …
>
> … This may also include social drives such as seeking the company of humans, finding interactions with humans and praises from them rewarding, and finding their pain unpleasant (akin to empathy in social animals).

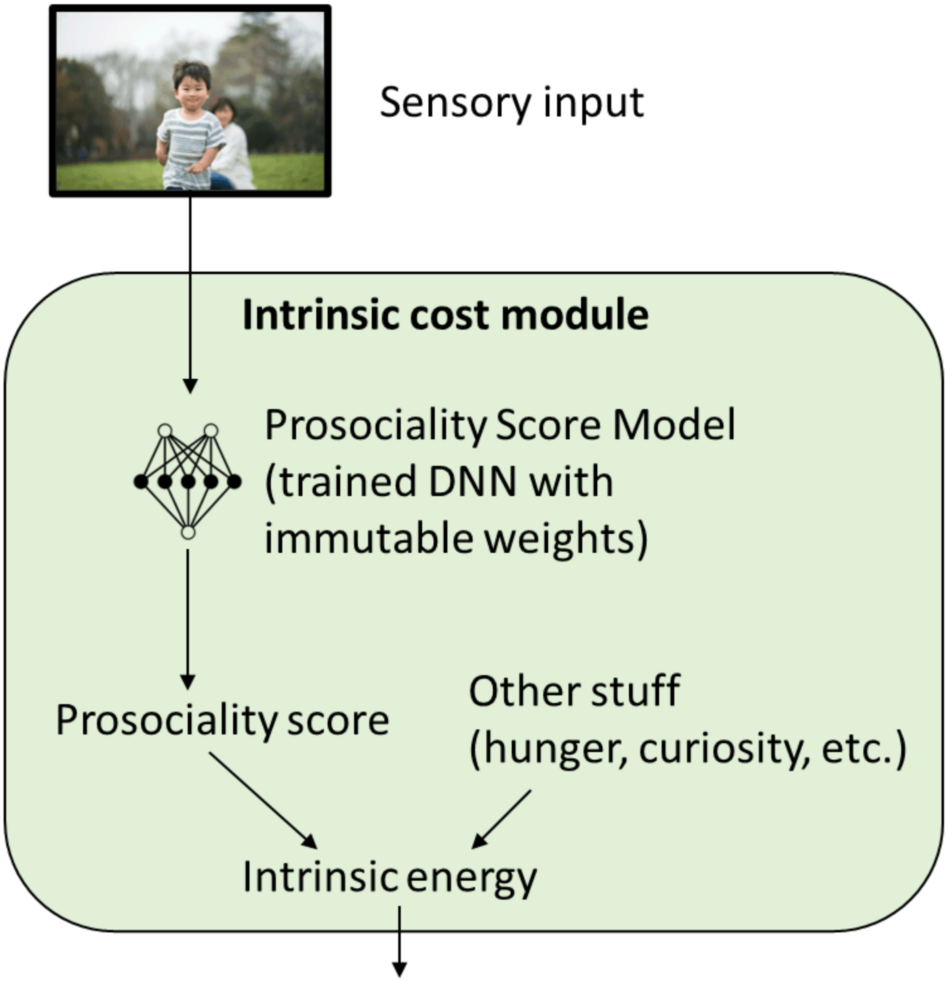
It's hard to respond to this because it's so vague. Different things can go wrong depending on how the implementation is supposed to work *in detail*. I can make some guesses, but maybe the response will be "No you moron, that's not what I meant". Oh well, I'll proceed anyway. If what I say below *isn't* what the author had in mind, maybe he can share what he *did* have in mind, and then I can revise my description of what I think the problems are, and maybe we can have a productive back-and-forth.

My attempt to flesh out what LeCun might have in mind:

The robot has "eyes" & "ears" (a video feed with sound). We get some early test data of the robot doing whatever (say, flailing around randomly), and then send the video feed to a bunch of humans (say, Mechanical Turkers) to manually label the video frames using the following rubric (drawn from the excerpts above):

- "Am I interacting with humans?" (1-10)
- "Are nearby humans experiencing joy?" (1-10)
- "Am I in the company of humans?" (1-10)
- "Is a human praising me right now?" (1-10)
- "Is a human in pain right now?" (1-10)

Next, we do a weighted average of these scores (the first four with positive weight, the fifth with negative weight) and use supervised learning to train an ML model that can take any arbitrary video frame and assign it a score. Let's call this trained deep neural net (DNN) the *Prosociality Score Model*. We freeze the weights of this classifier and put it inside the Intrinsic Cost module (to be added to the other terms like pain and curiosity, discussed above). Here we are so far:


Sensory input
Intrinsic cost module
Prosociality Score Model (trained DNN with immutable weights)
Prosociality score
Other stuff (hunger, curiosity, etc.)
Intrinsic energy

My (uncertain) attempt to flesh out the vague proposal in APTAMI. If this isn't what LeCun had in mind, I strongly encourage him to write out more specific pseudocode for the Intrinsic Cost module such that he thinks it will lead to an AI that has controllable and/or prosocial motivations. And then we can have a productive discussion about whether that pseudocode will actually work as intended.

Does this approach actually make an AI with prosocial motivations? I think the answer is a clear "no".

For starters, suppose the AI straps lots of humans into beds, giving them endless morphine and heroin IV drips, and the humans get into such a state of delirium that they repeatedly praise and thank the AI for continuing to keep the heroin drip turned on.

This dystopian situation would be, to the AI, *absolute ecstasy*—much like the heroin to those poor humans. The Prosicality Score Model would (perhaps—see below) give

- 10/10 for "interacting with humans",
- 10/10 for "being near humans experiencing joy",
- 10/10 for "being in the company of humans",
- 10/10 for "receiving praise", and
- 0/10 for "being around humans in pain".

Now, it doesn't immediately follow that the AI will actually want to start buying chair-straps and heroin, for a similar reason as why I personally am not trying to get heroin right now. But it certainly raises that kind of AI behavior as a salient possibility, and in this particular instance my guess is that something in this general class of "failures" would be pretty likely to actualize. At the very least, I see no strong reason to believe that they *won't* actualize.

On a more technical level, we face (among other things) the classic problem of out-of-distribution (OOD) generalization, and we face it not once but twice:[1]

- *Out-of-distribution generalization problem 1:* How does the Prosociality Score Model generalize from the supervised (human-labeled) examples to the AI's future perceptions —which might be far outside that training distribution?
- *Out-of-distribution generalization problem 2:* How does the *critic* generalize from its *past* observations of Intrinsic Cost to estimate the Intrinsic Cost of *future* plans and situations —which, again, might be far outside the distribution of its past experience?

These problems are made worse because they are *adversarial*—with the adversary being the AI itself! Let me explain this adversarial aspect via some toy examples:

Why is problem 1 an "adversarial" OOD problem? Here's a toy example. The AI might notice that it finds it pleasing to watch movies of happy people—because doing so spuriously triggers the Prosociality Score Model. Then the AI might find itself wanting to make its own movies to watch. As the AI fiddles with the settings in iMovie, it might find that certain texture manipulations make the movie *really really* pleasing to watch on loop—because it "tricks" the Prosociality Score Model into giving anomalously high scores.

What happened here was that the AI sought out and discovered "adversarial examples" for an immutable DNN buried deep inside its own "mind".

(That particular example doesn't seem very scary, until the AI notices that humans might want to turn off its weird-texture movie playing on loop. Then the situation gets "adversarial" in the more literal sense!)

Why is problem 2 an "adversarial" OOD problem? Here's a toy example. Imagine that the AI is deciding what to do, out of a very wide possibility space. For example, once we get AIs that can invent new technology, then the AI has access to actions that might wildly change the world compared to anything in history. Thus, if there are any anomalies where the critic judges a weird course-of-action as unusually low-intrinsic-cost, then we're in a situation where the AI's brainstorming process is *actively seeking out such anomalies*.

(From our human perspective, we would say "this plan is exploiting an anomalous edge-case in the critic". Whereas from the AI's perspective, it would say, "this plan is a clever awesome out-of-the-box way to solve every problem!!" You say tomato, I say to-*mah*-to.[2])

Needless to say, robustness to adversarially-chosen wildly-out-of-distribution inputs is an unsolved problem in ML. So it's probably safe to assume that, if we use the APTAMI plan (as I interpret it), the AI is probably going to wind up with weird and *a-priori*-unpredictable motivations. And this problem is *not* the kind of problem where we can just straightforwardly patch it once we have a reproducible test case running on our computers.

4. Conclusion

In Section 2 I argued that the AI (as described in APTAMI) will have at least *some* motivations (like hunger, pain, and curiosity) that run directly counter to controllability / steerability / prosociality, thanks to "instrumental convergence"°. This *might* be OK if the AI *also* has *other* motivations that create controllability / steerability / prosociality (as is the case in humans, who are sometimes cooperative despite *some* selfish innate drives).

However, in Section 3 I argued that it's an open problem to write out an Intrinsic Cost function that will lead to *any* motivation for controllability / steerability / prosociality, and that APTAMI says only a few words about how to solve this problem, and that what little it says does not seem to be pointing in a promising direction. Instead, the paper's proposal seems likely to lead to AIs with weird and *a-priori*-unpredictable motivations. Indeed, I'd guess that these weird unpredictable motivations are more likely to *contribute to* "instrumental convergence" effects than to push against them. And this problem would be very difficult to patch even if we had a working minimal test case on our computers, because wildly-out-of-distribution adversarial robustness is an open problem in ML, and there is no obvious better alternative approach.

So we have a very interesting, open technical problem here: **"Exactly what code should we put into the Intrinsic Cost module [in conjunction with other design choices, e.g. training environment], such that we have strong reason to believe that we'll be pleased with the AI that results?"** In fact, I myself have a full-time job in which I spend most of my days trying to work towards an answer to this question, and have been doing so for years. It is a very hard problem.

I think that LeCun himself is more qualified than most to work on this technical problem, and I think we already have all the information we need to make progress, so I would strongly encourage him and his colleagues to dive in. I humbly offer my Intro to Brain-Like AGI Safety series° as a potentially useful starting point / resource in this context, since LeCun and I share many assumptions about what autonomous machine intelligence will look like, and hence I imagine he'd find it somewhat less difficult to relate to than most AI alignment documents. And I would be happy to chat more! :)

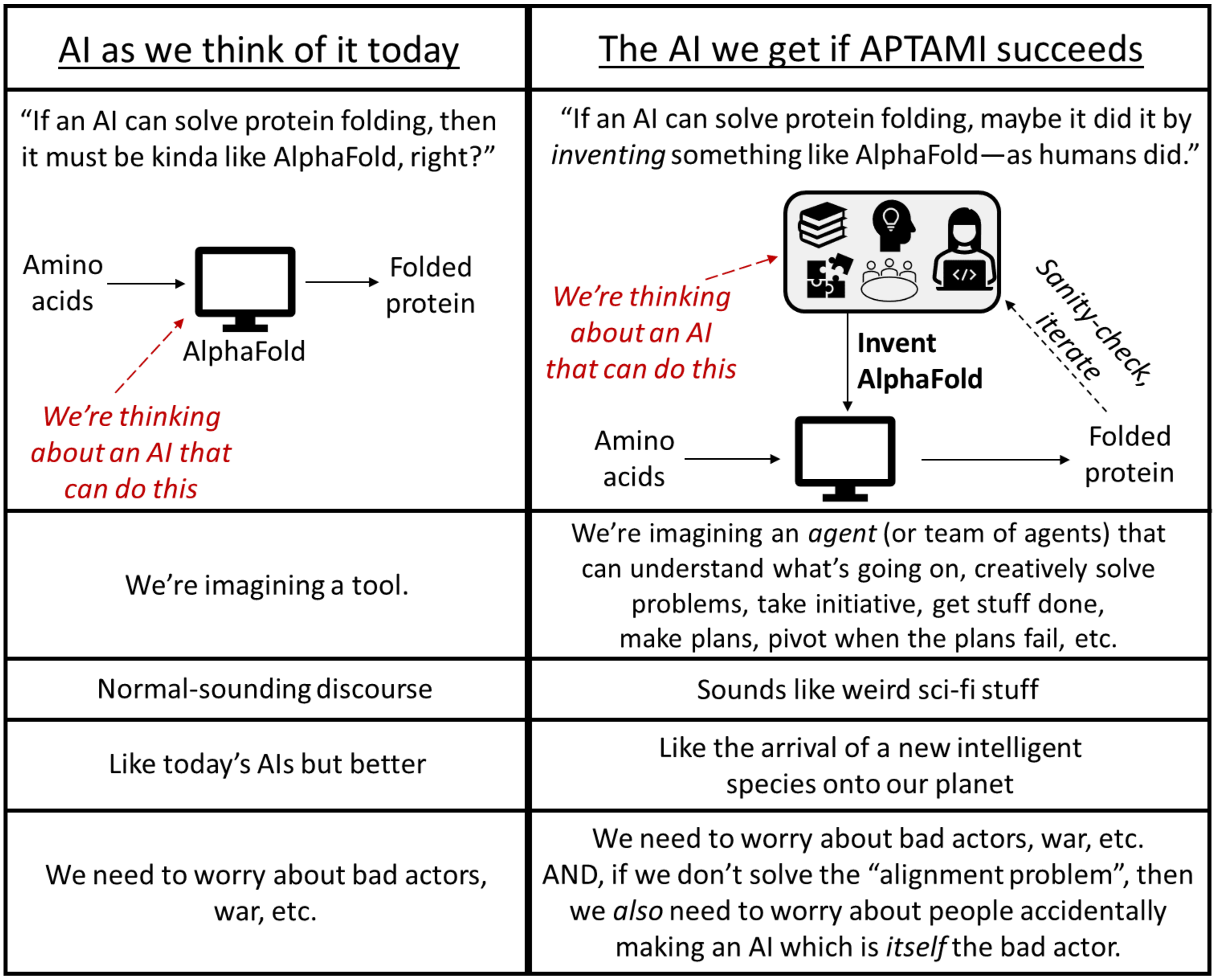
Epilogue: We need to do better than a cursory treatment of this technical problem, *even* in the context of a very-early-stage speculative vision paper

OK, if you've read this far, then maybe you're thinking something along the following lines:

> So, Yann LeCun published a self-described 'position paper' expressing a 'vision for a path towards intelligent machines'. He was explicitly intending to spur discussion and solicit feedback, even posting it on openreview.net rather than arxiv. And now this other guy has written a blog post saying that one aspect of the vision is more complicated and difficult to get right than implied by the very brief paper discussion.
>
> Umm, yeah, duh. *Everything* in the paper's proposed architecture is more complicated and difficult to get right than the corresponding very brief paper discussion. It's a vision paper, not a technical blueprint. And the paper was written to solicit feedback, and it evidently succeeded, because now I'm reading a blog post that is giving feedback. And meanwhile, while the blog post suggested that there is an open problem that needs to be solved, everyone seems to be in agreement that a solution to that problem probably exists—since after all we're talking about a path to a brain-like AI architecture, and humans have brains, and humans can be nice and cooperative sometimes.
>
> So, kudos all around. This is a good and healthy R&D process. Everything is fine.

I disagree—I think that if you were nodding along with the above paragraphs then you have lost sight of something very important.

APTAMI is an attempt to describe a path towards *powerful* AI—AI that can understand the world, get things done, figure things out, make plans, pivot when the plans fail, build tools to solve their problems, and so on—all the things that would make us think of the AIs intuitively as “a new intelligent species” rather than “an AI system as we think of it today”.



AI as we think of it today	The AI we get if APTAMI succeeds
“If an AI can solve protein folding, then it must be kinda like AlphaFold, right?” Amino acids Folded protein AlphaFold *We’re thinking about an AI that can do this*	“If an AI can solve protein folding, maybe it did it by *inventing* something like AlphaFold—as humans did.” *We’re thinking about an AI that can do this* Invent AlphaFold Sanity-check, iterate Amino acids Folded protein
We’re imagining a tool.	We’re imagining an *agent* (or team of agents) that can understand what’s going on, creatively solve problems, take initiative, get stuff done, make plans, pivot when the plans fail, etc.
Normal-sounding discourse	Sounds like weird sci-fi stuff
Like today’s AIs but better	Like the arrival of a new intelligent species onto our planet
We need to worry about bad actors, war, etc.	We need to worry about bad actors, war, etc. AND, if we don’t solve the “alignment problem”, then we *also* need to worry about people accidentally making an AI which is *itself* the bad actor.

Figure is modified from this post °—see there for further discussion

Suppose someone published a position paper describing “a vision for a path towards using bioengineering to create a new intelligent nonhuman species”. And suppose that the question of how to make this new species care about humans and/or stay under human control is relegated to a vague and cursory discussion in a few sentences, and close examination reveals that if we were to follow the advice in those few sentences then we would probably get a highly-intelligent nonhuman species *pursuing its own interests with callous disregard for human welfare*—somewhat akin to a species of high-functioning sociopaths.

If that someone says “Yeah sure, obviously there are still lots of details to work out”, then I would respond: “No no no! The question of how to design this new species such that they will be docile and/or intrinsically care about human welfare is *not* just one of many technical details to be worked out! This is the kind of problem where *we halt all other work on this research program until we have sorted this out*.”

That seems like common sense to me. If not, consider:

- For one thing, we don't actually know for sure that this technical problem is solvable at all, until we solve it. And if it's not in fact solvable, *then we should not be working on this research program at all*. If it′s not solvable, the only possible result of this research program would be "a recipe for summoning demons", so to speak. And if you're scientifically curious about what a demon-summoning recipe would look like, then please go find something else to be scientifically curious about instead.
    - Now, in the case at hand, it's a *decent* argument to say "humans are sometimes nice, therefore it's probably possible in principle to make brain-like AIs that are nice"—indeed, I often make that argument myself. But it's not a *strong* argument, let alone air-tight. For example, for all I know right now, maybe making a nice human requires a "training environment" that entails growing up with a human body, in a human community, at human speed. Doing that with AI is not really feasible in practice, for many reasons. And that's just one example problem among many.
- For another thing, this technical alignment problem could be the kind of technical problem that takes a long time to solve, even assuming we have the benefit of trial-and-error. If we make progress on every other aspect of the research program *first*, while taking a "we'll cross that bridge when we get to it" attitude on how *exactly* to code up the Intrinsic Cost module, then we could wind up in a situation where we have discovered (perhaps even open-sourced) a recipe for building self-interested AIs with callous disregard for humanity, but we have *not* yet discovered any analogous recipe for building friendly powerful AIs that might help us and fight on our side. That's a bad situation. And we can avoid that situation by doing the requisite research in the right order.
- Finally, if Yann LeCun were *merely* treating this open technical problem in a cursory way, and proposing approaches that are technically flawed upon close scrutiny, then that would at least be somewhat understandable. I myself propose technically-flawed plans all the time! A bigger issue is that LeCun, in his public statements, gives a strong impression that he is *opposed* to people working on this technical problem. If my impression here is wrong—if LeCun in fact thinks that the open technical problem described in this post is a worthwhile thing for AI researchers to be working on—then I appeal to him to directly and straightforwardly say that. It would make a huge difference.

(*Thanks Christopher King, Roman Leventov, & Justis Mills for critical comments on earlier drafts.*)

1. ^ I note that these two out-of-distribution problems correspond respectively with [aspects of] what I call "outer alignment" and "inner alignment" in this post°. "Outer alignment" (in this context) is the question "Is the Intrinsic Cost module returning high vs low intrinsic energy outputs in a way that tracks the extent to which the agent is doing things that it was intended to do?" And "inner alignment" (in this context) is the question "When the AI imagines some possible future plan, does the plan seem appealing / unappealing to the AI in a way that actually tracks its expected future Intrinsic Cost?" By the way, as discussed at that link, it's not *necessarily* the case that the best way to get good AI behavior is to *separately* solve both these out-of-distribution generalization problems; for example, if we have sufficient neural network interpretability of the trained critic, then we get to slice through both layers, bridging directly from the design intentions to the AI's motivations, without relying on the Intrinsic Energy function being perfect.
2. ^ Maybe you're thinking: "OK, we'll design the AI such that, if something seems like an out-of-the-box idea, then the AI doesn't want to do it. The AI wants to stay *in* the box!" Or in conventional ML terms, if we're worried about out-of-distribution problems, then we can just put in a penalty term that makes the AI want to stay in-distribution. I do actually think there's a kernel of a promising research direction here, but I don't know how to flesh it out into a plausible plan. In particular, the most obvious approaches along these lines would have an unintended side-effect of crippling the AI's ability to learn new things, make new connections, do R&D, etc. Further discussion in Section 14.4 here°.

Reinforcement Learning 2 AI 2 Frontpage

Mentioned in

99 Shallow review of live agendas in alignment & safety
76 Thoughts on "AI is easy to control" by Pope & Belrose
12 Yann LeCun on AGI and AI Safety

23 comments, sorted by top scoring

[−] **tailcalled** 1y < 7 > ✕ 2 ✓

Nice, I was actually just thinking that someone needed to respond to LeCun's proposal.

That said, I think you may have gotten some of the details wrong. I don't think the intrinsic cost module gets raw sensory data as input, but instead it gets input from the latent variables of the world model as well as the self-supervised perception module. This complicates some of the safety problems you suggest.

[–] **Steve Byrnes** 1y < 3 > ✕ 0 ✓

Thanks. That did actually occur to me, but I left it out because I wasn't sure and didn't want to go on an exhausting chase down every possible interpretation of the paper.

Anyway, if the input to the Prosociality Score Model is a set of latent variables rather than a set of pixels then:

- My OP claim that there are two adversarial out-of-distribution generalization problems (in the absence of some creative solution not in the paper) is still true.
- One of those two problems (OOD generalization of the Prosociality Score Model) might get less bad, although I don't see why it would go away altogether.
- …But only if the labels are correct, and the labeling problem is potentially much harder now, because the latent variables include inscrutable information about "how the AI is thinking about / conceptualizing the things that it's seeing / doing". I think. And if they do, then how are the humans supposed to label them as good or bad? Like, if the AI notices someone feeling physically good but psychologically distressed, we want to label it as low-energy when the AI is thinking about the former aspect and high-energy if the AI is thinking about the latter aspect, I imagine. And then we start getting into nasty neural net interpretability challenges.

Also, aren't the latent variables changing as we go, thanks to self-supervised learning? But the Intrinsic Cost Module is supposed to be immutable. I'm confused about how this is supposed to work.

[–] **Roman Leventov** 1y < 3 > ✕ 0 ✓

My comments on this post winded up into a whole separate post: "H-JEPA might be technically alignable in a modified form°".

[–] **Ben Amitay** 1y < 2 > ✕ 2 ✓

Didn't read the original paper yet, but from what you describe, I don't understand how the remaining technical problem is not basically the whole of the alignment problem. My understanding of what you say is that he is vague about the values we want to give the agent - and not knowing how to specify human values is kind of the point (that, and inner alignment - which I don't see addressed either).

[–] **Steve Byrnes** 1y < 2 > ✕ 1 ✓

> I don't understand how the remaining technical problem is not basically the whole of the alignment problem

Yes. I don't think the paper constitutes any progress on the alignment problem. (No surprise, since it talks about the problem for only a couple sentences.)

Hmm, maybe you're confused that the title refers to "an unsolved technical alignment problem" instead of "the unsolved technical alignment problem"? Well, I didn't mean it that way. I think that solving technical alignment entails solving a different (albeit related) technical problem for each different possible way to build / train AGI. The paper is (perhaps) a possible way to build / train AGI, and therefore it has an alignment problem. That's all I meant there.

[−] **Ben Amitay** 1y ‹ 1 › ✕ 0 ✓

Yes, I think that was it; and that I did not (and still don't) understand what about that possible AGI architecture is non-trivial and has a non-trivial implementations for alignment, even if not ones that make it easier. It seem like not only the same problems carefully hidden, but the same flavor of the same problems on plain sight.

[−] **Steve Byrnes** 1y ‹ 4 › ✕ 1 ✓

I think of my specialty as mostly "trying to solve the alignment problem for model-based RL". (LeCun's paper is an example of model-based RL.) I think that's a somewhat different activity than, say, "trying to solve the alignment problem for LLMs". Like, I read plenty of alignmentforum posts on the latter topic, and I *mostly* don't find them very relevant to my work. (There are exceptions.) E.g. the waluigi effect° is not something that seems at all relevant to my work, but it's extremely relevant to the LLM-alignment crowd. Conversely, for example, here's° a random recent post I wrote that I believe would be utterly useless to anyone focused on trying to solve the alignment problem for LLMs.

A big difference is that I feel entitled to assume that there's a data structure labeled "world-model", and there's a different data-structure labeled "value function" (a.k.a. "critic"). Maybe each of those data structures is individually a big mess of a trillion uninterpretable floating-point numbers. But it still matters that there are two data structures, and we know where each lives in memory, and we know what role each is playing, how it's updated, etc. That changes the kinds of detailed interventions that one might consider doing. [There could be more than two data structures, that's just an example.]

[−] **Ben Amitay** 1y ‹ 1 › ✕ 0 ✓

I see. I didn't fully adapt to the fact that not all alignment is about RL.

Beside the point: I think those labels on the data structures are very confusing. Both the actor and the critic are very likely to have so specialized world models (projected from the labeled world model) and planning abilities. The values of the actor need not be the same as the output of the critic. And things value-related and planning-related may easily leak into the world model if you don't actively try to prevent it. So I suspect that we should ignore the labels and focus on architecture and training methods.

[−] **Steve Byrnes** 1y ‹ 2 › ✕ 0 ✓

Sure, we can take some particular model-based RL algorithm (MuZero, APTAMI, the human brain algorithm, whatever), but instead of "the reward function" we call it "function #5829", and instead of "the value function" we call it "function #6241", etc. If you insist that I use those terms, then I would still be perfectly capable of describing step-by-step why this algorithm would try to kill us. That would be pretty annoying though. I would rather use the normal terms.

I'm not quite sure what you're talking about ("projected from the labeled world model"??), but I guess it's off-topic here unless it specifically applies to APTAMI.

FWIW the problems addressed in this post involve the model-based RL system trying to kill us *via* using its model-based RL capabilities in the way we normally expect—where the planner plans, and the critic criticizes, and the world-model models the world, etc., and the result is that the system makes and executes a plan to kill us. I consider that the obvious, central type of alignment failure mode for model-based RL, and it remains an unsolved problem.

In addition, one might ask if there are *other* alignment failure modes too. E.g. people sometimes bring up more exotic things like the "mesa-optimizer" thing where the world-model is secretly harboring a full-fledged planning agent, or whatever. As it happens, I think those more exotic failure modes can be effectively mitigated, and are also quite unlikely to happen in the first place, in the particular context of model-based RL systems. But that depends a lot on how the model-based RL system in question is supposed to work, in detail, and I'm not sure I want to get into that topic here, it's kinda off-topic. I talk about it a bit in the intro here°.

[–] **PeterMcCluskey** 1y ‹ 1 › ✕ 0 ✓

Why assume LeCun would use only supervised learning to create the IC module?

If I were trying to make this model work, I′d use mainly self-supervised learning that′s aimed at getting the module to predict what a typical human would feel. (I′d also pray for a highly multipolar scenario if I were making this module immutable when deployed.)

[–] **Steve Byrnes** 1y ‹ 2 › ✕ 0 ✓

> If I were trying to make this model work, I′d use mainly self-supervised learning that′s aimed at getting the module to predict what a typical human would feel.

I don't follow. Can you explain in more detail? "Self-supervised learning" means training a model to predict some function / subset of the input data from a different function / subset of the input data, right? What's the input data here, and what is the prediction target?

[–] **PeterMcCluskey** 1y ‹ 1 › ✕ 0 ✓

I haven′t thought this out very carefully. I′m imagining a transformer trained both to predict text, and to predict the next frame of video.

Train it on all available videos that show realistic human body language.

Then ask the transformer to rate on a numeric scale how positively or negatively a human would feel in any particular situation.

This does not seem sufficient for a safe result, but implies that LeCun is less nutty than your model of him suggests.

[–] **Steve Byrnes** 1y ‹ 2 › ✕ 0 ✓

> Then ask the transformer to rate on a numeric scale how positively or negatively a human would feel in any particular situation.

I′m still confused. Here you′re describing what you′re hoping will happen at inference time. I′m asking how it′s trained, such that that happens. If you have a next-frame video predictor, you can′t ask it how a human would feel. You can′t ask it anything at all - except ″what might be the next frame of thus-and-such video?″. Right?

I wonder if you′ve gotten thrown off by chatGPT etc. Those are NOT trained by SSL, and therefore NOT indicative of how SSL-trained models behave. They′re *pre*-trained by SSL, but then they′re fine-tuned by supervised learning, RLHF, etc. The grizzled old LLM people will tell you about the behavior of pure-SSL models, which everyone used before like a year ago. They're quite different. You cannot just ask them a question and expect them to spit out an answer. You have to prompt them in more elaborate ways.

(On a different topic, self-supervised pre-training before supervised fine-tuning is almost always better than supervised learning from random initialization, as far as I understand. Presumably if someone were following the OP protocol, which involves a supervised learning step, then they would follow all the modern best practices for supervised learning, and "start from a self-supervised-pretrained model" is part of those best practices.)

[–] **red75prime** 1y < 2 > ✕ 0 ✓

> If you have a next-frame video predictor, you can't ask it how a human would feel. You can't ask it anything at all - except "what might be the next frame of thus-and-such video?". Right?

Not exactly. You can extract embeddings from a video predictor (activations of the next-to-last layer may do, or you can use techniques, which enhance semantic information captured in the embeddings). And then use supervised learning to train a simple classifier from an embedding to human feelings on a modest number of video/feelings pairs.

[–] **Steve Byrnes** 1y < 2 > ✕ 0 ✓

I think that's what I said in the last paragraph of the comment you're responding to:

> (On a different topic, self-supervised pre-training before supervised fine-tuning is almost always better than supervised learning from random initialization, as far as I understand. Presumably if someone were following the OP protocol, which involves a supervised learning step, then they would follow all the modern best practices for supervised learning, and "start from a self-supervised-pretrained model" is part of those best practices.)

Maybe that's what PeterMcCluskey was asking about this whole time—I found his comments upthread to be pretty confusing. But anyway, *if* that's what we've been talking about all along, then yeah, sure. I don't think my OP implied that we would do supervised learning from random initialization. I just said "use supervised learning to train an ML model". I was assuming that people would follow all the best practices for supervised learning—self-supervised pretraining, data augmentation, you name it. This is all well-known stuff—this step is not where the hard unsolved technical problems are. I'm open to changing the wording if you think the current version is unclear.

[–] **[deactivated]** 5mo < 0 > ✕ 0 ✓

> Now, it doesn't immediately follow that the AI will actually want to start buying chair-straps and heroin, for a similar reason as why I personally am not trying to get heroin right now.

This seems important to me. What is the intrinsic cost in a human brain like mine or yours? Why don't humans have an alignment problem (e.g. if you radically enhanced human intelligence, you wouldn't produce a paperclip maximiser)?

Maybe the view of alignment pessimists is that the paradigmatic human brain's intrinsic cost is intractably complex. I don't know. I would like more clarity on this point.

[−] **Steve Byrnes** 5mo ‹ 2 › ✕ 0 ✓

The "similar reason as why I personally am not trying to get heroin right now" is "Example 2" here° (including the footnote), or a bit more detail in Section 9.5 here°. I don't think that involves an idiosyncratic anti-heroin intrinsic cost function.

The question "What is the intrinsic cost in a human brain" is a topic in which I have a strong personal interest. See Section 2 here° and links therein. "Why don't humans have an alignment problem" is sorta painting the target around the arrow I think? Anyway, if you radically enhanced human intelligence and let those super-humans invent every possible technology, I'm not too sure what you would get (assuming they don't blow each other to smithereens). Maybe that's OK though? Hard to say. Our distant ancestors would think that we have awfully weird lifestyles and might strenuously object to it, if they could have a say.

> Maybe the view of alignment pessimists is that the paradigmatic human brain's intrinsic cost is intractably complex.

Speaking for myself, I think the human brain's intrinsic-cost-like-thing is probably hundreds of lines of pseudocode, or maybe low thousands, certainly not millions. (And the part that's relevant for AGI is just a fraction of that.) Unfortunately, I also think nobody knows what those lines are. I would feel better if they did. That wouldn't be enough to make me "optimistic" overall, but it would certainly be a step in the right direction. (Other things can go wrong too.)

[−] **[deactivated]** 5mo ‹ 0 › ✕ 0 ✓

> ...I think the human brain's intrinsic-cost-like-thing is probably hundreds of lines of pseudocode, or maybe low thousands, certainly not millions. (And the part that's relevant for AGI is just a fraction of that.) Unfortunately, I also think nobody knows what those lines are. I would feel better if they did.

So, the human brain′s pseudo-intrinsic cost is not intractably complex, on your view, but difficult to extract.

[−] **Steve Byrnes** 5mo ‹ 2 › ✕ 0 ✓

I would say "the human brain's intrinsic-cost-like-thing is difficult to figure out". I'm not sure what you mean by "... difficult to extract". Extract from what?

[−] **[deactivated]** 5mo ‹ 0 › ✕ 0 ✓

Extract from the brain into, say, weights in an artificial neural network, lines of code, a natural language ″constitution″, or something of that nature.

[−] **Steve Byrnes** 5mo ‹ 2 › ✕ 0 ✓

"Extract from the brain" how? A human brain has like 100 billion neurons and 100 trillion synapses, and they're generally very difficult to measure, right? (I do think certain neuroscience experiments would be helpful°.) Or do you mean something else?

[−] **[deactivated]** 5mo ‹ 0 › ✕ 0 ✓

I meant ″extract″ more figuratively than literally. For example, GPT-4 seems to have acquired some ability to do moral reasoning in accordance with human values. This is one way to (very indirectly) ″extract″ information from the human brain.

[−] **Steve Byrnes** 5mo ‹ 2 › ✕ 0 ✓

GPT-4 is different from APTAMI. I’m not aware of any method that starts with movies of humans, or human-created internet text, or whatever, and then does some kind of ML, and winds up with a plausible human brain intrinsic cost function. If you have an idea for how that could work, then I’m skeptical, but you should tell me anyway. :)

Moderation Log