Ilya Sutskever's thoughts on AI safety in a world with superintelligent AI systems

Mishka (Michael Bukatin)

Dataflow Matrix Machines project

slides: `https://github.com/anhinga/2024-notes/tree/main/Ilya-SSI/`

AI lightning talk at Boston Astral Codex Ten Meetup

June 28, 2024

Context

Ilya left OpenAI on May 14, saying “I am excited for what comes next — a project that is very personally meaningful to me about which I will share details in due time.”

On June 19 he launched https://ssi.inc/

> **Safe Superintelligence Inc.**
>
> **Superintelligence is within reach.**
>
> Building safe superintelligence (SSI) is the most important technical problem of our time.
>
> We have started the world’s first straight-shot SSI lab, with one goal and one product: a safe superintelligence.

“The world’s first straight-shot SSI lab, with one goal and one product: a safe superintelligence.”

SSI is our mission, our name, and our entire product roadmap, because it is our sole focus. Our team, investors, and business model are all aligned to achieve SSI.

We approach safety and capabilities in tandem, as technical problems to be solved through revolutionary engineering and scientific breakthroughs. We plan to advance capabilities as fast as possible while making sure our safety always remains ahead.

This way, we can scale in peace.

Our singular focus means no distraction by management overhead or product cycles, and our business model means safety, security, and progress are all insulated from short-term commercial pressures.

Context 2

Ilya is the only person who played a key role in all three modern **"phase transitions in AI capability"**:

- AlexNet (2012)
- GPT-3 (2020)
- GPT-4 (2023 public release)

Many people think he can do it again.

He was sharing his thoughts on **existential safety for a world with superintelligent AI systems** quite a bit last year, and I made some notes of it:

https://www.lesswrong.com/posts/TpKktHS8GszgmMw4B/ilya-sutskever-s-thoughts-on-ai-safety-july-2023-a

Before presenting his thoughts, I'll discuss the situation.

AI self-improvement, timelines, and speed of changes

How to think about takeoff and transition to superintelligence?

The main factor: how much will AI systems contribute to making AI systems better/stronger/more capable.

Tom Davidson in his "Takeoff speeds presentation at Anthropic"

https://www.lesswrong.com/posts/Nsmabb9fhpLuLdtLE/takeoff-speeds-presentation-at-anthropic

redefines AGI by narrowing it down as

"AGI" (=AI that could fully automate AI R&D)

Rapid takeoff can start before that line is achieved, as AI can start accelerating AI R&D rapidly without full automation.

I am seeing *a lot of recursive self-improvement experiments*, but right now they all saturate rapidly, and acceleration of AI R&D by AI is still relatively moderate.

Superintelligence and self-modification

For our purposes: **superintelligent systems are systems which are much better than humans at AI R&D**.

Superintelligent system will self-modify rapidly
(experiment with their modified copies a lot).

Ecosystems containing superintelligent systems
will self-modify rapidly, producing

- novel AIs
- novel kinds of collectives
- novel collective dynamics
- and so on

A world with superintelligent AI systems is a moving target.

If we want it to have any particular properties we want, this is not easy. Strong potential for unpredictable blow-up.

Two classes of approaches to AI existential safety

Adversarial vs non-adversarial (collaborative)

“Traditional alignment” is very adversarial.

It wants to be technically able to impose arbitrary values and goals onto systems smarter than humans.

It then wants to use that technical ability to impose values and goals based on some kind of “extrapolated human values”.

Of course, ethics, feasibility, and wisdom of trying to impose our values and goals on smarter entities are highly questionable.

It is quite unlikely that arbitrary values and goals would survive radical self-modifications of AIs and of the overall ecosystem.

What could be the mechanisms preserving any particular values and goals through radical self-modifications of AIs and of the world?

non-adversarial (collaborative) approaches (how I see it)

Any feasible approach must take interests and rights of **everyone** into account

This notion of **“everyone”** definitely includes smart AI systems.

It’s not “us vs. them”

We would like to **collaborate with AIs** to figure out how to take interests of everyone into account, and *how to make it so that the situation does not deteriorate during radical changes and self-modifications.*

We would ideally like to formulate the desired “safety properties” in a *non-anthropocentric fashion*, so that AIs would have a lot of incentives to preserve those properties through self-modification.

Any anthropocentric properties we might particularly desire should be *corollaries* of general non-anthropocentric universal properties.

non-adversarial (collaborative) approaches 2

Any anthropocentric properties we might particularly desire should be *corollaries* of general non-anthropocentric universal properties.

We want to get what we need without trying to make humans central.

Humans are just individuals, and **individuals (human or not) have rights and interests which need to be protected**.

One possible way this could happen and be maintained:

If there are a lot of **AI individuals** on different levels of capabilities with a lot of combined power, they are likely to be interested in having a system *protecting individuals on all levels of capabilities*, because they don't know their future trajectory and each individual wants to be sure that its interests and rights are protected in various future scenarios.

Ilya's approach is a non-adversarial collaborative approach

OpenAI alignment efforts were always done in a traditional adversarial way.

When Superalignment was announced in July 2023, it was also formulated in an adversarial fashion:

https://openai.com/index/introducing-superalignment/

> We need scientific and technical breakthroughs to steer and control AI systems much smarter than us.

But when Ilya gave an interview a few days later, he was steering away from this and towards a non-adversarial collaborative approach:

https://www.lesswrong.com/posts/TpKktHS8GszgmMw4B/ilya-sutskever-s-thoughts-on-ai-safety-july-2023-a

Ilya: July 2023 interview

Radically redefining alignment:

> The concern number one has been expressed a lot and this is the scientific problem of alignment. You might want to think of it from the as an analog to nuclear safety. You know you build a nuclear reactor, you want to get the energy, you need to make sure that it won't melt down even if there's an earthquake and even if someone tries to I don't know smash a truck into it. So this is the superintelligent safety and it must be addressed in order to contain the vast power of the superintelligence. It's called the alignment problem.

This is not what people call alignment. This is just a constraint against an uncontrolled blow-up.

Ilya: July 2023 interview 2

Collaborate with AI to decide how to steer the reality:

> The Second Challenge to overcome is that of course we are people, we are humans, "humans of interests", and if you have superintelligences controlled by people, who knows what's going to happen... I do hope that at this point we will have the superintelligence itself try to help us solve the challenge in the world that it creates. This is not... no longer an unreasonable thing to say. **Like if you imagine a superintelligence that indeed sees things more deeply than we do, much more deeply. To understand reality better than us. We could use it to help us solve the challenges that it creates.**

Ilya: July 2023 interview 3

Pondering partial merge of people and AI:

> Then there is the third challenge which is the challenge maybe of natural selection. You know what the Buddhists say: the change is the only constant. So even if you do have your superintelligences in the world and they are all... We've managed to solve alignment, we've managed to solve... **no one wants to use them in very destructive ways**, we managed to create a life of unbelievable abundance, which really like not just not just material abundance, but Health, longevity, like all the things we don't even try dreaming about because there's obviously impossible, if you've got to this point then there is the third challenge of natural selection. Things change, you know... You know that natural selection applies to ideas, to organizations, and that's a challenge as well.

Ilya: July 2023 interview 3 continued

Pondering partial merge of people and AI:

> Maybe the Neuralink solution of people becoming part AI will be one way we will choose to address this. I don't know. But I would say that this kind of describes my concern. And specifically just as the concerns are big, if you manage, it is so worthwhile to overcome them, because then we could create truly unbelievable lives for ourselves that are completely even unimaginable. So it is like a challenge that's really really worth overcoming.

Speaking of merging humans and AIs, I'd prefer people to focus more on the intermediate solutions before jumping to Neuralink-grade ones. In particular, **high-end augmented reality** and **high-end non-invasive brain-computer interfaces** can go a long way and are much easier to accelerate rapidly, so I wish people would not gloss over those intermediate solutions, but would talk about them more.

Ilya: September 2023

https://time.com/collection/time100-ai/6309011/ilya-sutskever/

> There is a precedent, according to Ilya Sutskever, for a less intelligent being ensuring that radically smarter and more powerful ones act in their interests. That precedent is the human baby. “We know that it’s possible,” says Sutskever, chief scientist at OpenAI. “Parents care very deeply about the well-being of their children. It can be done. How does this imprinting work?”
>
> “The upshot is, eventually AI systems will become very, very, very capable and powerful,” he says. “We will not be able to understand them. They’ll be much smarter than us. By that time it is absolutely critical that the imprinting is very strong, so they feel toward us the way we feel toward our babies.”