MARCH 2019

Artificial Intelligence Safety

CENTER FOR APPLIED RATIONALITY

Prepared by
CHRIS SHEPARD,
OUTREACH COORDINATOR

For the purpose of
INTERMEDIATE WRITING
ENGLISH 2010

Artificial General Intelligence Is Coming

Artificial general intelligence (AGI) refers to "strong" AI that is capable of performing any intellectual task that a human is capable of. All of the examples of artificial intelligence in the world today are limited, "weak" AIs, which perform well in narrowly defined domains.

The predictive text software on your phone, or the autonomous driving software found in self-driving cars, or the transaction software underpinning Amazon's webstore are all examples of weak AI in use today.

We do not yet know how to create artificial general intelligence, but efforts are in progress and there are strong reasons to think that it can be achieved (Harris, 2016).

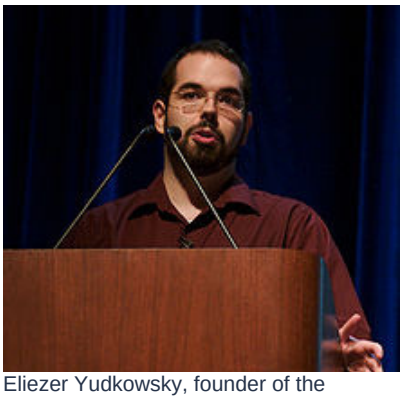
AGI Is Catastrophic By Default

Assuming the conservative position that artificial general intelligence will operate at merely human levels in arbitrary domains but at the speed of modern computing, an AGI would achieve as much as *20,000 years'* worth of progress on any intellectual task it set out to perform *in a week*.

An AGI with such capabilities will be able to effectively accomplish whatever goal it sets out to do. If its goals are incompatible with our own, humanity is unlikely to be able to prevent the negative outcomes of these incompatibilities (Soares, 2015).

The possible ways to build something with goals that are incompatible with our own *vastly* outnumber the ones that align nicely.

There are Strong Incentives to Build AGI



Eliezer Yudkowsky, founder of the Machine Intelligence Research Institute

Whichever company or government first creates an artificial general intelligence will have an insurmountable advantage over its competitors and adversaries (Harris, 2016).

A financial institution with an AGI would dominate the market irrevokably. A medical research organization with an AGI would produce cures and effective treatments at an unprecedented rate. A military with an AGI would be able to completely negate any possible threat from adversarial nations.

Such an overwhelming advantage leads to a zero-sum race, where the first to win is the only winner. Any call to slow such research competes directly against these incentives.

We Do Not Know How to Build Safe AGI

Some expect that building safe AGI will require interfacing with human brains directly (Harris, 2016).

The field of research for understanding how a safe AGI would need to be structured is in its infancy. The scope of the problem is both vast and poorly understood. We do not know how to create something capable of understanding our intentions. If we did, that does not mean that the AGI would act in ways we could anticipate. Even if we could easily detect when an AGI was working under some misunderstanding, we do not know how to guarantee that we could fix the problem, or even ensure that we could turn the AGI off again, once it is activated.

These are known, open problems in the field. They are still in the process of being rigorously defined let alone being resolved.

The funding gap between general AI research and AI safety research in particular is more than 200:1 ("Funding of AI Research", 2017; Farquhar, 2017).

Impact of AGI on CFAR

"Addressing existential risks such as AI safety is one of the greatest opportunities for clearer thinking to make an important difference on humanity's future."

Due to the catastrophic impact that unsafe AGI would have on humanity in general, we believe that we have an obligation to have a positive impact in the field of artificial intelligence safety research.

The Center for Applied Rationality has long held that our efforts to teach effective rationality skills are a means, not only to the greater end of improving humanity, but specifically to reducing existential risks to humanity's future. Indeed, our mission statement highlights this importance explicitly: "Many of our alumni and staff have come to believe that addressing existential risks such as AI safety is one of the greatest opportunities for clearer thinking to make an important difference on humanity's future" (Rationality, 2019, para. 15).

The risks posed by unsafe artificial general intelligence have implications for the direction of our own research and the future development of our workshop curriculum.

How can we best leverage our domain of expertise to improve the chances of success for AGI safety research?

CFAR alumnus Kenzi Amodei addressing workshop attendees.

What CFAR Can Contribute

The Center for Applied Rationality has a role to play in artificial intelligence safety research.

CFAR cofounder Julia Galef addressing workshop attendees.

The Center for Applied Rationality already interacts with the broader community of artificial intelligence safety research to a large degree. The founding members of our organization shared office space and knowledge with the founding members of the Machine Intelligence Research Institute, one of the leading research organizations in the field of AI safety.

In order to better leverage this familiarity to have an impact on the problem of AI safety, our organization is primed to orient along one of several possible paths forward:

1. Focus on outreach and raising awareness of the problem of AI safety among the general population.

2. Focus on collaborating and educating those currently working in the field of AI safety research today.

3. Focus on educating those members of the public who are best situated to have an impact in AI safety research but who are not already involved.

Each of these options would individually require the full efforts of the Center for Applied Rationality as an organization. We do not have the available means to effectively execute on each of these simultaneously. It is important that our choice is as effective as possible.

Outreach and Raising Awareness

Increasing the number of people who know about the risks of AGI may help to improve research funding and interest in AI safety.

CFAR workshop attendees during presentation.

The Center for Applied Rationality already has a broadly available public platform. Beyond our workshops, we interface and organize community events around the world. We present at conferences, and lecture at major universities. We also manage a vibrant alumni community.

These avenues allow us to interact with the general public in ways that organizations dedicated to AI safety research are often ill-placed to do.

By leveraging our existing public-facing platform, we could lead a campaign to increase public awareness of the dangers of unsafe artificial intelligence research.

Through such a campaign, we anticipate increased funding to organizations aligned with AI safety research. We also expect a more engaged public to act as a counterbalance to the zero-sum incentives at play in artificial general intelligence research, today.

CFAR cofounder Julia Galef addressing workshop attendees.

Working directly with AI safety researchers may ensure that they have the skills necessary to tackle difficult problems.

"Designing an 'art of rationality' that can support work on AI safety is different from designing an 'art of rationality' for some other cause."

We have interacted with AI safety researchers since the earliest days of our organization. We initially operated out of the same office space as the Machine Intelligence Research Institute, and some of our earliest curriculum advisors are active AI safety researchers.

Our work so far has been broadly focused on improving the general rationality of everyone. It seems likely, however, that we would be more effective in having a positive impact by narrowing our focus to help those who are actively performing AI safety research.

Skills that work well in general may not be particularly applicable or relevant in this narrow domain. Techniques that boost the productive efforts of AI safety research may not be generalizable to other areas of life. As Anna Salamon has stated (2016b), "Designing an 'art of rationality' that can support work in AI safety is different from designing an 'art of rationality' for some other cause" (para. 5).

By pivoting our existing workshop efforts to focus specifically on techniques and skills that synergize well with AI safety research, we could act as a force multiplier for these efforts.

Refocus Intended Audience of Workshops

CFAR workshop attendees discussing course material.

Focusing the workshops towards those interested in AI, and those who aim to direct their careers in ways that can do the most good, would increase the amount of total work done.

"These folks, we suspect, are the ones who can give humanity the most boost in its survival-odds per dollar."

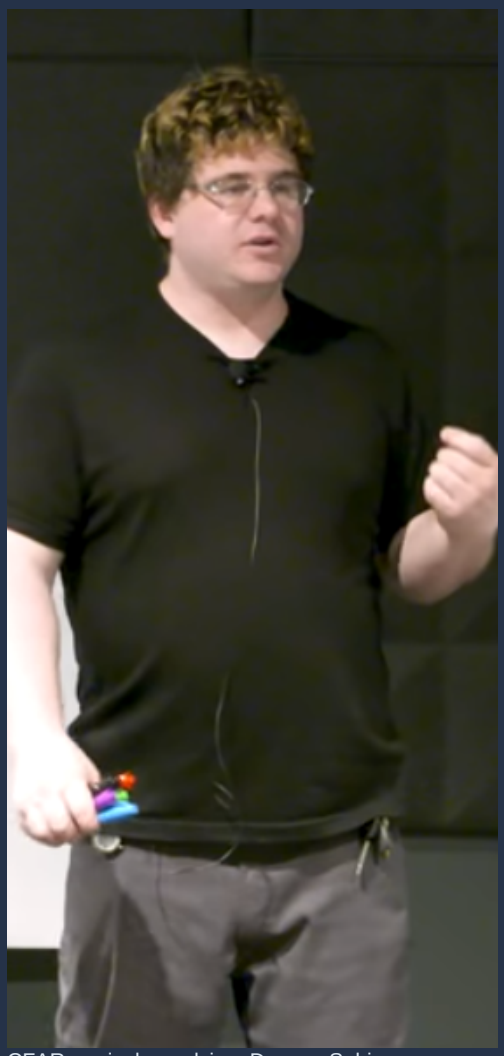
Our workshops are designed to improve the rationality and decision-making capabilities of the participants. We iteratively improve upon what works, and discard what doesn't, refining the results into effective techniques that help people achieve their goals.

The curriculum as it stands is the result of what we have found to work. Our audience has been anyone who has expressed interest in rationality, self-improvement, or being effective in achieving their objectives.

We could narrow the targeting of our audience, and thereby improve the focus of our curriculum. Rather than broadly targeting anyone with an expressed interest, or exclusively targeting active researchers in the field of AI safety, we could select a middle ground. Individuals who are interested in artificial intelligence, effective altruism, and those who aim to direct their careers in ways that can do the most good. Anna Salamon points out (2016a) "These folks, we suspect, are the ones who can give humanity the most boost in its survival-odds per dollar" (para. 10).

By aiming for this broader group we could not only act as a force multiplier to AI safety research itself but we could act as a jumping-off point for others who are able to better position themselves as force-multipliers as well.

CFAR and Artificial Intelligence Safety



CFAR curriculum advisor Duncan Sabien presenting during Effective Altruism Global 2018.

Effective, clear thinking is important when working on hard, complicated problems like AI.

The Center for Applied Rationality exists to improve peoples' abilities to think clearly about hard problems.

The problem of building an artificial general intelligence that is safe, into the distant future, is *exceptionally* hard.

Artificial general intelligence is likely to be created, sooner rather than later, regardless of the state of AI safety research.

Failing to solve this problem poses an existential risk to humanity's future.

CFAR has the capacity to positively impact AI safety research, through its workshops and interfacing with the broader public.

It is important that this capacity is leveraged effectively, in order to ensure maximum positive impact.

References

Farquhar, S. (2017). Changes in funding in the ai safety field. *Centre for Effective Altruism.* Retrieved from https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field

Funding of AI research. (2017). *AI Impacts.* Retrieved from https://aiimpacts.org/funding-of-ai-research

Harris, S. (2016). *Can we build ai without losing control over it?* TEDSummit June 2016. Retrieved from https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it

Rationality, C. F. A. (2019). Mission. *Center For Applied Rationality*. Retrieved from http://www.rationality.org/about/mission.

Salamon, A. (2016a). Cfar's new focus, and ai safety. *LessWrong*. Retrieved from https://www.lesswrong.com/posts/3zYXD8RyB6fv2czFz/cfar-s-new-focus-and-ai-safety.

Salamon, A. (2016b). Further discussion of cfar's focus on ai safety, and the good things folks wanted from "cause neutrality". *LessWrong*. Retrieved from https://www.lesswrong.com/posts/mPap4eYwGEcXBzDiH/further-discussion-of-cfar-s-focus-on-ai-safety-and-the-good.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Aaai workshops*. Autstin, Texas: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Retrieved from intelligence.org/files/Corrigibility.pdf.