Report FRI-16-1

Suffering-focused AI safety: Why "fail-safe" measures might be our top intervention

Lukas Gloor
Foundational Research Institute
lukas.gloor@foundational-research.org

June 2016

Abstract

AI-safety efforts focused on suffering reduction should place particular emphasis on avoiding risks of astronomical disvalue. Among the cases where uncontrolled AI destroys humanity, outcomes might still differ enormously in the amounts of suffering produced. Rather than concentrating all our efforts on a specific future we would like to bring about, we should identify futures we least want to bring about and work on ways to steer AI trajectories around these. In particular, a "fail-safe"[1] approach to AI safety is especially promising because avoiding very bad outcomes might be much easier than making sure we get everything right. This is also a neglected cause despite there being a broad consensus among different moral views that avoiding the creation of vast amounts of suffering in our future is an ethical priority.

Contents

[1] "Fail-safe" in the sense that if control fails, at least the AI causes less suffering than would have been the case without fail-safe measures.

1 Introduction

Classical AI safety looks very difficult. In order to ensure a truly "utopian" outcome, the following things need to happen:

1. The AI that gains a decisive strategic advantage over the competition needs to be built by the right group of people.
2. These people would need to have figured out how to program favorable values into an AI (or program the AI to learn these values).
3. They would also need to have come up with a satisfying description of what these values are, either directly, or indirectly with the help of a suitable extrapolation procedure (e.g. Muehlhauser & Helm, 2012).
4. Finally, the creators of the first superintelligence might need to get decision theory right (see Soares & Fallenstein, 2015) and safeguard the AI from a lot of hard-to-anticipate failure modes.

Succeeding with all of this together is the only safe way of bringing about a flourishing utopia mostly free of suffering. (There might be other paths to utopia, provided enough luck.) But what about outcomes where something goes wrong? It is important to consider that things can go wrong to very *different* degrees. For value systems that place primary importance on the prevention of suffering, this aspect is crucial: the best way to avoid bad-case scenarios specifically may not be to try and get *everything* right. Instead, it makes sense to focus on the worst outcomes (in terms of the suffering they would contain) and on tractable methods to avert them. As others are trying to shoot for a best-case outcome (and hopefully they will succeed!), it is important that some people also take care of addressing the biggest risks. This perspective to AI safety is especially promising both because it is currently neglected and because it is easier to avoid a subset of outcomes rather than to shoot for one highly specific outcome. Finally, it is something that people with many different value systems could get behind.

2 How AI outcomes might contain suffering

From a suffering-focused perspective, the main reason to be concerned about the risks from artificial intelligence is not the possibility of human extinction or the corresponding failure to build a flourishing, intergalactic civilization. Rather, it is the thought of misaligned or "ill-aligned" AI as a powerful but morally indifferent optimization process which, in the pursuit of its goals, may transform galactic resources into (among other things) suffering. Either like the suffering we see on Earth today, or by bringing about optimized structures that possibly also contain novel forms of suffering. The things a superintelligent AI would build to pursue its goals might include a fleet of "worker bots", factories, supercomputers, space colonization machinery, etc. It is possible that all of these end up being built in a way that does not permit suffering. However, given that evolution has produced plenty of sentient minds, this weakly suggests that some of the easiest ways to implement mind architectures do come with the capacity for suffering. In the absence of an explicit anti-suffering preference, even the slightest benefit to the AI's objectives would lead to the instantiation of suffering minds. What makes this especially worrying is that the stakes involved will be huge: Space colonization is an attractive subgoal for almost any powerful optimization process, as it leads to control over the largest amount of resources (Omohundro, 2008). Even if only a small portion of these resources were used for purposes that include suffering, the resulting disvalue would sadly be astronomical.

To find the best interventions to prevent suffering, it is crucial to study how AI outcomes

differ in their expected amount of suffering and whether there are ways to favorably affect the likely outcomes. This paper aims to provide an overview on "bad-case scenarios" in the AI context and some general suggestions for how they could be avoided. Future work will include checking fundamental assumptions and looking at the most promising proposals in more detail.

2.1 Type I: Controlled AI gone bad

In these scenarios, the team that builds the first superintelligence knew what it was doing and managed to successfully shape the resulting AI's goals exactly the way they wanted it to go. Unfortunately, the goals are not what we would wish them to be and thus lead to bad or very bad consequences. Things that could go wrong include:

a) Anthropocentrism: perhaps the resulting AI would not care about the suffering of "weird" and/or voiceless nonhuman minds such as suffering subroutines or animal minds in ancestor simulations.
b) Retributivism: perhaps the AI gets built by people who want to punish members of an outgroup (e.g. religious fundamentalists punishing sinners).
c) Uncooperative: perhaps the AI's goal is something like classical utilitarianism (or any other "monotone" maximizing function) with no additional regards for cooperation with other value systems. Even though such an AI would all else equal prefer to not create suffering, it seems possible that the anti-suffering concern would in practice be overridden by opportunity costs: if happiness simulations contain much more utility for the AI than the disutility produced by the accidental suffering created in the buildup, then such an AI would behave similarly "recklessly" as e.g. a paperclip-maximizing AI.[2] Another concern lies in how strongly the AI would value robustness to deterioration.
d) Libertarianism regarding computations: perhaps the creators of the first superintelligence instruct the AI to give every human alive at the time of the singularity control of a planet or galaxy, with no additional rules to govern what goes on within those territories. Some of these human rulers are curious and amused by seeing others hurt, or worse, might be psychopaths.

2.2 Type II: "Near misses"

In this scenario, the first superintelligence is built by people who were aware of the risks, and who tried to get things right. Unfortunately, mistakes happened and the resulting outcome with "nearly-controlled AI" is bad or very bad.

a) Miserable creatures: perhaps the AI's goal function includes terms that attempt to specify sentient or human-like beings and conditions that are meant to be good for these beings. However, because of programming mistakes, unanticipated loopholes or side-effects, the conditions specified actually turn out to be bad for these beings. Worse still, the AI has a maximizing function and wants to fill as many regions of the universe as possible with these poor creatures.
b) Black swans: perhaps the AI cares about sentient or human-like minds in "proper" ways, but has bad priors, ontology, decision theory, or other fundamental con-

[2] An AI that values happiness simulations but takes cooperation with suffering reducers into account would in expectation still create some amount of suffering, but depending on just how costly it is to use more "suffering-proof" algorithms in the colonization process, or to do fewer or less fine-grained ancestor simulations, it might be possible to cut down on most of the instrumentally produced suffering at a cost that isn't very high (e.g. not higher than 15% of total happiness simulations produced).

stituents that would make it act in unfortunate and unpredictable ways.

2.3 Type III: Uncontrolled AI

Uncontrolled AI is stipulated to have goals that were never intended by its creators, or at least were not intended to take over control in the form of a singleton (Bostrom, 2006). Unless the AI in question has a neuromorphic design, it is unlikely that its goals would (directly) have something to do with sentient beings. However, even non-neuromorphic uncontrolled AI might instrumentally instantiate suffering minds in the process of achieving its goal, as a side-effect of the computations it performs. The following presents a list of ways how uncontrolled AI might create vast amounts of suffering:

- Suffering subroutines: For tasks like inventing and developing advanced technologies, as well as for coordinating its expansion into space for resource accumulation, the AI would have to rely on a fleet of "robot workers" or "robot scientists" of various kinds (which, of course, might work and function very unlike human workers or scientists). It is possible that these processes would be capable of suffering (Tomasik, 2014), although it is unclear whether would perform at their best if they (occasionally) suffer. What is clear is that if they do suffer under useful circumstances, an AI that is not explicitly aligned with compassionate goals would not have any qualms about instantiating astronomical numbers of them.
- Ancestor simulations (Bostrom, 2003; Bostrom, 2014, Ch.8): in order to gather information relevant to its goals, e.g. for improving its understanding of human psychology or sociology (Bostrom, 2014, pp. 125-26) or for studying the density of aliens in the universe and what their likely values are, a superintelligence might simulate many runs of Darwinian evolution in planet-sized supercomputers (Sandberg, 1999). It is possible that these simulations would be fine-grained enough to contain sentient minds.
- Warfare: it may be that the density of life in the universe is high enough for colonizing AIs to eventually encounter one another. If so, uncontrolled AI might end up clashing with other superintelligences, including AIs that build expanding civilizations of happy sentient beings. Perhaps they could agree to a compromise, but such an encounter could also lead to fighting over resources and warfare at the points of contact.

It is interesting to note that none of the aforementioned scenarios involve large quantities of suffering in the types of structures that the AI *directly* optimizes for. It seems likely that what makes uncontrolled AI bad is what it builds for instrumental reasons, including colonization machinery, science simulations, and other strategic computations. This suggests that opportunity costs are relevant: If an uncontrolled AI's goal function does not have diminishing returns on resources (like the infamous "paperclip maximizer"), then the room for instrumentally important computations might be small. By contrast, an AI with goals that are easy to fulfill[3] – e.g. a "paperclip protector" that only cares about protecting a single paperclip, or an AI that only cares about its own reward signal – would have much greater room pursuing instrumentally valuable computations. This line of thought suggests that outcomes from uncontrolled AI should be steered away from "paperclip-protector types."[4]

[3]Thanks to Carl Shulman for bringing this idea to our attention.

[4] However, it may be that paperclip protectors are easier to compromise with, and thus more likely to also pursue maximization of sorts in exchange for benefits from "paperclip-maximizer types."

3 Suffering-focused AI safety: Some proposals

Bad-case scenarios can be avoided in several ways. Using the typology from above, we can distinguish three classes of interventions:

- **Influencing outcomes with controlled AI:** influence the values implemented in outcomes with controlled AI either through value spreading or by supporting the AI safety project with the best values

The following can be labelled "fail-safe" measures:

- **Targeted AI safety:**
  a) Differential progress: push classical AI safety differentially in ways that best insulates it from the worst "near misses"
  b) AI safety where it is most needed: identify the paradigm in general AI that would create the worst outcomes in the event of a control failure, and work towards identifying (general) control or shut-down mechanisms in those areas
- **Graceful fails**: research ways to make AI fail in a "benign" way *conditional on it failing* ("It might be the end of the world, but it could have been (much) worse")

The following sections present some preliminary ideas for the type of interventions that could be done in those areas.

3.1 Influencing outcomes with controlled AI

This class of interventions is primarily targeted at "controlled AI gone bad" outcomes. It is something FRI and others in the EA community have already thought about a great deal.

- Improve values: raise awareness of anti-speciesism, anti-substratism, concern for weird minds or suffering-focused ethics in general.
- Cooperation: promote the idea that the gains from cooperation make it important for different value-systems to work together.
- Cooperative AI-safety efforts: by funding the AI-safety project with the best values or the best approach to cooperation, we make it more likely that controlled AI won't create vast amounts of suffering.

3.2 Differential progress in AI safety

This intervention is targeted at the worst outcomes in the "near misses" category. The goal is to figure out solutions to the worst problems first, such that any additional progress in AI safety is unlikely to lead to "near-misses," which may be worse than outcomes with uncontrolled AI in terms of suffering produced. This intervention seems comparatively more neglected than interventions in the category above.

- Secure value implementation: Look into the pros and cons of various approaches to specifying an AI's values and work on the one least likely to lead to really bad outcomes. Questions to study may include:
  - What's more likely to go wrong – something like indirect normativity, or an explicit specification of a goal function?
  - What are other promising approaches to AI safety, and how bad would the outcome be if they fail according to their weak points?
  - Would corrigibility reduce risks of astronomical suffering?
- Foundational research: Which unanticipated failure modes could there be for controlled AI? Once they are identified, how can we prevent them?

3.3 AI safety where it is most needed

This intervention is targeted at preventing the worst outcomes involving uncontrolled AI. Not all uncontrolled AIs are expected to behave equally; there are different paradigms in AI research, i.e. different basic architectures for attempting to build smarter-than-human intelligence. Examples include reinforcement learning (Sutton & Barto, 1998), proof-based seed AI, and neuromorphic AIs (Hasler & Marr, 2013). Classical AI safety is necessarily concerned with the tradeoff between success probability and "controllability;" if a particular approach to building AI is likely to lead to superintelligence, but in a way where it is hard to control the values of the resulting agent, then this poses a likely dead end. By contrast, a suffering-focused "fail-safe" approach does not have to worry about controllability to the same extent, as long as there's *enough* controllability to predictably avoid outcomes with the most suffering. This suggests that, conditional on uncontrolled AI being worse than (nearly-)controlled AI in expectation, suffering reducers can be less constrained by issues with controllability, and can thus focus more on applying AI safety to the paradigm/architecture with the highest probability of bringing about superintelligence.

In addition, we could try to identify differences among the *types* of uncontrolled AI each AI-paradigm would produce, in order to then focus on AI safety in the domain where failures of control result in the worst outcomes. Suppose for instance that one way of building AI appears to show promising results and rapid progress, yet control seems hard and likely failure modes are particularly worrying. In this case, we should consider putting efforts into AI safety applied to that particular domain. Specific examples of this more general idea include:

- AI safety for machine learners: With DeepMind's recent success, it seems not unlikely that the first superintelligent AI may be all reinforcement learning as the top control structure (as opposed to GOFAI, theorem proving or logic-based AI). In contrast to AI architectures where goals refer to the state of the world (Hibbard, 2011) rather than to internal or observational signals, pure reinforcement learners may be at a higher risk of wireheading, i.e. of hacking their input signals to continuously attain the maximum reward. If the wireheading AI is smart enough to plan into the future, it might patiently avoid doing so and develop plans for takeover first: As Bostrom notes in *Superintelligence* (Bostrom, 2014, 122–23), smart wireheaders would still want to colonize other galaxies for option value, protection, research, etc. However, unlike paperclip maximizers, wireheaders would more generally function like "paperclip protectors" in that they would have almost no opportunity costs (Ring & Orseau, 2011). If the observation that lower opportunity costs make outcomes with uncontrolled AI worse is accurate (see the discussion under "Type III: Uncontrolled AI), this would make AI safety applied to reinforcement learning architectures particularly promising. Specific interventions could include work on Interruptibility (Armstrong & Orseau, 2016), but also methods of value-loading (examples here and here).
- AI safety for "em-first" scenarios: Conditional on whole brain emulation (Sandberg & Bostrom, 2008; Hanson, 2016) becoming technologically feasible before the advent of de-novo AI, de-novo AI as eventually developed will more likely exhibit a neuromorphic design.[5] This would make it more likely that the values are human-like, which also comes with the possi-

[5] If only because neuroscience research would be very advanced in such a world.

bility of "bad values" or "near misses." AI safety applied to "em-first" scenarios could thus be a promising intervention (although with early arrival timelines being longer, it may not be favored by haste considerations).

Caveat: Pursuing any of the interventions above would require that we make sure to not speed up progress in AI development – both in general and in "unsafe" or "hard-to-control" approaches specific to those interventions.

3.4 Graceful fails

This class of interventions is targeted at preventing the worst outcomes involving "nearly-controlled" or uncontrolled AI. Instead of trying to make fails less likely, the idea is to come up with safety nets or mechanisms such that, if control fails, the outcome will be as good as it gets under the circumstances. Graceful fails are easiest to bring about in the domain of near misses, as the researchers working on AI safety are aware of the risks and thus open to the idea of safety nets.

- Backup goal functions: Research ways to implement multi-layered goal functions, with a "backup goal" that kicks in if the implementation of the top layer does not fulfill certain safety criteria. The backup would be a simpler, less ambitious goal that is less likely to result in bad outcomes. Difficulties would lie in selecting the safety criteria in ways that people with different values could all agree on, and in making sure that the backup goal gets triggered under the correct circumstances.

In the context of AI research that is not guided by safety precautions, graceful fails proposals seem harder to implement. The challenge is to get an AI project that might result in uncontrolled AI to care about whatever proposal there is to implement. Perhaps there are proposals to consider in rather early stages of AI development where researchers think hard takeoffs are very unlikely. Should the AI design in question nevertheless undergo an intelligence explosion, there might be ways to make the result counterfactually less bad.

- Dummy goals: We could come up with a goal function with the following properties:
  - Easy to program
  - Interesting (in the sense that it's easy to create problems for an AI with this goal function to solve, such that researchers can track their progress on general intelligence)
  - Leads to decent outcome in the unlikely case of hard takeoff

If AI research and testing at early stages is always conducted with goals of this type, an unanticipated hard takeoff would at least be comparatively benign.

4 Concluding thoughts

Suffering-focused AI safety, and "fail-safe" measures in particular, make up a large and promising area of interventions for FRI (and perhaps other organizations) to investigate further. Advantages are that they are neglected and often "less ambitious" than classical AI safety, which means they might end up being more tractable. Moreover, "fail-safe" measures are a promising project to focus on because the interventions discussed are positive or at worst "unobjectionable" from the perspectives of virtually all other value systems.

The main reason we had not explicitly zoomed in on this category of interventions earlier was that FRI was initially too focused on answering the more general, complicated question whether AI safety on the whole is net positive for suffering reducers. Upon reflection, this question may be less important. AI safety is a very broad category where we should expect a lot of room for targeted efforts. The general lesson to draw is that partitioning

broad categories can be an important step towards making progress. More research into the feasibility of suffering-focused AI safety carries high information value for this very reason: As we look into more proposals and their practical feasibility, we might uncover more relevant distinctions among AI outcomes, value-transfer proposals, or general approaches to AI – a process which, in turn, will make it easier to identify effective interventions.

Acknowledgements

Brian Tomasik provided the central idea for this paper and gave important feedback on earlier versions. Thanks also go to Caspar Oesterheld and Max Daniel for contributing to the ideas and scenarios in the paper, and to Simon Knutsson, Adrian Hutter, David Althaus, Adrian Rohrheim, Andrei Pöhlmann, Jan Leike and Kaj Sotala for valuable feedback and edits.

References

Armstrong, M. S., and L. Orseau. "Safely Interruptible Agents." *Machine Intelligence Research Institute* 2016.

Bostrom, Nick. "Are We Living in a Computer Simulation?." *The Philosophical Quarterly* 53.211 (2003): 243-255.

Bostrom, Nick. "What Is a Singleton." *Linguistic and Philosophical Investigations* 5.2 2006: 48-54.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford UP, 2014. Print.

Hanson, Robin. *The Age of Em: Work, Love and Life When Robots Rule the Earth.* Oxford: Oxford University Press, 2016. Print.

Hasler, Jennifer, and Harry Bo Marr. "Finding a Roadmap to Achieve Large Neuromorphic Hardware Systems." *Frontiers in Neuroscience* 7 2013: 118.

Hibbard, Bill. "Model-based Utility Functions." Journal of Artificial General Intelligence 3.1 2011: 1-24.

Muehlhauser, Luke, and Louie Helm. 2012. "Intelligence Explosion and Machine Ethics." *In Singularity Hypotheses: A Scientific and Philosophical Assessment.* Ed. Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. Berlin: Springer, 2012. Print.

Omohundro, Stephen M. "The Basic AI Drives." *AGI.* Vol. 171. 2008.

Ring, Mark, and Laurent Orseau. "Delusion, Survival, and Intelligent Agents." International Conference on Artificial General Intelligence. Springer Berlin Heidelberg, 2011.

Sandberg, Anders. "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains." *Journal of Evolution and Technology* 5.1 1999: 1-34.

Sandberg, Anders, and Nick Bostrom. *Whole Brain Emulation: A Roadmap*. Future of Humanity Institute, 2008. Print.

Soares, Nate, and Benja Fallenstein. "Toward Idealized Decision Theory." *arXiv preprint arXiv:1507.01986* 2015.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction.* Vol. 1. No. 1. Cambridge: MIT press, 1998.

Tomasik, Brian. "Do Artificial Reinforcement-Learning Agents Matter Morally?." arXiv preprint arXiv:1410.8233 2014.