List of current research projects

The projects detailed below highlight my current research interests and could be used to set the research agenda for a focused research organization.

My previous work can be found here.

Table of Contents

International spending agreements

[Link]

Summary

Here, I argue that we should advocate for an international agreement to get countries to spend ~1% of GDP on reducing global risks, similar to the agreement to spend 0.7% on aid.

I proceed as follows:

1. I highlight three consequential properties of international agreements to spend a fixed percentage of GDP on both global public goods and bads (global risks):
    1.1. They are large in scale.
    1.2. Their game theoretical features solve public good problems like the free rider problem.
    1.3. Their coupling to growth affects differential technological development.

2. I make the case for the 1% of global risk reduction target. This would lock in a very large, consistent flow of funding to reduce global risks. Also, because funding is pegged to GDP, it hedges against risks from emerging technologies.

In the appendix, I review already existing international agreements to spend part of GDP on military, R&D, aid, global governance, etc. and how they affect global risks.

Moravec's law: AI will progress faster than we think

[Link]

Summary

Cognitive biases affect how we judge existential risks.[1]

Moravec's paradox states that, surprisingly, reasoning (which is high-level in humans) needs little computation, but sensorimotor skills (comparatively low-level in humans) needs a lot of computation.' Put simply: 'For the brain, easy things are hard, and hard things are easy.'[2].

Here, I introduce *Moravec's bias*, systematic inability for humans to overcome Moravec's paradox.

Moravec's bias makes us believe that sensorimotor tasks (e.g. robotics, computer vision) are easier to automate than cognitive tasks (e.g. mathematical cognition, language). But, in fact, it's the other way round: cognitive tasks are easier to automate. This biases us to underestimate how challenging it is to automate tasks like driving or cleaning, and underestimate how easy it is to automate verbal and mathematical intelligence.

Moravec's bias also makes us believe that AI will develop more slowly than it actually is, because we systematically overestimate how much compute is needed for higher-level human cognition.

Just like optimism bias causes 'Hofstadter's law' – Moravec's bias causes its corollary *Moravec's law*: *'Automating sensorimotor tasks (e.g. self-driving cars) always take longer than you expect and cognitive tasks (e.g. language) never take as long as you expect—even if you take into account Moravec's Law.'*

---

[1] Cognitive Biases Potentially Affecting Judgment of Global Risks

[2] Another explanation of Moravec's paradox from Josh Greene's dissertation "There's a saying among cognitive scientists: "Easy things are hard," (Minsky, 1985, pg. 29). That is, the most computationally complex things we do are most often the things we do with the least effort and the least understanding of how we do them. For example, people have no trouble at all determining whether or not a given human face is male or female. We don't know how we do this. We just do it, quickly and effortlessly. Except in rare instances, we have no experience of figuring out whether or not the human faces we encounter are male or female even though these judgments are in fact extremely complex (Wiskott et al., 1995). From our point of view, human faces just look male or look female. The complexity of these judgments is buried deep in the subconscious processes that make these judgments on our behalf. In contrast to such "easy" tasks as sorting faces by sex, we find tasks such as memorizing the state capitals or finding the square root of 2,345 considerably more difficult. But these tasks, which require from us a relatively large amount of effort, are actually much simpler than the aforementioned facesorting task in terms of their computational demands. They can be performed very quickly by very simple computers running very simple programs. Why are so many of the things we find easy actually quite hard and vice versa? Part of the answer is that the forces of natural selection that have shaped our brains have lavished more R&D on some of our abilities and less on others. Sorting conspecifics by sex is a longstanding challenge to sexually reproducing species, and it's no surprise that our abilities in this regard reflect many millions of years of practice. Having the names of the state capitals at one's disposal and an ability to calculate square roots in one's head can be very useful under the right circumstances, but our use for such skills is a relatively recent development, and our lack of practice is revealed in our bumbling efforts to accomplish these tasks. It's worth noting that some people are more bumbling than others, and in different ways. Consider autistic savants who lack the most basic of social skills, who lack "common sense," but who can perform inordinately complex mathematical calculations in their heads (Dehaene, 1997). The existence of these individuals gives us insight into the sorts of tradeoffs that mother nature "

Thus, we might underinvest in AI governance and safety.

The key claims here are:

1. Most of the brain's computations are for sensorimotor processes in rich real-world environments that are not needed for human-level AIs in abstracted environments.
2. This is evidenced by:
    a. Some animals with brains with as much compute as human brains do not have higher level cognitive abilities.
    b. Some animals with brains much smaller than humans do have higher level cognitive abilities.
3. Human-level cognition does not need much compute.

Thus, AI will solve cognitive tasks (e.g. scientific discovery), before solving sensorimotor tasks (e.g. driverless cars). Because we're close to solving sensorimotor tasks, we are closer to AI solving cognitive tasks. I show that this prior is pervasive even amongst experts who haven't updated their model, and keep being surprised by lack of progress in robotics ('self-driving cars are always 5 years away'), which they overinvest in, while also being surprised by AI progress in cognitive areas (e.g. GPT-3).

Diversification to reduce global risks

[Link]

Summary

What if everyone owned a little bit more of everything? Here, we argue that we should diversify the ownership of corporations, in order to mitigate the dangers of unaligned transformative artificial intelligence (TAI).

In Part 1, we highlight three benefits of diversification:

1. Power becomes decentralized, reducing the risks from unaligned malevolent actors.
2. Actors developing TAI face fewer incentives to profit from negative externalities.
3. Greater global collaboration and less adversarialism in TAI development.

We believe diversification is tractable, as diversification both:

4. Hedges against risks to the domestic economy

And:

5. Increases the expected returns of investors

In Part 2, we propose three concrete policies that diversify the ownership of corporations. All countries should:

1. Create sovereign wealth funds (SWFs) that passively invest in the global market portfolio.
2. Create tax incentives to invest in the global market portfolio and reduce distortions and disincentivizes from doing so.
3. Improve corporate taxation through self-assessed taxes.

We close with a series of drawbacks, risks, and reservations of these policies.

Effective Altruism and Philanthropy

Summary

Here, I write about the role of philanthropy in Effective Altruism. More specifically, I look at the role giving plays for small donors and how it relates to their involvement with EA. I argue that this has changed in recent years and has some implications for the community.

The initial excitement about EA came from people discussing donations. Having 'skin in the game' and needing to 'put your money where your mouth is' led to lots of critical engagement, independent thinking, analysis and discussion of ethics and prioritization (I'll call this 'deep engagement'). In other words, what made EA special was that there was not just empty talk about what to prioritize, like in many other political communities, but there was a very concrete implication—donating to a certain cause or charity over another—that distinguished it from it.

A lot of the initial excitement about EA came from active donating, forcing people to engage with difficult questions in ethics and prioritization. Prioritization and discussion of ethics used to be 'Task Y' the 'scaleable use of the people' in the movement. As a result of this deep engagement, some early EAs became well-rounded generalists with some in-depth knowledge of different causes, ethics, and prioritization. Donations have this special quality in that, unlike with ethical career choice, donations can be continuously updated and discussed.

As one commenter puts it:

'**EA feels stagnant both intellectually and socially** [...] Newer EAs seem to be looking to the older EA intellectuals to tell them the answer to what they should do with their lives and how they should think about the world. Something I liked about the vibe of the EA community in the past was the sense of possibility; the sense that there were many unanswered questions and that everyone had to work together to figure things out.' How have you become more (or less) engaged with EA in the last year?

But now the community discourages deep engagement about prioritization through practices like:

- **Donor lotteries**
- Outsourcing donation decisions to experts through funds

- Charity evaluators nudging donors towards 'fire and forget' standing orders and unrestricted funding to regrant at their discretion
- **Direct work**, which generally includes some salary sacrifice, and is theoretically equivalent to a standing order to one's employer (the donative labor hypothesis)

These practices emerged because of perverse incentives and goodharting for short-term, easy-to-evaluate metrics like 'money moved' over deep engagement metrics such as whether the people make more effective donations.

The benefits of these practices are small: though it makes small donors more effective, due to wealth inequality, the overall amounts moved are trivial.

But the costs are large: it decreases deep engagement and increases information cascades, groupthink, 'buzztalk', and what Weyl calls technocratic blind spots (see Appendix). There's also less discussion across causes as people are more siloed into their causes.

There seems to be something missing in EA discussions these days, and perhaps this is it. Newcomers seem really motivated to do something and help, but they end often up discussing trivial things (see for instance the rise of productivity advice on the EA forum 'Effective Egotism', which is not neglected at all by the wider world and not EA's comparative advantage).

Given that the costs outweigh the benefits, some of these practices are net negative for community health.

But if what we 'really want is for people to spend more time discussing why they make donations, prioritize certain causes above others'[3], if we really 'want to give prospective donors the choice' and if 'we're very strongly encouraging people to try to think this through for themselves, at least a little bit.'[4], and have 'skin in the game', then perhaps we need to give people back control over their own donations.

I argue that the community has changed because the community emphasizes recent developments. I show that donor lotteries and outsourcing donation decisions to experts through philanthropic funds (as well as similar related practices) have negative externalities—because these practices reduce deep thinking, critical engagement, learning, analysis, and discussion of ethics and prioritization.

In brief, I make the following claims:

1. The EA community encourages certain practices like lotteries and funds (see **Table 1**) that cause small donors (<$10k/y) to think less about where to donate and thus think less about prioritization and (perhaps) ethics.

2. The main benefits of these practices are that they make donations more effective, because lotteries leverage economies of scale and funds leverage expertise.

---

[3] https://forum.effectivealtruism.org/posts/YsH8XJCXdF2ZJ5F6o/i-want-an-ethnography-of-ea?commentId=6dYR4xQfwtq86vXob

[4] Effective Altruism is trying to save mankind from extinction — Quartz

3. However, due to wealth inequality, small EA donors, even when pooling their resources, will only ever donate trivial amounts of money.
4. Thus, the total benefits of all small EA donors who outsource their donation decisions are very small.
5. The costs of these practises is that they decrease the amount of thinking and discussion about prioritization.
6. Thus, *how* small donors donate is more important than *where* they donate.
7. Because the costs outweigh the benefits, these practices are net negative from a community health perspective.

Implications: I propose three ways on how to improve current practices to improve the situation:

1. changing current donation practices to cause more deep engagement
2. offsetting the lack of deep engagement by funding (student) prioritization projects or giving games
3. deemphasizing small donor philanthropy entirely

Innovating on mechanism design in philanthropy is important and I applaud the hard work that goes into lotteries, funds, etc.—the following is meant to be productive criticism.