A judgmental estimate of the number of people in AI safety

Nuño Sempere

July 12, 2024

Abstract

Estimating funding sources and salary sizes allows to come up with an order of magnitude estimate for the size of the AI safety community of ~600 to ~3.3K. That range excludes academia, the size of whose contribution remains uncertain but still potentially very large.

Contents

Estimation strategies

Here are various estimation strategies we can use:

1. Direct judgmental estimation of the size. We could ask forecasters, or some other population, for their best guess as to what the size could be.

2. Enumeration. We could try to enumerate specific people working on AI safety. This has the advantage of being at least a solid lower bound. However, it is also more labor intensive.
    - You could enumerate a number of people, and then later somehow estimate what fraction you've captured.
3. Do a Fermi estimate by multiplying different factors to arrive at the final estimate. This has the advantage that the relative error of the final estimate will be lower[1].
    - One specific Fermi estimation strategy is to first estimate what fraction of total funding goes to salaries, and then divide total funding available by average salary.

Throughout this brief, we will use all three strategies:

- Direct judgmental estimation for the number of people working on AI safety at AI labs
- Enumeration + adjustment to estimate the contribution of academia
- A Fermi estimate around dividing total funding by average salary to estimate the contribution of the nonprofit and government sectors
- Enumeration of the number of authors in the AI alignment forum as a lower bound sanity check

Models

Desiderata

When writing a model in a specific stack, we have various trade-offs:

- Easily readable. Can the target audience read the model?
- Easily modifiable. Can the target audience modify the model?
- Reliability. Is the tool reliable? Are there likely to be bugs, or is it likely for errors to be
    - Simplicity of the stack. Is the method by which the results are arrived at straightforward? or are we relying on a jenga tower of dependencies?
- Reproducibility. Can a dedicated operator reproduce results now? What about a year from now? What about in ten years?
- Powerful or expressible. Different types of calculations and manipulations, of thoughts, should be expressible.
- Fast.

Links to the models

These goals pull in different directions. To have broad coverage for these goals, we present the model in three forms[2]:

[1] See the last section here

[2] Some other alternatives I considered were UseCarlo, Guesstimate, squigglepy and squiggle.c.

1. A google sheet / csv with presents model figures and sources. This is readable and accessible, but not particularly modifiable. It presents sources and comments for the various estimates.
2. A squiggle model which replicates some of the underlying calculations, and which should be easily editable by readers. This is powerful and modifiable, but not so simple and not so fast as one requires more samples.
3. A model in a simple custom DSL, that is simple, fast and reliable, but not very accessible.

Auxiliary programs used to gather data

Auxiliary programs used to gather data are on the "sources" folder of this repository.

Model results

I estimate the size of the AI safety field as ~900 to ~5K without including academia, or ~1.4K to ~12K including it.

Discussion

Quality adjustments

These numbers are of raw headcount. However, it's unclear what kinds of decisions one can make with reference to those—as opposed to with reference to some quality-adjusted weighting.

One casual approach to make such quality adjustments might be to pick a level of impressiveness and genius, e.g., a von Neumann, and estimate how many "von Neumann-equivalents" of brainpower the AI safety field has at its disposal. For illustration, one might arrive at an estimate that the nonprofit-funded part of AI safety has half a von Neumann, the AI safety work done at labs has two von Neumanns, and academia has a tenth of a von Neumann.

One reasonable place to apply quality adjustments to seems AI labs against everyone else, as they seem to have better talent, and definitely have more compute. If one went down this route, more accurate estimates of the size of safety teams at labs would become more important.

Civilizational neglectedness Still, from the raw headcount, one could maybe try to make some arguments about whether AI safety is "civilizationally neglected".

As a few points of comparison, I thought it would be interesting to look at the size of various militaries. Some numbers here are the size of the armies of:

- Kuwait (17,500)
- Paraguay (13,950)
- Ireland (9,500)

- New Zealand (9,000)
- Estonia (7,100)
- Latvia (6,200)
- Papua New Guinea (3,600)
- A Spanish tercio (~3,000)
- Montenegro: (2350)
- Equatorial Guinea (1,450)
- Luxembourg (900)
- The Seychelles (420). For a historical comparison,
- The royal guard of Sparta (300)

I also thought it would be interesting to look at the size of various US government agencies. Some of their headcounts are: Department of State—13,963, Department of Education—4,269, Securities and Exchange Commission—2875, Office of Personnel Management—2730, Federal Communications division—2,359, the National Labor Relations Board—1,637), the Commodity Futures Trading Commission—726, the Farm Credit Administration—338, the Federal Election Commission—303, or the Holocaust Memorial Museum—121.

Overall, the field seems to have the size of the army of a smallish nation, or the size of a medium-sized federal agency in the United States. Although one can argue that its size is still small relative to its importance, I think this now *does* have to be argued, since the field now does have many more than just a few PhDs.

Steep quality adjustments and discounting factors

> Once I saw this guy on a bridge about to jump. I said, "Don't do it!" He said, "Nobody loves me." I said, "God loves you. Do you believe in God?"
>
> He said, "Yes." I said, "Are you a Christian or a Jew?" He said, "A Christian." I said, "Me, too! Protestant or Catholic?" He said, "Protestant." I said, "Me, too! What franchise?" He said, "Baptist." I said, "Me, too! Northern Baptist or Southern Baptist?" He said, "Northern Baptist." I said, "Me, too! Northern Conservative Baptist or Northern Liberal Baptist?"
>
> He said, "Northern Conservative Baptist." I said, "Me, too! Northern Conservative Baptist Great Lakes Region, or Northern Conservative Baptist Eastern Region?" He said, "Northern Conservative Baptist Great Lakes Region." I said, "Me, too!"
>
> Northern Conservative Baptist Great Lakes Region Council of 1879, or Northern Conservative Baptist Great Lakes Region Council of 1912?" He said, "Northern Conservative Baptist Great Lakes Region Council of 1912." I said, "Die, heretic!" And I pushed him over.

There is this perspective that efforts outside some narrow slice should be greatly

discounted. For instance because research requires very focused attention, because those outside that slice are not smart or rational enough, because those outside that narrow slice don't truly get it, etc.

That stance will greatly affect estimates of the number of people that are working "on the part of the problem that matters", or doing so in a way that seems likely to succeed.

I see some sense in those perspectives, but personally I think that discounting factors are sometimes just too steep.

Could the academic contribution really be that large?

Here are some mechanisms through which the contribution of academia could be larger than the contribution of the rest of the AI safety field:

- If notions about fairness and short-run bias avoidance end up being important
- If interpretability efforts are scaled up in academia
- If academia develops different concepts and strategies that end up being useful
- If work from other fields, like philosophy or law, ends up being relevant

Leech in 2020 does a cursory look at the contributions of academia and concludes it is plausible that its contributions could be as large or larger than that of the rest of AI safety. This still seems plausible four years later.

Definitional ambiguities

The above estimate refers to the number of people. Like in many real-life estimates, there is some ambiguity about exactly what it refers, and whether some specific case would be included or not included.

Those ambiguities are sometimes resolvable. For example, it is meant to include operations and support people at AI safety organization, since we are considering total budget divided by average salaries. And when the AI safety field leverages itself—for instance by convincing Biden to sign an executive order, or the State of California to enact some law—then we are not including the headcount of the national security establishment or of the California civil service.

Funders may want to appear larger than they are

Initially when doing these estimates, one tricky aspect was that researching and talking to people about small funders initially gave me the impression that they were giving away more money to AI safety than they actually were, or with more certainty. In practice, "the Oilmoneybags foundation, which has an endowment of $123B, is interested in AI safety" might translate in them, maybe, doing exploratory grants of $200K to $2M. Ditto for the "McGoldman Family Foundation", but for smaller amounts. Eventually, either of these could turn

into significant money out of the door, but my sense is that this is not the case in the year 2024.

But at the same time… not all donations have to be reported publicly, and if someone was trying to stay quiet, I don't naturally expect to know about. Overall this does bring the tail end of the estimate for smaller funders and individuals up, and is also generally a limitation of a model based mostly on public information.

Confidence in the model

Overall, the estimate of ~900 to ~5K for non-academia contributions seems in the right order of magnitude. Looking at one order of magnitude below, the amount of people working on AI safety is definitely higher than Dunbar's number. Looking at one order of magnitude above, it's definitely lower than 50K.

Correlation of unobserved variables One particular potential issue is that we are estimating the size of different funding sources separately, and then adding them up. But it seems possible that they each have bias in the same direction[3], which could result in bias in the final estimate vs a situation in which their errors were uncorrelated.

To account for this, I multiply times a holistic adjustment o 0.6 to 1.5, to make the funding distribution wider.

A sanity check using the alignment forum We can use the alignment forum API to fetch the number of people that have written at least one post over 2023. If we do so, we arrive at a number of 244. So, on the lower end of our previous estimate, about half of the AI safety community (excluding academia) would be posting—or is being crossposted tom—the AI alignment forum. This seems like a high fraction, but perhaps not outside the realm of possiblity.

The most sensitive variables

The two most important factors in the model are:

1. Share of the budget that is dedicated to salaries
2. Fraction of academic research on AI that is safety relevant

If one wanted a narrower estimate, it might be worth polilng a few organizations for the first number.

[3]This is similar to how polls in the 2016 election in different states had each a bias against Trump.

Algebra of distributions

When I was starting to manipulate distributions a few years ago, one feature I found counterintuitive is that the distribution of the sum of a variable whose 90% confidence interval is 1 to 10, and the sum of a variable whose 90% confidence interval is 3 to 30 is, by an large, not 4 to 40. For instance for two lognormally distributed variables, it is 5.7 to 34.9.

Conclusion

An estimate for the size of the AI safety community of 900 to ~5K members seems reasonable, though subject to the nuances we explored above. The contribution of academia remains potentially large, but it seems so uncertain as to be worth reporting separately.

Acknowledgments

Thanks to Benjamin Hilton of 80,000 hours for commissioning this estimate.

References

Stephen McAleese, 2023 (updated 2/January/2024), *An Overview of the AI Safety Funding Situation*, https://www.lesswrong.com/posts/WGpFFJo2uFe5ssgEb/an-overview-of-the-ai-safety-funding-situation

Gavin Leech, 2020, *The academic contribution to AI safety seems large*, https://forum.effectivealtruism.org/posts/8ErtxW7FRPGMtDqJy/the-academic-contribution-to-ai-safety-seems-large

Vilhelm Skoglund, Jona Glade, 2023, *Observations on the funding landscape of EA and AI safety*, https://forum.effectivealtruism.org/posts/RueHqBuBKQBtSYkzp/observations-on-the-funding-landscape-of-ea-and-ai-safety

See also sources referenced here