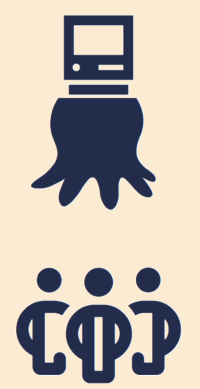
Impact Assessment: AI Safety Camp



An Arb Research report

Sam Holton,
January 2024

Authors: Sam Holton, Misha Yagudin

Data collection: David Mathers, Patricia Lim

Conflict of interest: Arb was commissioned to produce this impact assessment by the AISC organizers. Further, Arb's directors (Misha and Gavin) are AISC alumni and have friends in the community. Sam's investigation was independent, but Misha, Gavin, and the current AISC organizers (Linda and Remmelt) were invited to comment on this report before publishing.

Executive Summary

- AI Safety Camp (AISC) connects people interested in AI safety (AIS) to a research mentor, forming teams that last for a few weeks and then write up their findings.

- To assess the impact of AISC, we first consider how the organization might increase the productivity of the Safety field as a whole. Given its short duration and focus on introducing new people to AIS, we conclude that AISC's largest contribution is in producing **new AIS researchers that otherwise wouldn't have joined the field.**

- We gather survey data and track participants in order to estimate how many researchers AISC has produced, finding that 5–10% of participants plausibly become AIS researchers (see "Typical AIS researchers produced by AISC" for examples) that otherwise would not have joined the field. **AISC spends roughly $12–30K per researcher.**

- We could not find estimates for counterfactual researcher production in similar programs such as (SERI) MATS. However, we used the LTFF grants database to estimate that the **cost of researcher upskilling in AI safety for 1 year is $53K**. Even assuming all researchers with this amount of training become safety researchers that wouldn't otherwise have joined the field, AISC still recruits new researchers at a similar or lower cost (though note that training programs at different stages of a career pipeline are compliments).

- We then consider the relevant counterfactuals for a nonprofit organization interested in supporting AIS researchers and tentatively conclude that funding the creation of new researchers in this way is slightly more impactful than funding a typical AIS project. However, this conclusion is highly dependent on one's particular views about AI safety and could also change based on an assessment of the quality of researchers produced by AISC.

- We also review what other impacts AISC has in terms of producing publications and helping participants get a position in AIS organizations.

Our Approach

To assess impact, we focus on AISC's *rate of net-new researcher production*. We believe this is the largest contribution of the camp given their focus on introducing researchers to the field and given the short duration of projects. In the appendix, we justify this and explain why new researcher production is one of the most important contributions to the productivity of a research field. For completeness, we also attempt to quantify other impacts such as:

1. Direct research outputs from AISC and follow-on research.
2. Network effects leading to further AIS and non-AIS research.
3. AISC leading to future positions.

AISC plausibly has several positive impacts that we were unable to measure, such as increasing researcher effort, increasing research productivity, and improving resource allocation. We are also unable to measure the quality of AIS research due to the difficulty of assessing such work.

Data collected

We used 2 sources of data for this assessment:

1. **Survey.** We surveyed AISC participants from all camps, receiving 24 responses (~10% of all participants). Questions aimed to determine the participants' AIS involvement before and after camp as well as identify areas for improvement. To ensure honest answers, we promised respondents that anecdotes would not be shared without their direct permission. Instead, we will summarize common lessons from these responses where possible.

2. **Participant tracking.** To counter response biases in survey data, we independently researched the career path of 101 participants from AISC 4-6, looking at involvement in AI safety research before and after camp. We further identified individuals who increased AIS research after attending camp and assessed whether those individuals would have succeeded without AISC.

Based on this data we:

1. Estimate how many counterfactually-new researchers AISC produces
2. Provide a glimpse into "typical researcher produced"
3. Compare that to other opportunities to produce researchers (specifically, LTFF upskilling grants).
4. Compare the value of producing a new researcher versus funding an existing project

Impact assessment: new researcher production

Assumptions

1. We assume that more research in the field of AIS is good. This may not be true if AIS research is ineffective or if such research also increases AI risk.
2. Relatedly, we assume that there are no negative effects of AISC on participants or the field as a whole.
3. We assume that already-established researchers get no post-camp benefit.
4. We assume conversion of new researchers is the most important effect.

Potential Issues

1. AISC draws from people already interested in AIS, so researchers who appear to have a step-change in participation may not have needed AISC to break into AIS research in the first place.
2. Survey bias: surveys tend to obtain both highly positive and highly negative responses
3. Small sample size that makes most estimates noisy

Estimating the rate of new researcher production

To estimate the number of individuals that became AIS researchers after AISC, we examined the publication history of participants in AISC 4–6, looking for individuals who went from no publications before AISC to at least one publication after AISC (not including their AISC project or follow-on work from that project).

We decided to limit our focus to these camps for two reasons; first, these camps were far enough in the past that we can observe participants' subsequent research in AIS, second, all three camps were run virtually, which should ideally reduce variance associated with camp location and organization.

In total, 21 / 101 (20.8%) studied individuals have post-AISC publications relating to AI/AIS while having none before camp. Optimistically, these individuals would not have had AIS publications if it were not for AISC.

To obtain a more conservative estimate, we looked more closely at these 21 individuals to filter out people with prior research experience in AI or related fields. Of these, we identified 8 / 101 (7.9%) individuals who plausibly changed their career trajectory towards AIS after attending AISC.

Turning to our survey, 4 of 24 respondents (16.7%) believed that AISC was pivotal in getting them to start work in alignment research, with 8 / 24 (33.3%) mentioning that AISC provided them with a nudge in that direction (but believed they were already headed towards safety research before starting AISC). Note that survey data can be biased towards extreme positive and negative responses. So the observed rate of researcher conversion in the survey is likely

too high. If we take the conservative assumption that none of the non-respondents were converted into AIS due to AISC, we get a conversion rate of 4 / 249 (1.6%).

Of all these estimates, the 7.9% figure seems like the most reasonable given the biases in survey data. Based on this, I estimate a 5–10% rate of conversion if AISC would be run as is (p=70%).

Dollar cost per new researcher produced by AISC

- The organizers have proposed $60–300K per year in expenses.
- The number of non-RL participants of programs have increased from 32 (AISC4) to 130 (AISC9). Let's assume roughly 100 participants in the program per year given the proposed size of new camps.
- Researchers are produced at a rate of 5–10%.

Optimistic estimate: $60K / (10% * 100) = $6K per new researcher
Middle estimate 1: $60K / (5% * 100) = $12K per new researcher
Middle estimate 2: $300K / (10% * 100) = $30K per new researcher
Pessimistic estimate: $300K / (5% * 100) = $60K per new researcher

Typical AIS researchers produced by AISC

Looking at the 5 survey respondents who claimed that AISC was pivotal to their move to AIS, Gavin Leech (also co-founder of Arb research) is representative of a typical AIS researcher produced by AISC, he is currently a PhD in AI at the University of Bristol. The most impactful of these researchers appears to be Lucius Bushnaq, who currently works as a research scientist at the safety organization Apollo Research.

Looking at the 8 studied individuals who plausibly changed their career trajectory towards AIS after attending AISC, Fabien Roger, now at Redwood Research, seems to be representative of the caliber of new AIS researchers produced. On the high end, Alex Mallen, now at EleutherAI, appears to be the most impactful researcher that plausibly had a trajectory change due to AISC.

Several other participants have gone on to have successful careers in AI safety, but it is likely that AISC played a smaller part in their career trajectory. These include Rai (Michael) Pokorny who was a software engineer at Google before AISC and transitioned to the Superalignment team at OpenAI.

Dollar cost per new researcher produced by other means

When considering other ways to create AIS researchers, we can think of individuals following a sequence of steps (graduate school, research projects, etc.) to increase their ability and experience in AIS. People start with little experience in AIS and develop skills to become an established researcher. At any stage of researcher development there are two ways to encourage people to continue further on the path: pull mechanisms and push mechanisms.

Push mechanisms directly assist someone in completing a particular stage. This could be via education, financial support for upskilling, or assisting people with applications. On the other hand, pull mechanisms typically offer money or prizes to people who have completed a particular stage, which creates an incentive to complete that stage. For instance, offering high-paying positions to experienced AIS researchers creates an incentive to upskill in AIS research. Naturally, we would like to determine which approach is more cost effective at producing established AIS researchers.

We don't have comparable data for programs like MATS that "push" new researchers into the field of AIS. However, one way to "pull" researchers into the field is by offering jobs in AIS positions. The annual salary for an AIS researcher ranges from $60K for a junior researcher working independently to $200–300K base salary for a member of technical staff at various private organizations to $1M if equity at OpenAI/Anthropic is priced-in. In net present value terms, the cost to pull a new researcher into AIS is much higher than the cost to push one via AISC. For government funding sources, the cost to support a graduate student is roughly $60K per year (NSF GRFP offers a total of $53K per year, universities often provide additional funding) and AIS projects are roughly $2 million per grant.

However most AIS research is funded by nonprofit entities. Assuming that the number of researchers is a bottleneck, what is the typical cost to "pull" a new researcher into AIS with non-profit grants such as LTFF? A rough calculation on the LTFF database suggests that it costs $80K per year to fund an AIS researcher.

We can also examine the LTFF database for mention of researcher "upskilling". The annualized average salary for grants that involve upskilling is $53K. Even making the optimistic assumption that every upskilling project creates a new AIS researcher that otherwise wouldn't have succeeded, AISC's cost per researcher compares favorably with this value.

Note that it is difficult to directly compare these estimates since they operate at different stages of researcher development. AISC typically operates earlier on the talent pipeline, while industry positions typically pull more experienced researchers into AIS positions.

Counterfactual analysis

Is a new researcher produced by AISC valuable relative to supporting existing projects? A organization considering funding AISC must choose between:

1. Funding an existing researcher for ~$80K for 1 year.
2. Funding the creation of a new AIS researcher for ~$40K.

In option 2, the new researcher then enters the pool of existing researchers, and may get support from academia, industry, or nonprofits. Alternatively, the funder that supported their transition to AIS also has the option of continuing to support their research.

If the new researcher is able to obtain outside funding from government or industry, then the organization has essentially obtained all of their subsequent research for "free". If the organization chooses to directly support the new researcher, then the net value depends on how much better their project is than the next-most-valuable project. Essentially, this is the marginal value of new projects in AI safety research, which may be high or low depending on your view of the field.

Regardless, a funder wishing to support AIS research may not value the creation of new AIS researchers if the number of researchers is not a bottleneck for the field. In AIS, there are many open questions with no supported researchers working on them. This could indicate that there is either a bottleneck in the number of researchers or in the amount of funding in AI safety. If funding is the bottleneck then producing more researchers will not advance the field. More work is needed to distinguish these possibilities.

Other impacts of AISC

Research outputs from AISC and follow-on research

Dozens of projects were completed at camp. Paraphrasing their funding case, AISC organizers note that alumni authored several important publications such as:

Goal Misgeneralization

AI Governance and the Policymaking Process

Detecting Spiky Corruption in Markov Decision Processes

RL in Newcomblike Environments

Using soft maximin for risk averse multi-objective decision-making

Reflection Mechanisms as an Alignment Target

Participants have been hired for dozens of positions in AIS organizations. Quoting the same funding case considering participants across all camps, the organizers list the following jobs:

> "*FHI* (1 job+4 scholars+2 interns), *GovAI* (2 jobs), *Cooperative AI* (1 job), *Center on Long-Term Risk* (1 job), *Future Society* (1 job), *FLI* (1 job), *MIRI* (1 intern), *CHAI* (2 interns), *DeepMind* (1 job+2 interns), *OpenAI* (1 job), *Anthropic* (1 contract), *Redwood* (2 jobs), *Conjecture* (3 jobs), *EleutherAI* (1 job), *Apart* (1 job), *Aligned AI* (1 job), *Leap Labs* (1 founder, 1 job), *Apollo* (2 founders, 4 jobs), *Arb* (2 founders), *AISS* (2 founders), *AISL* (2+ founders), *ACS* (2 founders), *ERO* (1 founder), *BlueDot* (1 founder)"

Follow-on projects also gathered roughly $600K in outside grants with the median funded project receiving $20K in initial funding and some projects receiving over $100K.

Survey respondents also note several post-camp projects with collaborators from AISC including: Apollo Research, AI Safety Fundamentals, and AI Standards Lab. This last project is a direct result of work done during AISC.

Network effects producing further AIS and non-AIS research

AISC also introduced like-minded individuals to one another, leading to follow-on projects both within and outside of AIS. The median respondent has interacted with 5 members of AISC after camp, with several reporting 10–15 such interactions.

In terms of how many people a participant would feel comfortable reaching out to, the median respondent said 5, with several feeling comfortable contacting 10 or more people.

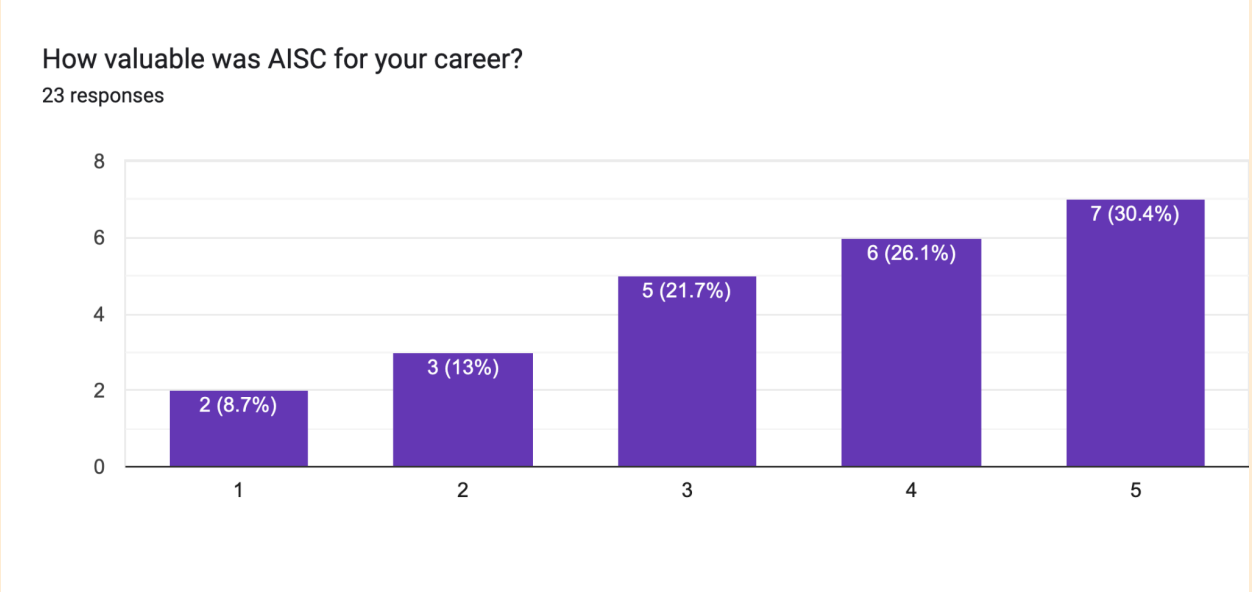
Beyond the follow-on research noted in the last section, two respondents mention new collaborations in AIS that were unrelated to their AISC project, but with people they met through AISC.

AISC leading to future positions

While it's not feasible to determine if a participant's AISC project led to them obtaining a new position in AI/AIS, we can examine a related question: did participants believe their project was substantial enough to include on applications to new jobs? In other words, did they believe that AISC provided a boost to their applications?

In our survey, 14 of 20 (70%) of participants listed their work with AISC on subsequent applications. Two of these 14 believe that their work at AISC was crucial to receiving AIS grants and safety-related jobs.

Additionally, 30% of respondents believed AISC greatly helped their career:

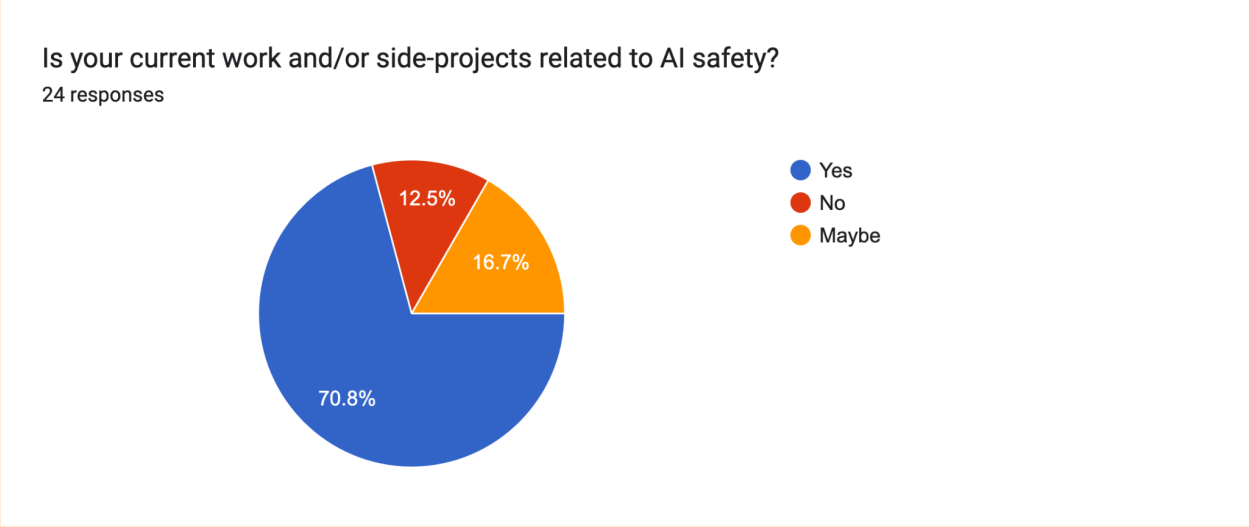


5= Greatly helped my career, 1=Not valuable at all

Other data

Fraction that pursue AIS

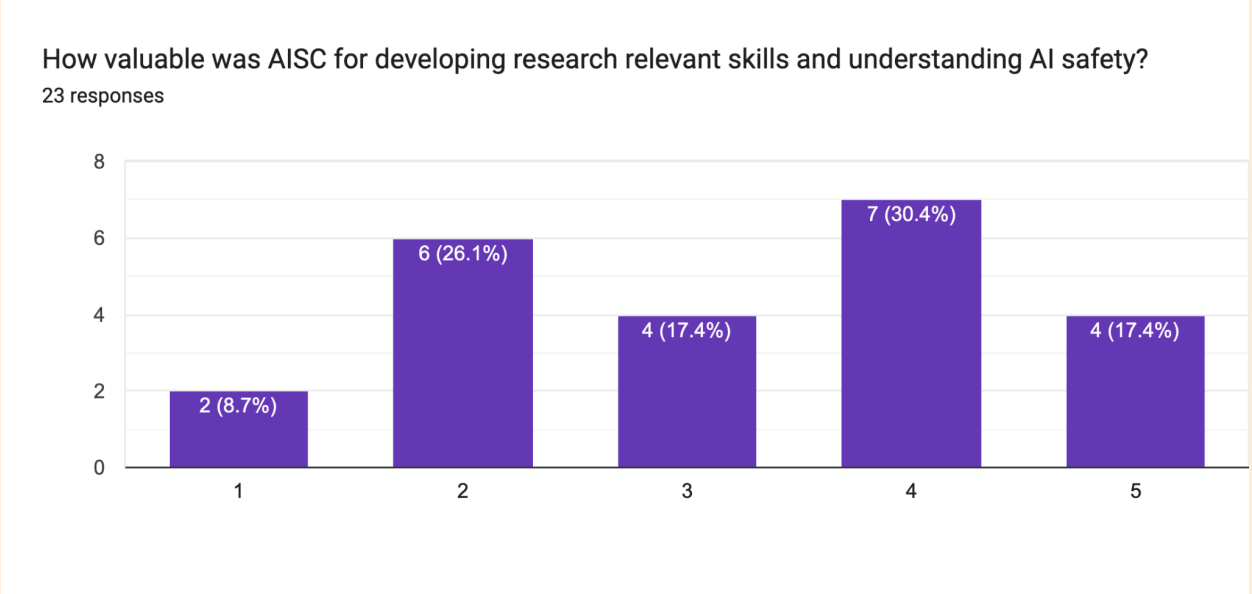
Of the respondents, 17 / 24 (70.8%) work in AIS or have side projects in AIS.



Looking at participants from camps 4-6, 67/101 (66.3%) have some sort of written work related to AI or AI safety (including posts on LessWrong), and of these, 48/101 (48.5%) have some publication in AI or AIS.

- 8 surveyed were beginning their transition into AIS before camp, using camp to assist that transition.
- 1 surveyed left AIS for various reasons while 1 is still aspiring to work on AIS

In-camp experiences: summary of positive and negative experience



5= Greatly helped my research skills, 1=Not valuable at all

Looking at the written responses, people generally had a positive experience of camp, appreciating the opportunity to work with like-minded individuals and have the support needed to start on an AIS project. Very few reported negative experiences from camp and these involved frustrations with project success and team organization.

Conclusion

Our all-things-considered assessment

Overall I (Sam) was surprised by the number of researchers who owed their position in AIS research to AISC. My expectation was that virtually all participants would be on a path to AIS without AISC and that evidence of a true trajectory change would be hard to find. However, there are several clear examples of individuals who changed their career trajectory towards AIS after camp and on surveys several respondents claim that this was directly because of AISC.

Programs like AISC and MATS have the effect of “pushing” new researchers into AIS which can be contrasted with programs like corporate, government, and nonprofit roles that “pull” new

researchers in by offering funding. In other words, “push” programs tend to support potential researchers by providing them the training and experience to take on roles in AIS research while “pull” programs create a financial incentive for people to take on these roles. These approaches are complementary. The effectiveness of spending on these “push” programs depends on who bears the subsequent costs of supporting a new AIS researcher. If a researcher produced by AISC is able to draw subsequent funding from the government or industry for their work, their subsequent research has been obtained for “free” since an organization only needs to pay the startup cost for creating that researcher.

However, if their subsequent funding comes from the same organization that activated them, then the organization must trade off between funding new researchers and funding more projects from established researchers. If an organization is funding constrained, it may be better to focus funds on established researchers. If an organization has much more funding than promising projects, producing new researchers may be more valuable.

Note that it’s difficult to assess the quality of the marginal AIS research produced by these new researchers and this is compounded by the difficulty of assessing the value of a given work in AIS more generally.

I (Sam) would guess that producing more AIS researchers is more valuable to the field than giving more funding to established researchers, especially given the fact that new researchers can help with existing projects and can draw outside support to the field via corporate or government support for their research. The fact that creating a new researcher via AISC is comparable or smaller than the cost of an established researcher’s annual salary, suggests that AISC is an effective way to boost AIS research.

Areas for further research

1. Direct assessment of the quality of research produced at AISC
2. Assessment of research quality post-AISC
3. Better comparison to similar training programs like MATS. What counterfactual benefit does MATS provide for producing AIS researchers? At what cost?
4. Learning why some participants didn’t transition into AIS. Did they lack interest? Did they miss a critical funding source? What could have been done better?

Appendix A: Details on our approach

A model for the productivity of a research field

Simple models of economic growth break innovation down into several inputs such as the number of researchers, the stock of ideas, and human capital. These models allow us to account for different sources of economic growth and suggest policies to boost growth.

Conceptually, these models of innovation can apply equally well to a single field and suggest a simple heuristic for that field's productivity:

Field productivity = (number of researchers) x (researcher effort) x (researcher productivity) x (researcher allocation)

Number of Researchers is relatively straightforward, referring to the total number of individuals capable of participating in the field. **Researcher effort** is analogous to research intensity and accounts for things like the number of hours worked per week. **Researcher productivity** refers to the quality of work produced by a researcher in a unit of time, one could imagine that training and research experience contribute significantly to this factor. **Researcher allocation** refers to how effectively researchers are assigned to projects, having good signals of researcher skill and ensuring that available researchers have an assigned project would help lower misallocation. It's usually a factor that ranges from 0 to 1, with 1 denoting a perfect allocation of resources.

Crucially, note that only the number of researchers can increase without limit. There's a finite number of working hours each day, a maximum level of productivity, and the allocation factor reaches 1 in the best-case. This is analogous to growth models, where Chad Jones notes:

> "Many of the sources of growth that have been operating historically—including rising educational attainment, rising research intensity, and declining misallocation—are inherently limited and cannot go on forever. The key source of sustained growth in the semi-endogenous setting is population growth."

AISC's impacts on the productivity of the AI Safety field

Programs like AISC can plausibly have an impact on all of these factors:

- It can increase the number of researchers by giving people the training and background to do safety research.
- It can increase researcher effort by providing the inspiration and community to work more hours per day. AISC temporarily increases working hours in AIS for the duration of the program.
- It can boost researcher productivity by training students in effective research habits.
- It can improve researcher allocation by providing subsequent funders with a signal of researcher quality and effectively allocates new researchers to projects.

The relative value of these different contributions depends on how the camp is designed. Given the short duration of AISC, it probably can't change researcher effort, productivity, or allocation in the long term. These factors are also very difficult to measure, and will have to be ignored for the rest of this assessment.

However, AISC probably *does* have influence on the total number of AIS researchers, helping people break into the field. This “0 to 1” effect of creating/bringing new researchers is likely the largest impact of AISC.

As noted above, the other factors are bounded in size, meaning that raising the total number of researchers is one of the most important contributions to long-term productivity. For these reasons, we will focus on estimating the number of new AIS researchers AISC produces per unit of input.

AISC’s other impacts on AIS

For completeness, AISC also has other direct impacts on the AIS field such as:

1. Direct research outputs from AISC and follow-on research.
2. Network effects leading to further AIS and non-AIS research.
3. Network effects leading to future positions.

We also attempt to quantify these impacts.

Note on the difficulty of assessing research quality

Assessing the value of a particular research work in AI Safety is very challenging. The alignment problem itself may not be solvable and even then, it’s often not clear how much a particular work contributes to safety versus AI capabilities. For these reasons, we will avoid direct assessments of the quality of research produced during or after AISC, focusing instead on simpler (though flawed) metrics such as number of publications.