**GiveWell:** First let's discuss SIAI's activities. Can you lay out for me everything SIAI is doing?

**SIAI:**

- Michael Vassar is working on an idea he calls the "Persistent Problems Group" or PPG. The idea is to assemble a blue-ribbon panel of recognizable experts to make sense of the academic literature on very applicable, popular, but poorly understood topics such as diet/nutrition. This would have obvious benefits for helping people understand what the literature has and hasn't established on important topics; it would also be a demonstration that there is such a thing as "skill at making sense of the world."
- Eliezer Yudkowsky is currently writing two books on rationality - basically taking the Sequences [a large collection of online essays on rationality-related topics] and turning them into readable, marketable books. He'll also be attending Google's AGI [Artificial General Intelligence] 2011 Conference in the hopes of building the connections necessary to build an FAI [friendly artificial intelligence] team down the line.
- Anna Salomon and Carl Shulman are working on getting our group's ideas out to academia - wherever they might fit and be able to get published. The idea is to improve our prestige and engagement with the broader intellectual community.
- Peter de Blanc (research fellow) is trying to get a paper or two into Google's conference, with a similar general aim. Steve Rayhawk is also working on a position paper for this conference; before that he was surveying the literature on AGI.
- Jasen Murray is running the Fellows program, which has served a lot of different purposes. Mostly it's been a way to bring interesting people to interact with us for a few months, and it's almost paid for itself in terms of donations. Currently the Fellows are working on modeling friendliness; they're also thinking about issues related to the Fermi paradox (and its implications for humanity's chances of colonizing the stars).

Overall the broad categories are:

- Decision theory, friendliness theory, etc. - trying to lay the theoretical groundwork for later work on FAI. One of the main questions here is "How can you build something that can keep modifying its own source code while holding a constant overall goal?"
- Getting published in order to improve our prestige & engagement with the broader intellectual community.
- Evangelism: getting more people interested in what we do, laying the groundwork for building an FAI team in the future.
- Also, Michael Vassar is interested in moving us in the direction of organizing the efforts of good thinkers to solve important problems in the world beyond AI. Persistent Problems Group falls under this category.

**GiveWell:** And what is your financial situation - how much does all this cost and do you have a funding gap?

**SIAI:**

- All fellows (and Jasen) get room and board here plus $1500/mo. The total cost of the Fellows program is about $15k/mo. Peter, Steve and Jasen get room and board plus $1500/mo; the others get just room and board.
- Full-time employees: Amy (Chief Operating Officer) makes around $50k/yr, Michael Vassar $60k, Eliezer Yudkowsky $80k, Anna Salomon $36k.
- The Singularity Summit is (slightly) revenue-positive.
- So total expenses are around $500k/yr. Over the last year we've brought in $500k-700k (of which $250k was from Peter Thiel).

---

**Jasen Murray**

**Comment [1]:** The Persistent Problems Group as I understood it would be doing experimental science in addition to literature review. It would consist of both recognizable experts working with us on and off or collaborating at a distance and promising young scientists working full time. An important component of the idea is to give a group (or several) of scientists the opportunity to focus exclusively on their work without worrying about teaching or grant applications while surrounded by people who want and expect to do revolutionary science.

**Jasen Murray**

**Comment [2]:** I assume AI safety outreach is also significant goal.

**Jasen Murray**

**Comment [3]:** Both Peter and Steve were previously working on the AGI literature review. Both are now working on demos for the conference (Peter submitted a paper).

**Jasen Murray**

**Comment [4]:** The current incarnation of the fellows program is ending at the end of March. Our next program, "Rationality Bootcamp" will begin the first week of June and last until mid August. The fellows were also working on formalizing and Revising Updateless Decision Theory and Intelligence augmentation research.

- We think it's important to have some funding available for opportunities that might come up - conferences, etc. - as well as great people that we might run into and want to hire. So having a slight surplus is a good thing.

**GiveWell:** Do you feel that substantially more funding would translate into substantial expansion / more activities?

**SIAI:** At the moment, we don't have immediate plans for more funding - however:

- Michael Vassar's Persistent Problems Group idea does need funding, though it may or may not operate under the SIAI umbrella.
- We're thinking about upgrading our living/working facilities.
- Our needs and opportunities could change in a big way in the future. Right now we are still trying to lay the basic groundwork for a project to build an FAI. At the point where we had the right groundwork and the right team available, that project could cost several million dollars per year.

**GiveWell:** OK. So it sounds like you aren't yet in a position where you're asking to be recommended by GiveWell to new/outside donors - though you may be later. I'm still wondering, though, about the question of what we should tell your existing donors, i.e., whether SIAI is what one might call a "hold" (keep donating if you're donating; don't if you're not) or a "sell" (don't donate).

I do agree with many of your most controversial views. I agree that UFAI is an existential risk and that FAI, if it could be created before UFAI, would be astronomically beneficial. However, my intuitions say that the specific work you're doing is not going to be helpful in bringing about this goal. My intuitions don't matter much because I know so little about artificial intelligence research and other relevant issues; but when I look at the actions of the people who seem to me that they ought to know better, they mostly seem to ignore your arguments (in the case of the more famous people) and to disagree with you along the same lines I do (in the case of the people I know personally who seem best positioned to evaluate your arguments). Many impressive people are loosely connected with you, e.g., speaking at the Singularity Summit, but few explicitly endorse your mission and capability to follow through on it

**SIAI:** I'm not sure whether you're aware of all the people who do endorse us. There are some who are quite impressive. The biggest two are probably Peter Thiel and Jaan Tallin. The latter is a co-Founder of Skype, who spent several days going around the Bay Area looking for anyone who had a good intellectual rebuttal to the ideas we're promoting, and concluded that he couldn't find any; now he's a donor and explicitly a supporter.

There are some smaller signs of external validation as well. A friend of the community was hired for Google's AGI team, and another may be soon.

**GiveWell:** That's helpful, thanks. I did know about Peter Thiel but not about Jaan Tallin. A conversation with him might push me toward changing my mind.

In addition to affiliations and endorsements, there are other things your group could do to improve its credibility in my eyes. Broadly, I have little sense of whether you have the right team for what you're trying to do. I see little in the way of achievements that I'd call "impressive," i.e., accomplishing something that many have tried to do but few have succeeded at. Certainly you have arguments that seem interesting to me and that I've seen few really knockdown answers to, but that could be because the people who are best positioned to work on these issues are more or less ignoring you. What you don't have much of is "impressive" things like patents, publications (and the publications you have are in philosophy, which is of questionable relevance in my view), and commercially viable innovations.

It's not that I think accomplishing your mission *intrinsically* requires any of these things. It's more that my view of your organization hinges heavily on the question of whether you have people who have rare insights, capabilities, and general rationality, and if in fact you have those things, you ought to be able to translate them into impressive achievements/affiliations. And to challenge yourself to do so would be a way of holding yourselves accountable to something other than your own intuitions; I'd hope you are challenging yourselves in that way.

**SIAI:** We are. Our push to get published in traditional academic journals is one of these challenges. We are also working to gain more prestigious affiliations. The Persistent Problems Group would be another way to demonstrate our overall competence, rationality, etc.

One reason we haven't generated much along these lines to date is that we've done a lot of changing paths. For example Eliezer was creating a new programming language, Flare, that could have taken off and been "impressive," but he later decided that this was the wrong problem to be solving.

**GiveWell:** I understand that and completely understand changing direction when you change your mind. But as an outsider, I am holding back and waiting to see the actual impressive achievements, even if there are good reasons that they haven't been forthcoming so far.

**SIAI:** That's a fair perspective as an outsider, and that's why we are working to prove ourselves.

As to patents and commercially viable innovations - we're not as sure about these. Our mission is ultimately to ensure that FAI gets built before UFAI; putting knowledge out there with general applicability for building AGI could therefore be dangerous and work directly against our mission.

**GiveWell:** I'm skeptical that that's the right call. It seems to me that if you have unique insights into the problems around AGI, then along the way you ought to be able to develop and publish/market innovations in benign areas, such as speech recognition and language translation programs. Doing so could benefit you greatly both directly (profits) and indirectly (prestige, affiliations) - as well as being a very strong challenge to yourselves and goal to hold yourselves accountable to, which I think is worth quite a bit in and of itself. I'm skeptical that the risks outweigh the benefits here.

**SIAI:** We disagree, broadly, but it's something we continue to discuss internally and could change in the future.

**GiveWell:** OK. Well that's where I stand - I accept a lot of the controversial premises of your mission, but I'm a pretty long way from sold that you have the right team or the right approach. Now some have argued to me that I don't need to be sold - that even at an infinitesimal probability of success, your project is worthwhile. I see that as a Pascal's Mugging[1] and don't accept it; I wouldn't endorse your project unless it passed the basic hurdles of credibility and workable approach as well as potentially astronomically beneficial goal.

**SIAI:** We agree with you on that. Others may invoke Pascal's Mugging on our behalf but we don't.

**GiveWell:** OK. So probably the best short-term step I could take to further think about your organization would be a conversation with Jaan Tallin. Longer-term, it sounds like you and I are broadly on the same page - you are actively looking for ways to achieve impressive things, improve your affiliations, and generally demonstrate your credibility to outsiders such as myself.

**SIAI:** That's right.

**GiveWell:** So we'll definitely keep in touch, and I'll be watching for more evidence regarding the issues we've discussed, as well as for updated funding needs.

---

[1] http://lesswrong.com/lw/kd/pascals_mugging_tiny_probabilities_of_vast/