Below is a heavily edited version of an email conversation between Holden Karnofsky, Eliezer Yudkowsky, and Luke Muehlhauser. The original email conversation took place in the early months of 2014. Some errors of continuity/coherence may have been introduced when assembling the emails into one streamlined dialogue.

Holden: GiveWell and its followers have been accused of being tight-lipped and ambiguous regarding the value of the far future. I think we have, in fact, been somewhat tight-lipped and ambiguous on this front. Speaking for myself, this hasn't been for PR purposes, but because it's been legitimately hard to convey my epistemic state on these matters, particularly in the language the questioners tend to use, which (I now feel, though I wouldn't have been able to articulate this until recently) does not do a great job describing or modeling Knightian uncertainty.

My state could roughly be characterized as follows. "I have no clue what the probability is that we colonize the stars or how much that's worth in utilons. I've read the things people have pointed me to on these subjects and remain basically clueless. I don't want to say the probability that the hypothesis of 'Astronomical Waste' is correct is 10^-10; that's something I definitely do not have the confidence to say. Does that mean the probability is greater than 10^-10 and I should accept its qualitative conclusion? That doesn't seem right either. Here's what does seem right. I have looked into the situation in Africa, have some understanding of the situation in Africa and see a path to doing a lot of good in Africa. I don't know how to look into the far future situation, don't understand the far future situation, and don't see a path to doing good on that front that I feel good about (and the people who are insisting that I do so seem unreasonable in multiple ways). Given this state, it seems like it must be more rational and good-accomplishing to give to Africa than 'give to the far future' in the ways that have been suggested. Does that mean I implicitly place an absurdly tiny probability on various claims I can't be that confident about? Maybe, maybe not. Does that mean I don't care about future-people? Certainly not. I value future-people equally to present-people, though I'm not sure I value creating a future-person equivalently to saving the life of a present-person. Does that mean giving to Africa is the best way to help the far future? From my perspective, this doesn't seem crazy; to the extent our far future prospects look bright, it's more thanks to the efforts of past people who thought like me (choose a relatively well-understood opportunity to do good over opportunities that are backed up essentially only by confusions over probabilities) than to people who thought like I would have to think to support e.g. MIRI given the information that I have. Others who share my epistemic state should do similarly."

I think the above is better than I was able to put it at the time. When people asked me about the far future I was tight-lipped and ambiguous because I felt uncertain about the probabilities I was being queried on: unhappy endorsing either a small or a high figure. I think I sometimes tried harder than was warranted to translate my thoughts into the language my questioners were using, which could have led to exaggeration of my beliefs on e.g. the impact of bednets on the far future. Though I don't think my post on Bayesian updating was an example of such overreach; I still stand by that post though I feel it has been misread (people have confused expected value with future value, translating e.g. "Extraordinary claims require extraordinary evidence" to "Extraordinarily good outcomes are extraordinarily unlikely").

Luke: You say that "to the extent our far future prospects look bright, it's more thanks to the efforts of past people who thought like me."

Do you also think that P: "to the extent our far future prospects look bright, it's more thanks to the efforts of past people who chose relatively well-understood near-term opportunities to do good over opportunities that looked high-leverage but less well-understood — e.g. most long-term (20+ yrs) planning — than to the people who put their efforts into relatively speculative long-term efforts"? That could be a more relevant comparison, since you seem to have been persuaded that some of the people working on x-risk aren't doing so for "confused about probabilities" reasons.

Holden: P is distinct from the thing immediately above it (let's call that Q). I endorse both P and Q, but much more strongly endorse Q than P. I find Q more relevant to this particular discussion about how GiveWell has communicated in the past. P is simply a consideration, and one that can potentially be outweighed by an inside view, and perhaps justifiedly is for some MIRI supporters. Q is something I feel strongly about, and that I feel applies to many MIRI fans (not all of them - but some who have read all of the Sequences). It's because of Q that I believed a lot of people to be pretty unreasonable for the way they were thinking about AMF vs. MIRI.

Here's a conceptual dialogue that may be illuminating. It's how things felt to me circa 2009 or something. "LESSWRONG" should be interpreted as an ambiguous hybrid of some LWers I discussed this stuff with. Some of it is different from what you and Eliezer have said in our recent conversations, some of it is not.

GIVEWELL: We found ways to save lives! These are the best giving opportunities we know of.

LESSWRONG: GiveWell's recommendations are interesting but I don't follow them because I think future people have value.

HOLDEN (in some conversation): Huh? I think future people have value.

LESSWRONG: Then why aren't you recommending x-risk?

HOLDEN: Because ... what? I don't know anything about x-risk. I have so many questions. I don't even know how to start answering them. Maybe some day I will. We did a bunch of work to find charities that can save lives! That's pretty good! (Enter "Bayesian adjustment" post: "pretty good, seriously" beats "astronomically good, according to wild speculation")

LESSWRONG: What about MIRI? They've spelled out their case in great detail. Can you evaluate that?

HOLDEN: I find the case weak. ("Thoughts on SI" post)

LESSWRONG: But if there's even a chance!

HOLDEN: That reasoning is invalid, even Eliezer says so.

LESSWRONG: What do you believe the value of the far future is? With what probability will we colonize the stars and how many utilons would that be worth? All possible answers imply either that you don't care about future people, that you assign a probability to space colonization too low to be plausible given the precision of your brain, or that you think bednets are the best path to space colonization. I'm excited to start quoting you as saying one of those three things!

HOLDEN: Boy, I'm really feeling in the mood for some tight-lipped ambiguity.

One good place for this to have gone less frustratingly would have been in the second line, if Less Wrongers had instead said to me "GiveWell's recommendations are interesting, but I've been thinking about MIRI and kicking its tires for a long time, and that's where I'm giving. GiveWell is a product offering commoditized giving opportunities to uninformed donors, and I think I'm an informed enough donor to diverge."

Another would have been for us to communicate that better in the first line. We did try to communicate this — for example here (that's from 2009; I think 2007 was similar but don't have a link) — and I really think we tried hard to communicate this generally, though it's a hard thing to communicate, and of course our priority is attracting our target audience (of high-uncertainty donors), not repelling our non-target audience (which we figure will repel itself fine).

Here's something we may still disagree about: I think giving to AMF is rational for a sufficiently disengaged, uninformed effective altruist with a few thousand dollars to give. I don't think this requires said person to discount the value of future generations. I think I understand this person's epistemic position probably better than you guys can. This person is the explicit target audience of GiveWell's traditional product. This includes many LW readers and enthusiasts who are still pretty far from where they'd need to be to be rationally sold on MIRI.

Luke: On this topic, you might enjoy (or be enraged by) a debate Robin and Eliezer had here and here. In particular, see Robin's comment:

> The question is what should rational outsiders believe, given the evidence available to them, and their limited attention. Ask yourself carefully: if most contrarians are wrong, why should they believe *your* cause is different?

I think MIRI hasn't *remotely* optimized for giving non-experts — including most AI folk — good reason to accept our line of thinking. People have limited time and attention, and instead of spending 5,000 hours to learn AI, learn forecasting, learn x-risk, learn moral philosophy, learn MIRI's arguments, etc., they've got to evaluate MIRI's claims with *quick* heuristics like:

1. Do accomplished academics in AI or tech forecasting seem to agree with MIRI?
2. Do MIRI researchers have normal credentials in the relevant fields?
3. Does even the upper crust of humanity have a track record of being able to figure out the kinds of things MIRI claims to have figured out?
4. Does MIRI submit its arguments to peer review?

5. Does MIRI seem to know why normal, credentialed, accomplished academics disagree with them? Then, does it represent their arguments fairly, and show clearly why its own position is better justified?

So non-experts do a model-combination on (1)-(5), plus a few other models, and the answer is "Mostly, no" for all of them, and they think "Well, maybe I'd change my mind if I knew more about this topic, but given my ignorance I can't see a good reason to think *these* contrarians are going to be among the relatively few contrarians who end up being right about their contrarian views."

Meanwhile, I think it's *simultaneously* true that, as Eliezer wrote:

> by default, world class experts on various topics in narrow AI, produce their beliefs about [intelligence explosion] by snap judgment rather than detailed modular analysis...
>
> ...Science only works when you use it; scientific authority derives from science. If you've got Lord Kelvin running around saying that you can't have flying machines because it's ridiculous... the problem is that he's running on naked snap intuitive judgments of absurdity and the Wright Brothers are using actual math.

So, non-experts have little reason to believe Eliezer, and Eliezer has little reason to doubt his models merely from the fact that credentialed AI or tech forecasting people (who know almost nothing about intelligence explosion or FAI) disagree with him.

And because MIRI hasn't done much of (1)-(5) yet, academics can't take the issue seriously enough to invest the time to make their own detailed, modular analyses (with rare exceptions like Bostrom and Chalmers, who were thinking about intelligence explosion before MIRI existed), and thus Eliezer can't learn much from what outsiders *might* know that he's missing.

Eliezer's usual reaction is something like "I don't have time to optimize for getting most people on-board via quick heuristics, I'm just going to do actual FAI research, and hopefully enough people will grok what I'm doing to fund it."

My reaction is more like "The 'what does MIRI look like on quick heuristics' problem is tractable, and it's critical we solve it. It's probably not what *Eliezer* should focus his time on, but *I'm* going to do something about it."

Holden: That all sounds right to me, except perhaps that you frame it more as a PR problem and less as a learning problem than I would, but clearly you're doing some of both. I think there's a great deal of value, for an insider contrarian, in being determined to get more outsiders on board, and the benefits down the line come in learning and not just PR. At the same time, it makes sense to me (and even seems right) that Eliezer should be doing what he's doing and leaving that activity to someone else ... as long as he recognizes the importance of that activity and at least engages with the person doing it for him enough to help them do their job well.

Luke: Yeah, I was focusing on the PR issue in that email, but clearly there's learning value to be

had as well. For example, I'm currently interviewing Gerwin Klein about formal verification (he did the seL4 formally verified microkernel), and he just taught me that probabilistic formal verification is further along than I had realized, and also pointed me to the best paper I've seen yet about the challenge of correctly translating intuitively desirable safety properties into formal requirements.

Holden: I believe that someone who has read every word of the Sequences, plus engaged in a fair amount of conversation with e.g. Michael Vassar and Carl and Luke, and read all of the various things that MIRI has put out that are supposed to summarize its arguments (including Bostrom's book, which I've only skimmed, but I think my position wouldn't change if I read word for word) ... still can rationally to be in the "high-uncertainty rationalist" camp, and might (defensibly) not affirmatively agree with the MIRI view of AI. This is assuming lack of a technical background and deep familiarity with AI.

In order to rationally consider MIRI an outstanding giving opportunity, I think one generally ought to have (a) deeper knowledge of MIRI's personnel, their strengths and weaknesses, and what they've done to challenge themselves than I have (or than is available from the materials above); OR (b) done some independent investigation making serious effort to extract the best possible critiques from people who are knowledgeable about AI.

If you disagree with this, I'd be interested to hear what it is you'd guess I've misunderstood or underweighted in the materials above, though I recognize that's a pretty difficult guessing game.

I feel like I've had, and seen, quite a few discussions of "why doesn't GiveWell focus on x-risk?" Nearly all the responses have said something like "They don't care about the far future" or "They don't understand Bayesian reasoning" or "[insert mangled interpretation of my 'Bayesian adjustment' post]." I don't think I've ever seen someone advance so much as the hypothesis that "GiveWell might just not have done the amount of work rationally necessary to buy into MIRI's claims. It isn't necessarily being irrational even given full comprehension of the Sequences."

When I've said to people that giving to AMF can be consistent with accepting Astronomical Waste considerations, they've pretty consistently boggled at me. So that's why it hadn't previously occurred to me that people like you two might actually understand the points I've made about Bayesian adjustment and flow-through effects, and the underlying common-sense position that one shouldn't give based on sufficiently uncertain arguments regardless of how big the numbers in them are.

Eliezer: You wrote that "When I've said to people that giving to AMF can be consistent with accepting Astronomical Waste considerations, they've pretty consistently boggled at me."

I still do boggle at that, unless by "can be consistent" you mean "could theoretically possibly be consistent even though it's not" or "saying the words 'astronomical stakes' does not of itself refute AMF".

It still seems to me wildly unreasonable to think that someone seeking the maximum of short-term well-studied lives-saved-per-dollar would happen upon the maximum of

marginal-astronomical-quality-or-probability purchased per dollar. It is entirely reasonable to think that this maximum is not MIRI. It is reasonable to think that this maximum is not MIRI or CFAR or FHI. It is reasonable to think that this maximum is not at a charity which has anything explicitly to do with x-risk. It is reasonable to think that this maximum is at a charity which has never heard of x-risk. It is *not* reasonable to think that this charity is AMF, and so I think the basic critique of "Holden, who isn't really sure that astronomical stakes are even worth significantly more than the present-day world, is making recommendations that clearly are not galactic-optimization-driven" holds. It doesn't hold as strongly nor as obviously as it would hold if galactic optimization necessarily meant explicit x-risk, but it still holds.

Holden: But "someone seeking the maximum of short-term well-studied lives-saved-per-dollar would happen upon the maximum of marginal-astronomical-quality-or-probability purchased per dollar" is not my claim. (Though I think that claim is not as unreasonable as you think.)

My claim is that someone who knows a lot about the former and not about the latter - because they've studied the former and not the latter - ought to give to the former and not the latter until and unless they develop better understanding of the latter.

If a perfectly good dentist who knew nothing about Africa told me he wanted to take next week and do the work that would most benefit Africa (putting aside "earning to give" considerations), I'd likely tell him to spend that week working as a dentist.

A similar situation holds for a full-time financial analyst who wants to give $10k this December.

The disagreement, I think, is whether reading the Sequences puts this person in position to rationally believe they can give competently regarding x-risk.

Eliezer: That seems way more reasonable — not necessarily true but vastly more reasonable — but can you please put some sort of effort into being very very clear about the distinction between the two claims? Despite our previous conversation being not too long ago, that intended distinction still wasn't clear to me without benefit of hindsight while reading your previous message.

Holden: This seems like a really hard thing to be clear about.

I will put a bit of thought into how I can do so habitually. I think the higher-return action, though, is to write a post re: my thoughts on the far future that includes this, and then you can quote from it.

Eliezer: I am not without sympathy. Many things are really hard to be clear about.

In fact, I suspect that while it's not the only effect in play, a lot of the difference in MIRI and GiveWell audiences may come from Givewell's main message depending mainly on points which are vastly easier to communicate, understand, and simplify. Not easy in any absolute sense, but far easier than MIRI content.

It also does seem increasingly to me that you might be more sympathetic to direct-from-source Yudkowskian content than to all the folks who've tried to explain their revised versions to you. E.g. my current writings on Pascal's Mugging, including the post which originated the phrase:

- Pascal's Mugging (original post, about a potential AI problem I called "Pascal's Mugging")
- Being Half-Rational About Pascal's Wager is Even Worse (after people started using "Pascal's Mugging" to mean something wholly different, I offered this historical example of Szilard vs. Fermi and how I think Rabi-quality face-value reasoning about risks is robust without panicking over small probabilities or wielding other fallacies to drive the apparent probability down)
- Pascal's Muggle (my best current proposal for an underlying epistemology which refutes Pascal's Mugging, and my demonstration that this has few or no implications for reasoning about modern x-risk one way or another)

I suspect that reading through this, you would find yourself thinking that the author sounds more like Holden than like some of the other LWers you've spoken to. I shall quite understand if you don't have the time, but if you're thinking of writing anything yourself about small probabilities it might save significant effort.

Holden: I've read all of those pieces; I've always thought that your epistemology is more similar to mine than your fans' epistemologies, though I haven't been confident of this. All of my writings on epistemology have been directed at the people I've had dialogue with earlier, not at you specifically.

Luke: I agree that all three of us seem to be having some difficulty being clear about our views, since we are constantly misinterpreting each other and then needing to be corrected. I do think this stuff is just hard to be clear about, though it gets easier as people write down well-thought-out frameworks and vocabularies for talking about the issues, e.g. *Superintelligence* for long-term AI futures, *Knowledge in a Social World* for social epistemology, etc.

Eliezer: I'll make a slight subject change to the problem of uncertainty-induced inaction.

Here's a possibly relevant example of a case where I was rereading a comment of mine from 2009 on LW, in an unusual beliefs/predictions thread, and noticed myself being clearly wrong in empirical hindsight:

Eliezer_2009:

> I have a suspicion that the best economic plans developed by economists will have no effect or negative effect, because the ability of macroeconomics to describe what happens when we push on the economy is simply not good enough to let the government deliberately manipulate the economy in any positive way.

In response to a further comment, Eliezer_2009 clarified:

> I was including the Federal Reserve Board in "economists". Forgive me if that was a

> mistake. Let me be more concrete: I suspect that the Obama stimulus plan won't accomplish anything positive, not because of any particular flaw I could name, but because the models they are using to organize their understanding of macroeconomics are just wrong - somehow or other. The amount of chaos here seems so great - so many things going differently than predicted, so many plans failing to have their intended constructive effect - that I suspect a chaotic inversion: it's not chaos, we're just stupid.

Eliezer_2014 (me) has spent a lot more time reading econblogs, following the links from smart-seeming people talking about other people who seem smart, and has at least temporarily settled on a widely-praised but ultimately non-mainstream viewpoint, the NGDP-centric viewpoint of market monetarism, which seems to me to (a) have the simplest convincing story about what has happened during many different economic episodes and (b) to have made pretty good predictions for as long as I've been reading the blogs, e.g. with respect to Japan.

Eliezer_2014 does not think that mainstream economics as currently practiced is particularly optimal.

Eliezer_2014, with the benefit of economic theories which only gained significant popularity after 2009, can now argue in much greater detail why the Obama stimulus plan was the wrong move and the correct move would have been for the Federal Reserve to commit to an NGDP level target for the next several years with shortfalls or excesses made up in future years in order to keep to the level path, and how this move would also have obviated any need to bail out the Too-Big-To-Fail entities.

But.

If the folks at the White House and the Federal Reserve had listened to the epistemological despair of Eliezer_2009 and thrown away their then-intended policy responses, then according to the best economic models I know, and also mainstream economics as well, we would currently be in the middle of the next Great Depression.

Now of course Eliezer_2009 was not really *advocating* that the White House and Federal Reserve give up, he was trying his best to make a mere prediction about the effect of policies already adopted, which he thought would be "not much". I think I would have *advised* the White House and Federal Reserve to question the economics theories which had led them into trouble and to look for alternative means of control... and yes, if I'd actually been in charge, I do think I'd have thought to ask mainstream economists how the mainstream models worked...

But nonetheless the fact was that Eliezer, in his wise doubt, in his perfectly-reasonable sounding and well-motivated doubt about an academic field and a bureaucracy which had to all appearances just failed to control its optimization target quite spectacularly, did nonetheless end up saying "I think these things might not have as much effect as their wielders think" when the true answer was that the Federal Reserve did have it within its power to turn a Great Depression into a Great Recession, and the Fed could (thinks Eliezer_2014) have significantly moderated the Great Recession just by *doing more of what they were doing, earlier and faster*. Mainstream economics was less wrong than Eliezer_2009 was guessing, and the up-and-coming successor to

mainstream economics did not say that the Federal Reserve had less power than widely thought, but rather asserted the opposite. (It did happen to say that stimulus would be pointless if the Fed did its job well enough, though this was also an older critique.)

The general point is that despite having what seemed like good reasons to distrust the professional economists and maybe even adopt an attitude of general epistemic skepticism pending detailed investigation, and even though many economic theories in use were in fact wrong, the reality was that what the Federal Reserve and government thought they were doing, actually had effects roughly in the direction that mainstream economics said they should be. And while it would have been possible to do better with better economic theory, it also would have been possible to do far worse by thinking you couldn't do much and all your theories were too wrong to use.

So as not to repeat that mistake again, I shall be more cautious in the future about claiming that some discipline is so hopeless that doing nothing might well be as effective as doing what that discipline tells us to do, unless I can point to specific known bad effects. E.g. multivitamins have been experimentally found not to help much on average, but this is because the typical multivitamin includes genuinely helpful micronutrients plus a few ingredients (e.g. inorganic manganese) which are quite harmful. The Federal Reserve was not in a known situation like this, and so in the harsh light of 2014, it seems that Eliezer_2009 really should not have shrugged and thought that maybe lowering the target rate would have no effect on aggregate demand, though of course Eliezer_2009 didn't think in those terms.

I find it annoying when people argue from their own past errors to claim that I must be making the same mistake as they. And perhaps you are not making the same mistake as I did then. But I do think that it seems suspiciously similar to argue that trying to do FAI research is equally likely to produce an equal quantity of antiFAI as FAI, and that we know so little about x-risk that our current actions are likely to have little effect on it. In other words, I think this kind of epistemic despair really *doesn't* work in real life. I was surprised to have found my 2009 self wielding it (I must not have encountered that fallacy applied to MIRI in 2009, hence not generalized its fallaciousness); and I was grimly amused to see the unambiguity of its hindsight-informed misguidedness.

Holden: I still feel like you are leaving the "observer" parameter out of your rational belief formation/action-taking function. Different things can be rational for you given your evidence vs. me given my evidence. (I'd like to pre-emptively state that I'm aware of the Aumann agreement stuff.)

I am not claiming: "trying to do FAI research is equally likely to produce an equal quantity of antiFAI as FAI, and that we know so little about x-risk that our current actions are likely to have little effect on it."

I am claiming: "I, Holden, do not know enough about the FAI situation to rationally use resources for it that I could use to make progress on something I understand better. The same applies to many MIRI fans whose main information source is the materials MIRI has produced." (This holds in the context of this year's donation, taking it as given that I'm going to make one.

This statement is actually false on other fronts; I can and should be learning more about AI and x-risk.)

I would not claim: "The Fed should consider its actions hopeless and do nothing." The Fed is the best-positioned institution to know what to do and to do it! I would instead claim: "Given Eliezer's 2009 state of knowledge, if he were offered control of the Fed by fiat, he should simply turn it back over to Bernanke."

My "broad market efficiency" function includes the observer as a parameter. Markets are usually ~perfectly efficient from the perspective of the ignorant, but often ~0% efficient from the perspective of the world's leading expert.

Eliezer: Ah. Well, if that's your argument, I definitely get to say that you have fans who have misunderstood you and have argued the misunderstood version with me in the name of GW-style EA. Which of course doesn't count for anything, and I shall endeavor to keep in mind your correct version henceforth.

I use Aumann on rare occasions with rare friends (not average friends), and consider the argument from Aumann a valid proof that "Something somewhere on this planet is going wrong relative to ideal rationality" but terrible advice for "You should behave like this and pretend that other people are behaving like this, in order to correct this flaw" so I certainly wouldn't try to use it to bludgeon you into anything. Did someone else try? That's another fallacy I can document having spoken out against since at least 2006.

It sounds to me like a plausible description of our current differences would be that they stem from:

1) Eliezer thinks that AMF is obviously "not it" from an astronomical-stakes perspective, that someone trying to optimize for astronomical stakes ought to continue looking, and this continuation should definitely not stop short of something which is not-obviously-not-it, because multiple things with that property sure look findable to him. Holden thinks that the current maximum of modern fires which can be cheaply put out using modern water, is with great plausibility that thing which an underinformed stockbroker ought to donate to this year if their sole goal is to ensure a happy future for the galaxies.

2) Eliezer thinks that relatively simple arguments about AGI, taken at face value, promote FAI work into the class of things that is plausibly-it for astronomical optimization; Eliezer also thinks it's basically epistemologically okay and nonsinful to go on taking these arguments at face value as successively deeper layers of argument and counterargument are examined. I'm not exactly sure how to characterize the Holden position except as "These things all seem really iffy to me", which may in turn be a position that has something to do with the dichotomy in (3).

3) Eliezer also thinks it's sensible to give money to an organization which is not super sleek if they're literally the only ones trying to do the most important thingy. Some of Eliezer's meta-heuristics revolve around questions like, "Does this proposed policy mean that we inevitably lose due to its sage advice, in the event that Reality has in fact handed us a real

x-risk?" and trying to figure out how smart people should behave in a world where real x-risks are a thing and somebody has to handle them. Holden doesn't regard MIRI as being especially privileged as an astronomical-optimization charity compared to AMF, so he thinks it's relevant to compare MIRI to AMF along dimensions like knowability, solidity, administration, etc. Since MIRI and AMF and many other charities are all, to Holden, playing on a level field of 'trying to do important things', it seems very odd to Holden that MIRI would claim to be knowably-to-stockbrokers better than AMF based on MIRI's grade of evidence. Another way of phrasing this dichotomy is that Holden sees a well-populated field of charities that are plausibly in the running as candidates, whereas from Eliezer's perspective it is obvious that the planet is on fire for 2-ish reasons and then only a tiny handful of charities are even trying to put out the fire, nor is it the kind of fire AMF is likely to put out accidentally without explicitly trying to do so.

Holden: No one tried to bludgeon me with Aumann. I was confused as to why you seemed to be skipping over the "available evidence to the agent" parameter in the "rationality of agent" function, so that was my pre-emptive guess. Now I think there was nothing special going on, I just wasn't being clear.

Apparent disagreement (1) sounds accurate. The "sure look findable" part is the key disagreement. It seems hard to argue with my take on this when talking about a finance or software professional who has 2 hours a year to decide where to donate (many of our fans), or GiveWell_2009. It's more contentious in the case of "semi-engaged fan of both GW and LW who has read 50% of the Sequences," or "where Holden should have donated in 2013."

Apparent disagreement (2) also sounds accurate. Let me try to clarify my position.

Presumably we agree that there are many possible get-rich-quick schemes that would initially sound right on the object level and would be very hard to find the object-level problems with, and that one should reject anyway. The Sequences' claims have much in common with a get-rich-quick scheme: their recommended action is (to an altruistic person) incredibly exciting (the literal definition of "incredibly" has relevance here) and also has direct benefits to the advertiser. The advertiser seems more likely to be an eloquent rationalizer and self-deluder with a hero complex than a calculating and self-conscious thief, but the former could easily be functionally equivalent to the latter.

When I read a description of a good-sounding get-rich-quick scheme, I don't need to find the object level problem to be skeptical. I can simply apply outside views like "smart people think get-rich-quick schemes are usually nonsense, and I don't know anyone smart who endorses this one." If I study deeply and broadly enough I may actually discover that the get-rich-quick scheme works, but for a mild to moderate amount of study the former belief is more rational.

I think a big difference here comes down to who counts as "smart people," or more broadly, whom one trusts. I think my disagreement with many MIRI fans is less likely to be about how good the object-level arguments in the Sequences are, and more likely to be about who counts as a "rational person that we should take seriously," because I think they are reasoning, "All the smartest people I know — such as LW person 1, LW person 2, etc. — agree with this stuff; it's good" whereas I am reasoning, "Most of the people who agree with this stuff seem unreasonable

to me in various ways; I will tread carefully." These differences are self-reinforcing to a degree, because my straw man friend has then spent a lot of time with the LW community whereas I have purposefully limited my interactions, leading to stronger bonds forming for the former. However, I have consistently decided I should spend *some* time with the LW community and certainly take it more seriously now than I used to (though I'm sure still less seriously than you guys do).

Then there's the expert endorsement thing. This is one of the best heuristics for evaluating lengthy technical arguments that you don't have time to obsess over sufficiently. MIRI has never had the sort of impressive endorsements I think it ought to have if it's right, and this goes beyond absence of evidence to evidence of absence given how long MIRI has been around, how much notoriety it has, and how novel and important-if-true its claims are. Then, once I started meeting impressive technical people via GiveWell, they thought MIRI's object-level arguments were wrong too. Add lack of concrete achievements (again, prima facie striking for someone believing themselves to be unusually insightful about AI) and general lack of organizational strategy (the latter was more relevant in 2009 and is less relevant now) to the cluster.

This may have seemed like a tangent but I don't think it really is one. Bottom line, it probably seems rational to you for someone to agree with you based on reading the Sequences, but I think that's because you know your arguments too well and can't simulate beginner's mind. I think people who come to agree with you *just* based on that really are more likely to be doing so as a result of a mistake. OTOH, I think it is perfectly reasonable to become *intrigued* by you just based on the Sequences and to do more investigation, and I concede that some such avenues of investigation *may* lead to rational agreement.

Moving on…

I don't disagree with "it's sensible to give money to an organization which is not super sleek if they're literally the only ones trying to do the most important thingy."

And sure, AMF is unlikely to put out your 2-ish fires, but "it is obvious that the planet is on fire for 2-ish reasons and then only a tiny handful of charities are even trying to put out the fire" is quite a claim. I don't think reading the Sequences alone ought necessarily to convince a rational person of it, for reasons outlined above.

One more thing. Luke, in your shoes I would be strategizing differently about the role of MIRI. My goal would be "Increase interest in AI safety among AI experts via deep direct engagement with said experts" rather than "Work on decision theory in order to draw in mathematicians."

Luke: If we *combine* those two goals, that's roughly the strategy I consider myself to be pursuing these days. Except I'm starting with "get AI *safety* people to care about long-term AI safety via deep direct engagement with them" (G1) rather than "get non-safety-interested AI people to care about long-term AI safety" (G2). I think if we can get AI safety people to care about long-term AI safety, they'll be better able to accomplish G2 than MIRI+FHI can alone.

Also, I usually find it more productive to talk to AI safety people than 'normal' AI researchers —

not surprisingly, AI safety people are more likely to have accurate intuitions about AI safety. E.g. AI safety people grok that strong safety guarantees are super hard for complex systems, that you don't get desirable constraints on behavior by default, that it's easier to design a system from the ground up with safety in mind than to design a system with other optimization criteria and then slap safety onto it, that translating intuitively desirable behavior constraints into formal requirements is extremely hard, etc. When I talk to someone who's just really good at machine learning or whatever and they don't interact with AI safety issues at all, they often don't have these intuitions, because those considerations aren't relevant to their machine learning work.

But in order to *show* AI safety people concrete examples of what long-term AI safety could look like and why one would do it, I need mathematicians to make novel progress on that issue. Also see these reasons for recruiting mathematicians.

Holden: Interesting, I wasn't aware that you're talking to AI safety people. Also, I'm open to the idea that there's a class of people better than the class of people that seems most relevant to me; I just don't think that's mathematicians.

Luke: BTW, "mathematicians" is our shorthand for "people who can invent new math quickly, whatever field they're in." E.g. our workshops have been attended by a roughly-equal portion of mathematicians, computer scientists, and formal philosophers, with the latter slightly less common than the other two.

Eliezer: Holden, let me jump back to this paragraph you wrote:

> The "sure look findable" part is the key disagreement. It seems hard to argue with my take on this when talking about some random finance or software person who has 2 hours a year to decide where to donate (many of our fans), or GiveWell_2009. It's more contentious in the case of "semi-engaged fan of both GW and LW who has read 50% of the Sequences," or "where Holden should have donated in 2013."

Actually, a lot of my worries about the cognitive style of Holden Karnofsky's fans derives from the point that if I'd never heard of MIRI it would still feel obvious to me that the key lay somewhere other than AMF. Maybe I'd be pushing on nanotech. If I'd never heard of x-risk, maybe I'd be advocating for funding for Bussard's Polywell. If near-term global poverty were my actual values priority, I might be pushing on Romer's special economic development zones.

Now there's also this other problem where the best things may sound weird to everyone else, or be accused of being "stuff white people like" or whatever... but of course this is the central problem of deploying unusual sanity into altruism; the important things with the most headroom for funding will occupy peaks of utility but not have the cognitive properties required to attract funding. In some cases the peaks of utility themselves will scare people off.

And I think there's a level on which you appreciate something like this and chose near-term global poverty charities to be photogenic, and that you always intended to branch out into mildly less-photogenic-to-mundanes things which don't scare you off. But I also think that our background visualizations of what constitutes The Problem, on the most general level, are pretty

different.

Now, in response to this paragraph:

> Then there's the expert endorsement thing. This is one of the best heuristics for evaluating lengthy technical arguments that you don't have time to obsess over sufficiently. MIRI has never had the sort of impressive endorsements I think it ought to have if it's right, and this goes beyond absence of evidence to evidence of absence given how long MIRI has been around, how much notoriety it has, and how novel and important-if-true its claims are. Then, once I started meeting impressive technical people via GiveWell, they thought MIRI's object-level arguments were wrong too. Add lack of concrete achievements (again, prima facie striking for someone believing themselves to be unusually insightful about AI) and general lack of organizational strategy (the latter was more relevant in 2009 and is less relevant now) to the cluster.

It seems to me very important to distinguish the following three theses:

A) There is a big thingy approaching; an intelligence explosion which must be gotten right and which will kill us if we don't.

B) The world is on fire in the sense that current visible mechanisms are not on track to handle (A) correctly.

C) MIRI is relevant to (B).

As far as I can tell, the "impressive people" you spoke to about MIRI's object-level arguments have argued mainly against (B), and as far as I can tell this is mainly because they don't understand details of why FAI is hard and think that a relatively straightforward effort will suffice for it.

I get the impression that you lack personal experience with modern academia qua academia. Is this correct? A lower opinion of modern academia would probably change a lot your estimate of how important it is that we don't have academic endorsements relative to the importance of e.g. the content of the Robust Cooperation paper, which nobody has really gotten around to trying to publish anywhere.

Also, the two years of blogging which I assume you have not read and would not ask you to read, do make a fairly detailed case for (A) and (B) and may also serve to legitimately convince reasonable people beyond a reasonable doubt that I am, at worst, a reasonable person who made reasonable mistakes, rather than a get-rich-quicker. This should not be an unlikely claim if you do not know what is inside the two years of blogging.

In response to:

> I don't think reading the Sequences alone ought to convince a rational person of it ...

But do you know what's in there? Sure, you don't have time to read it, but it seems a lot less "robust" to claim that you know what other people ought to conclude from reading it. Many steps of the big inference such as "No other visible party is even trying to handle this problem" are immediately observable. Others such as "You can have a rational agent which only tries to make paperclips", despite being argued a lot online, would be widely agreed with by analytic philosophers. Others such as "intelligence explosion / FOOM" are taken seriously by many people and have been for decades. Your personal knockdown objection of tool AI is not taken seriously by anyone but you. Putting a lot of strong pieces together to make a new image is not something that smart people ought to run away from.

A hypothesis I have formed about you is that you put way too much trust in your own absurdity heuristic and commit some version of the typical mind fallacy in thinking that everyone else should find the same things absurd, and those who don't must be dysfunctional. I find AMF absurd as effective altruism, but I don't think anywhere near as badly as you for being involved in it as you appear to think badly of anyone who has read two years of blogging you haven't and concluded that the planet has been legitimately established to be on fire with >50% probability.

Holden: I would estimate that I've read 80% of the Sequences. Most of my memories of the Sequences are of thinking "Yeah, that sounds right. It's a more articulate version of something I basically already believed" and sometimes "It describes a way in which people can be irrational that I don't particularly think I or Eliezer is less prone to than people who haven't had this thought." Most of it is about rationality, not anything about the current empirical state of the world and whether it's on fire. So anyway, all I'm saying is that I don't think that what I'm missing comes down to some theoretical point (like could a rational agent theoretically want to maximize paperclips while still being rational - I say yes) that early in the inferential chain.

I disagree with this: "And I think there's a level on which you appreciate something like this and chose near-term global poverty charities to be photogenic, and that you always intended to branch out into mildly less-photogenic-to-mundanes things which don't scare you off." Especially the last part. MIRI is not even *close* to the weirdness threshold I would be uncomfortable with if I believed its arguments to be correct ... not even in the ballpark. I think it is *highly* likely that I will end up pushing MIRI *or something at least as weird*. And the more I learn, the less weight I put on models like expert opinion and conventional wisdom, so just because I'm putting weight on them now doesn't mean they form a solid barrier to my becoming convinced over the long run, if you engage with me enough and if you're right.

Re: academia. I think much of academia is dysfunctional, many academics are not people whose views I would put weight on, and publication of that particular paper would have zero impact of my views on it. But I think there are some very smart people in academia that you ought to have gotten on your side by now if your arguments are right.

I think one basic disagreement is how much of this problem you've defined robustly and how much is in the category of "smart people with more knowledge than us will redefine the problem for us once they care about it."

It's one thing to do things that make sense conditional on (A) "AGI might be here in 50 years."

I'd agree with Luke that we should accept (A).

It's another to do things that make sense conditional on (B) "AGI might be here in 50 years and will be a world-shattering development when it comes." I actually think that is far more suspect from an outside-view perspective, thinking about how predictions like that have turned out in the past. I also think there are inside-view arguments that this is a bit shaky. However, I think it makes sense for MIRI to go ahead and assume this because it's good for organizations to specialize in exploring potentially important theories, even shaky ones, and to raise the profile of such possibilities. I also see a high probability that I'll end up thinking this is a good hypothesis to engage in philanthropy around.

It's a third thing to do things that make sense conditional on (C), a very particular view of what the precise risks will be and what the precise skills needed to address them will be. This is where I think you guys are on shakiest ground (and where the outside-view track record looks worst). Given that the broader (B) is still far from mainstream, I think you guys should be more focused on (B) and less on (C).

I think one of our biggest disagreements comes down to broad-market-efficiency-related stuff. I've repeatedly had the experience of thinking, "OMG an awesome idea that no one is working on!" and then gradually refining the idea, talking to more people about it, and eventually discovering that in fact the *smarter version* of my idea (which is described in a completely different language from my original idea, thus making it hard to have found) is being worked on by what seem like the logical people to be working on it — people who are smart, who have thought about it a lot, and who are well-positioned to work on it. Now that doesn't mean the problem is "taken care of." In many cases there is a small set of good people working on an issue, shouting into a world that is ignoring them. In many cases it may be true that *if* I were to become fully engaged with the topic, I would have many insights that the current crew doesn't have. However, what does seem true upon this finding is that it would have been ill-advised to try to "exploit" my original idea by starting a nonprofit or something.

This is a version of "broad market efficiency" that doesn't claim anything like "resources correlate strongly with problem importance" but does claim something like "You can't just come up with a problem, superficially observe there's no one working on it, and conclude that there's a hole you should fill."

Examples of where I've been through this dynamic: results-based aid, results-based giving (note on both of these: it's true that the specific thing GiveWell was trying to do turned out to be unique, but I originally thought we were more broadly unique), meta-research related to reproducibility and data sharing and preregistration, asteroid risk (originally claimed by LWers to be an actual promising area as opposed to a toy example), identifying distortions in the relationship between the diseases the NIH devotes the most funding to and the diseases with the highest burdens ...

For a long time, my picture of MIRI was that it was a set of people who had formulated a problem in their heads and promptly gotten to work on their badly formulated version of the problem, not bothering to engage with subject-matter experts who could help them understand

the better formulation and find the people already working on the better-formulated version, and instead attracting an audience through posts about rationality and (later) HPMoR. That's the kind of thing that you can't falsify by reading the Sequences.

I've since changed my view substantially though not fully. I see that Luke is trying hard to engage relevant experts (though this wasn't true at the time); I recognize that you have more technical talent and intelligence in your camp than I had thought (the Sequences are largely exposition and philosophy, and don't seem to me to demonstrate this); and I now believe that *if* there is a "set of relevant experts you haven't found yet" it's because they're extremely hard to find. That said, my object-level understanding (and my most-trusted technical friends) still tells me you are formulating the problem badly; also, I'd bet that there *are* "relevant experts" that are extremely hard to find, and that they're at DARPA/DoD and/or Google, the two US orgs that would seem to have the highest odds of producing AGI. (If true, that still doesn't call for inaction on your part.)

Re:

> Many steps of the big inference such as "No other visible party is even trying to handle this problem" are immediately observable. Others such as "You can have a rational agent which only tries to make paperclips", despite being argued a lot online, would be widely agreed with by analytic philosophers. Others such as "intelligence explosion / FOOM" are taken seriously by many people and have been for decades. Your personal knockdown objection of tool AI is not taken seriously by anyone but you. Putting a lot of strong pieces together to make a new image is not something that smart people ought to run away from.

I've explained why the "No other visible party" thing is not, in my view, immediately observable. Moreover, the Sequences don't document expert support on these issues, they merely argue them at the object level. That's the whole thing I'm complaining about. I still don't know where expert opinion stands on the foom idea. The "can have a rational agent ..." thing has never been a point of disagreement.

As an aside, I think many take the "tool AI" idea seriously; if what you mean is that they don't take it seriously *as a knockdown objection*, that's true and I don't either (and never did), but took it as an example of how you were doing a bad job formulating the problem, and took the arguments I had on it with LWers as evidence that people were not thinking intelligently about these things. For an example of this see the comment discussions on Tool-AI in my post. Your response made technical claims that still seem unsupported to me but at least it made sense: it didn't flat-out confuse the concept of tool-AI with Oracle AI or deny a conceptual distinction. Anyway, we've been over this - your supporters are not you, but they're who I had access to at the time.

Regarding:

> Actually, a lot of my worries about the cognitive style of Holden Karnofsky's fans derives from the point that if I'd never heard of MIRI it would still feel obvious to me that the key

> lay somewhere other than AMF. Maybe I'd be pushing on nanotech. If I'd never heard of x-risk, maybe I'd be advocating for funding for Bussard's Polywell. If near-term global poverty were my actual values priority though this is hard to imagine, I might be pushing on Romer's special economic development zones ... A hypothesis I have formed about you is that you put way too much trust in your own absurdity heuristic and commit some version of the typical mind fallacy in thinking that everyone else should find the same things absurd, and those who don't must be dysfunctional.

I just think that the more exotic and less verifiable something is, the higher the chance that upon coming to deeply understand it I'll see why it's a bad idea or why it's already being worked on. Of course there *are* really important ideas that are neither, but most random stuff I hear of like "economic development zones" is going to disintegrate if I look into it harder, and the straightest and shortest path to something that holds up is to explore things that can be explored in straight and short paths. Now we have the resources and knowhow to explore longer and more winding paths.

I do expect the best giving opportunity to look more like MIRI or Charter Cities than like AMF. But I don't know *which* MIRI/Charter-Cities like thing it is, and I think AMF beats a random guess.

I think you have a poor model of how my mind works, and I am trying to help you understand it better, because your arguments are well within the "I could imagine becoming convinced of this" space. If I could ask one thing of you, it would be as follows: lower your confidence that you understand how I am reasoning, and express this to people when asked. Instead of saying "Holden likes AMF because [something]," say "Holden likes AMF, and I honestly don't know why. I have some theories, but before I present them to you, please understand that I am genuinely confused by Holden's mental model; many of my theories of him have either turned out to be wrong or are asserted explicitly to be wrong by him at the moment. So here's a model of Holden, but don't take it to the bank."

Also, I should mention that I don't think I ever proactively tried to talk someone out of donating to MIRI. That would have been a bad thing to do for multiple reasons. What I did do was defend the idea of donating to AMF when people asked "You can't be serious? Isn't it clear from the Sequences that the world is on fire and we should donate to MIRI?" I would probably stand by most of the things I said in response to that. The two are importantly different, in my view.

To be clear, my guess now is that you guys are onto something important that calls for at least some sort of action. I want to check things out more, but that's my guess. I definitely do not remotely believe my comment about DARPA and Google should be taken as an argument that "it's all good." It was more of a side comment. I don't think it's definitely right, or that even if it were right it would be much of an argument against taking further action.

Eliezer: I don't think I've ever tried to talk someone out of donating to Givewell. Though my memory is not perfect. But I do feel like I would have heuristics that would stop me from doing that. I have previously had conversations along the lines of someone else spontaneously bringing up the subject in the form of, "But if this is really a good idea, why doesn't Givewell endorse it?"

to which I previously replied with "Givewell endorses AMF, they're not playing on the same field."

When I read that about "LWers" talking about "asteroid risk" then lo I nearly banged my head against the wall while internally shouting, "Asteroid risk? Who the hell has been talking about asteroid risk?" I really don't think your image of LWers is at all like my image of LWers.

Luke: Regarding:

> I do expect the best giving opportunity to look more like MIRI or Charter Cities than like AMF. But I don't know *which* MIRI/Charter-Cities like thing it is, and I think AMF beats a random guess.

Yeah, many of our differences might just come down to differences of knowledge. The more we talk, the more I think I'll be surprised if, after 1000+ hours of investigation of GCRs, you end up with a radically different picture of basic-GCR-landscape than we do. Like, my confidence is increasing over time that you'll end up thinking (like we do) that asteroids & climate change & nuclear security just aren't the big looming problems that biosecurity and AI risk are (and maybe nano, but we get to see that one coming to a greater extent).