Phone conversation, 5/10/11 - paraphrased transcript, edited by both Holden Karnofsky and Jaan Tallinn

Holden Karnofsky: The first question concerns the output of SIAI. Do you feel that SIAI has produced specific papers and/or tools that are likely to be useful in ensuring that artificial intelligence is friendly?

Jaan Tallinn: My knowledge of SIAI's research output is limited. In my view, the most important contribution, by far, is pointing out the problem of Friendly AI in the first place. At least for me, SIAI was the first group to bring this problem to my attention. Based on my conversations with AI developers, they aren't necessarily aware of these issues.

As Moshe Looks – the lead AGI developer at Google – said at the Winter Intelligence Conference in Oxford: "If you are an AI developer out there, you *really* should listen to what Eliezer Yudkowsky says, even if it's annoying to admit that he's right!"

I also think that Singularity Summit is valuable because it serves as a venue for discussing and disseminating these ideas.

In terms of research contributions, I would like to see more insights and tools that would make AGI development safer. For example, developing the idea of "tripwires" – automated mechanisms that are designed to remain under the radar of an AGI and shut it down if its behavior deviates from expected norms.

Of course it's hard to assess the exact value of such tools today as there remains an unknown gap between now and where this work might be needed.

Holden Karnofsky: That's exactly my issue though. It seems a bit like trying to design Facebook before the Internet was in use, or even before the computer existed.

Jaan Tallinn: True, but I'm not approaching this as an ordinary technological challenge aimed at contributing to the economy. I estimate the probability of us dying from an AGI disaster in double digits – which gives me a different mindset. It's like having a terminal disease, where you're simply trying to save yourself, not so much improve your current life. So even if it turns out that now is not the ideal time to be building such tools, I think the alternative of sitting and doing nothing is much worse.

Holden Karnofsky: Would you rather see SIAI working on raising awareness of the problem or on developing tools to address the problem?

Jaan Tallinn: Clearly tools are more valuable than awareness.

Holden Karnofsky: But the question is whether this team, at this time, is capable of building tools, or whether they should be focusing on awareness.

Jaan Tallinn: I'd like to see them trying to build tools. One of your questions in your interview with SIAI was how they could use money to scale activities; to me an obvious thing to do is to think about how to scale the search for the answers, if there were more funding. It does seem like they have enough funding for what they're doing now.

Holden Karnofsky: OK. The next line of questions has to do with the probability of a "hard takeoff" playing out as SIAI envisions, and as it would need to in order to make SIAI's work relevant.

One thing I think is that for any given domain of intelligence, I'd expect a human-computer team to surpass human ability before a computer does so on its own. This includes the domain of understanding intelligence itself - I think there's a good chance we'll develop new tools for thinking about intelligence and AGI design that will render the work of SIAI, done with today's tools, moot before we actually have the AGI itself. This also includes domains relevant to developing super-powerful technologies; I think there's a good chance some human, working with a computer, will gain control over a super-powerful technology, which could make SIAI's work moot in a different way.

Jaan Tallinn: Certainly this is possible. The question is how that possibility factors into our probability estimate of a hard takeoff. Myself I think that we're still left with a probability in double digits this century. If you believe differently, the proper thing for us to do would be to sit down for a day or so, go over all the evidence and reasoning that influences our estimates, and see if they could be brought closer together.

Dario Amodei: Another possibility is that Google or Facebook or someone in that area, over the next few years, comes up with insights that change our opinion on the nature of intelligent systems. It seems like there's a huge amount of research into AI in a broad sense, and I wonder if the things SIAI is working on are going to turn out to be relevant in the face of all that.

Jaan Tallinn: I don't believe such companies are likely to produce this sort of thing, with the possible exception of Google which actually is working on AGI. Myself I would look to neuroscientists and AGI startups - they absolutely could find things that would render the work of SIAI moot, but still, the possibility that SIAI's work is not moot is very real, and very important given the stakes. To me it's as though we're trying to defuse a bomb.

Dario Amodei: I hear the argument about wanting to be conservative about dangerous possibilities. My concern is more over whether the whole "defusing a bomb" framework is completely off. I hear SIAI's concepts of "friendly AI" and "unfriendly AI," and to me this whole framework is sufficiently vague that the leading possibility in my mind is that these concepts just won't apply at all to what actually happens.

I think the image of trying to defuse a bomb assumes too much concreteness and understanding of the situation. The way I think about it, there may or may not be a bomb, or maybe there is one but it looks very different from what we think it looks like, maybe the actions we associate with defusing the bomb are the ones most likely to set off the

bomb.

Holden Karnofsky: The way I think about it, it's as though there is a bomb, but none of us knows anything about bombs, and there might or might not be a bomb squad on the way … there's such a thing as having little enough understanding of how the bomb works that you're really not helping by trying to come up with ideas that will be helpful to the bomb squad when it comes. And you should direct your energy elsewhere, especially when there are other important problems to contend with as well.

Jaan Tallinn: To me it seems obvious that intelligence is highly explosive, because we witness the disruption that it has caused on our planet in the last 100 thousand years. Hence I find it hard to imagine a situation where we create systems whose intelligence compares to ours as ours compares to that of an insect, and the world/environment would *not* completely change as a result. So the burden of proof shifts to those who say that SIAI's activities *cannot* be useful – I think we should be doing everything we can.

And we should not use the trend of technological progress producing nice things as an excuse for waiting. Nick Bostrom put it really well: "When dumb, smarter is safer… however, when smart, smarter is more dangerous."

Holden Karnofsky: Another question I have for you concerns the nature of the "hard takeoff." It seems like if you could design a computer to calculate what the world will look like in response to different possible actions, you can design it to report this information rather than act on it - what some call "Oracle AI." I see no reason why this should be infeasible.

Jaan Tallinn: The main challenge with Oracle AI is that we must ensure that the "calculating the answer" process will be completely safe (and, for example, not turn the entire planet into computronium in order to produce the answer or test its correctness). More generally, before we let a powerful AGI program loose to do *anything*, we must be sure that it has no detrimental side effects. Just think of the number of species that we have killed as a side effect of our activity! We did not really give them a choice, did we.

And the lack of side effects becomes very hard to ensure if the system can inspect and rewrite its own source code (hence the need for reflectively consistent decision theory – another tool that SIAI is working on).

Holden Karnofsky: That's fair, but it's just one more assumption you have to make to believe that SIAI's work won't be moot. You have to believe not only that we're going to see a "hard takeoff" (i.e., overnight development of a super-powerful, super-intelligent machine) but that this hard takeoff will take a specific programmatic form: a computer rewriting its own source code.

Jaan Tallinn: Indeed, there could be different forms of hard takeoff that SIAI's current work is not addressing. For example, a group of uploaded scientists using their

superhuman abilities to develop their own AGI project in very short timeframe.

Importantly, I think all forms of hard takeoff are dangerous, because they are one-way transitions that we can't course-correct nor reverse. So if there are multitude of ways how they could come about we need to think about *all* of them.

Holden Karnofsky: One thing I've argued to SIAI is that they ought to be producing commercially viable innovations at some point. My reasoning is that it seems to me that if they have unique insights into the problems around AGI, then along the way they ought to be able to develop and publish/market innovations in benign areas, such as speech recognition and language translation programs, which could benefit them greatly both directly (profits) and indirectly (prestige, affiliations) - as well as being a very strong challenge to themselves and goal to hold themselves accountable to, which I think is worth quite a bit in and of itself. Do you think this is fair?

Jaan Tallinn: Myself I care about the final results, not necessarily about commercial byproducts. Going back to my analogy of searching for a cure to terminal disease, saying that SIAI should produce commercial byproducts is equivalent to requiring that the cure (or its byproducts) should also have an application in some other area.

Of course, the tactic of producing saleable products is completely feasible and an obvious one for a commercial AGI startup. In general, I believe the world needs a diversity of research teams and tactics in order to produce the best answers.

Holden Karnofsky: I'm largely struggling for a way to evaluate the SIAI team. Certainly they've written some things I like, but I don't see much in the way of technical credentials or accomplishments of the kind I'd expect from people who are aiming to create useful innovations in the field of artificial intelligence. What do you think?

Jaan Tallinn: I have been programming computers for most of my life, and based on my interactions with SIAI, I can vouch that they are indeed a) very intelligent and b) aware and immersed in the problems that we need to solve before we can develop safe AGI.

Like I said, the world needs many more safety-conscious research groups like SIAI – be they commercial or not. Currently I'm only aware of 2-5 of them, depending on how you count.

Email from Holden Karnofsky, 5/12/11

… I have a couple of questions regarding your edited answer to the question about Oracle A.I.

Your edited answer states, "there could be different forms of hard takeoff that SIAI's current work is not addressing. For example, a group of uploaded scientists using their superhuman abilities to develop their own AGI project in very short timeframe."

In my view, no such elaborate scenario is needed to envision Oracle A.I. appearing overnight. My understanding is that once we figured out how to get a computer to do arithmetic, computers vastly surpassed humans at arithmetic, practically overnight ... doing so didn't involve any rewriting of their own source code, just implementing human-understood calculation procedures faster and more reliably than humans can. Similarly, if we reached a good enough understanding of how to convert data into predictions, we could program this understanding into a computer and it would overnight be far better at predictions than humans - while still not at any point needing to be authorized to rewrite its own source code, make decisions about obtaining "computronium" or do anything else other than plug data into its existing hardware and algorithms and calculate and report the likely consequences of different courses of action. Of course this is only one possible scenario, but to me it is a fairly simple one, and highlights that "Fast takeoff" does not equal "non-Oracle / unsafe takeoff."

Do you think that's fair?

Another question on this section: I believe that on the phone you said that you think it's very important to prevent a hard takeoff if we possibly can, where "hard takeoff" is defined as "takeoff that doesn't give humans time to react." My notes reflect this but your edit does not - it merely says that we should think about all the possible forms of hard takeoff. Was that a purposeful edit or do you think we should put in a statement along the lines of "We need to make sure this happens in such a way that we have time to react."?

Best,
Holden

Email from Jaan Tallinn, 5/12/11

re oracle: ah, i see, you mean a scenario where a system would get really good at predicting after we invent and implement some new kind of prediction algorithm (and that algorithm does *not* involve self improvement, unsupervised learning and other interaction with the environment)? i think this brings 2 interesting and non-trivial questions: 1) how powerful such system can be in the first place (eg, is it powerful enough to program computers better than humans)?, and 2) is such algorithm easier to devise than kicking off a self-improvement cycle, and then sitting back and watching the fireworks?

i don't have immediate answers to those questions -- thanks for proposing a scenario that i hadn't even considered! from the notes perspective, perhaps you should insert a new question to address that and i'll try to figure out an answer?

re removing the "hard takeoff not giving time to react" sentence: i rephrased it to - what i thought was more precise - "hard takeoff[s] are dangerous, because they are one-way transitions that we can’t course-correct nor reverse". if you think the original wording was clearer, perhaps we could put it this way: "Importantly, I think all forms of hard takeoff are dangerous, because they are one-way transitions that we can’t course-correct

nor reverse -- by definition we simply won't have time to react!"

- jaan

Email from Jaan Tallinn, 5/13/11

ah, i figured out the answer!

basically, there are 3 possible scenarios with such "limited oracle":

1. the oracle is not powerful enough to predict what the changes to its own source code and hardware will do -- in which case it's not really more powerful than its creators, so i would not call it a takeoff, much less hard takeoff, scenario;

2. the oracle is powerful enough to predict the effect of its own source code and hardware changes (to a greater degree than its developers!), and proceeds to implement those changes in order to gain even greater power. this is the classical recursive hard takeoff scenario;

3. the oracle is, in principle, powerful enough to come up with self-improvements, but refrains from doing so because there are some protective mechanisms in place that control its resource usage and/or self-reflection abilities. i think devising such mechanisms is indeed one of the possible avenues for safety research that we (eg, organisations such as SIAI) can undertake. however, it is important to note the inherent instability of such system -- once someone (either knowingly or as a result of some bug) connects a trivial "master" program with a measurable goal to the oracle, we have a disaster in our hands. as an example, imagine a master program that repeatedly queries the oracle for best packets to send to the internet in order to minimize the oxygen content of our planet's atmosphere.

- jaan

Email from Holden Karnofsky, 5/13/11

I agree with your analysis, though I still don't why there would need to be special "protective mechanism" in #3. I'm just picturing a program like

```
$prediction = prediction_function ($data, $hypothetical_action);
print $prediction;
```

And if prediction_function is as good an algorithm as what humans implement - but implemented in better hardware and fed more data - this would seem to be an AGI.

I agree that this would be an enormously dangerous entity, because it would help its user get whatever that user wanted. I'd be more worried about a human using it to gain control of the world than about its getting hooked up to a paperclip maximizer type user. In any

case it seems that the state of research on "Friendliness" would be unlikely to be a major factor in the outcome under #3.

What do you think?

Email from Jaan Tallinn, 5/15/11

yes, you're right to question the need for protective mechanisms in the 3rd scenario -- my wording was too ambiguous there. here's an improved one:

3. the oracle is, in principle, powerful enough to come up with self-improvements, but refrains from doing so for whatever reason (either it lacking the "drive" to do so, it being hindered by some special protective mechanisms, etc). however, it is important to note the inherent instability of such system -- it's just a matter of connecting a trivial "master" program to it in order to create a powerful optimizer that can either proceed to start rewriting itself (because the master program asks the predictor for code improvements and proceeds to implement them. in which case we're back to scenario #2), or change the future in any other irreversible way (eg, imagine a master program that repeatedly queries the oracle for best packets to send to the internet in order to minimize the oxygen content of our planet's atmosphere).

i agree that, under scenario #3, it's also possible to have a "non-explosive" oracle that someone would use to take over the world.. however, it's not necessarily trivial to do so without inadvertently destroying the world in the process due to a bug in your master program or ambiguity in your questions that you pose (see http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/).

i also agree that the friendliness research would probably not help much once the world would come to such scenario, but it can help to prevent such scenario in the first place by increasing the probability of an FAI being constructed first. in other words, i would lump scenario #3 together with other existential risk scenarios such as nano- or biotech getting out of control.

in addition i would argue that scenario #2 is more likely than #3, because it's easier to construct a superintelligent program by tapping into the power of its early versions in order to develop subsequent ones.

- jaan

Email from Holden Karnofsky, 5/15/11

Thanks. I think we're having more success learning about each others' views over email than we did over voice ... I'm happy to continue this informally and work on a formal version at the end rather than asking you to create formal responses as we go. What do you think?

I'm familiar with the "hidden complexity of wishes" analysis and think it applies well to an A.I. that is designed to bring about an outcome / maximize some parameter in the world. However, the A.I. I described is simply designed to report. I picture it working something like Google Maps:

- You tell it your goal (for Maps, get from point A to point B; for AI program, become powerful)
- It gives you a set of instructions for reaching your goal, plus helpful tools for visualizing the intermediate and final outcomes of your following those instructions (for Maps I'm talking about the route map)
- You can then impose further constraints and/or modify your original parameters and it will modify its instructions to you (in Maps you can specify that you want to take a certain road or, of course, plug in a different destination if you realize that it's taking you to the wrong place)
- And that's it. Google Maps doesn't drive your car unless you manage to hook your car up as a "master," something that is neither easy nor tempting to any reasonable person.

As to #2 vs. #3, my intuition disagrees with you here.

- For a specialized program operating in an environment whose rules we fully understand, it is possible to simulate the environment and get the program to "learn" very quickly via trial and error. Thus, you could, in theory, build a superhuman chess player without yourself having a deep understanding of how to win at chess (though you do have to know the rules).
- However, it's hard for me to imagine this process working to build an AGI. It seems you'd need a perfect simulation of the entire real world to allow it to learn what intelligence is by trial and error. That seems infeasible.
- Instead, in order to build a program that is better at writing source code for AGIs than we are, it seems like you'd likely need to fundamentally understand and formalize what general intelligence consists of. How else can you tell the original program how to evaluate the "goodness" of different possible modifications it might make to its source code?
- As I mentioned in the last email, it seems like one with this understanding could jump to building the Google Maps style AGI anyway just by implementing a human-quality general intelligence algorithm in superior hardware. In other words, a computer that can write an AGI program better than a human is probably already an AGI.
- Another note is that even if the real world is more like chess than I think ... the actual story of the development of superhuman chess intelligences as I understand it is much closer to "humans writing the right algorithm themselves, and implementing it in hardware that can do things they can't" than to "a learning algorithm teaching itself chess intelligence starting with nothing but the rules." More generally, I believe that the overwhelming majority of useful superhuman programs have been closer to "written by humans" than to "written by themselves, given only the rules, the goal and a simulated world in which to play." So while it may be intuitive that it's easier to write a self-improving machine and "watch the fireworks," in practice it seems that it's more common for humans to figure out the right algorithm themselves (with help from computers, but not computers running the whole show).

Interested in your thoughts. For now I'm just interested in thinking through these things; we'll figure out how to formalize & publish the exchange later.

Email from Jaan Tallinn, 5/17/11

…1. sure, working on this over email and formalising the notes later is fine by me

2. re hidden complexity of wishes not applying to AI that's not designed to maximize some parameter: i'm not too well versed on the intricacies there (hopefully i'll meet nick bostrom next week, so can run your arguments by him!). however, i seem to remember that it is not trivial to delineate such class of AI-s, because of a) implicit goals (eg, an oracle maximizing the probability that the answer is true), and b) instrumental goals (eg, pretty much every AI system includes sub-functions that have "maximize X" as their explicit goal).

more generally, it seems to me that any powerful optimizer (which AGI is) must, in order to be safe, be a) parameter-sufficing (both in entirety and subsystem-wise) and b) have limited interactions with environment. furthermore, both criteria can be hard to meet and seem to require developers who are safety-aware (parameter-maximizer is algorithmically easier than parameter-sufficier and interaction control is effectively an exercise in AI-boxing).

3. re the "google maps AGI" there are 2 pending questions: 1) is it safe? 2) is the work (the likes of) SIAI are doing relevant for such scenario? we seem to be in rough agreement about the former (it's *not* safe because the risk of misuse - either due some bug or malice - is too high). about the latter, i would argue that yes it is, because a) such oracle is just another specimen in the zoo of potential AGI-s whose likelihood and properties must be researched before it's too late, b) SIAI's work would increase the chances that such system would be developed by safety-aware developers, and c) such developers (and by extension entire world) would benefit from safety-related tools and techniques.

(for similar reasons i think SIAI's work can be relevant for whole brain emulation scenarios: even though they don't involve recursive improvements by default, there's still the question of how to control/contain/align a mind that's more capable in reaching its goals than we are in preventing it from doing so)

4. btw, just thought of a "test question for seemingly safe oracle designers" (this email exchange is indeed really useful -- it's just unfortunate how badly it eats into the time i need for my presentation): what would the oracle do if asked a *really* hard question (such as "is p=np?" or "does this program halt?")? in particular, given that it a) obviously prefers a future where that question is answered, b) has to have powerful sub-systems that work towards that goal, and c) possibly knows the laws of physics better than we do, explain the mechanism that guarantees the lack of detrimental side-effects to that computation.

5. re the probabilities of scenarios #2 and #3 in my previous email: the reason why i think the recursive improvement scenario is more likely is similar to your (or was it dario's?) observation of teams-of-men-and-machines being the most powerful optimizers we have today. i think the main reason behind that is that the machines are *different* than humans. therefore, i think it's valuable to use early iterations of a would-be-AGI as a tool in its own development process -- not unlike compilers are used to compile themselves today.

finally, let me make the meta-point that this email exchange really is an instance of the class of discussions the likes of SIAI are engaging in -- so if you see value in this particular instance, i hope you see value in the whole class :)

- jaan

Email from Holden Karnofsky, 5/18/11

Thanks for the thoughts & no problem about the responsiveness. My responses below.

> 1. sure, working on this over email and formalising the notes later is fine by me

Great.

> 2. re hidden complexity of wishes not applying to AI that's not designed to maximize some parameter: i'm not too well versed on the intricacies there (hopefully i'll meet nick bostrom next week, so can run your arguments by him!). however, i seem to remember that it is not trivial to delineate such class of AI-s, because of a) implicit goals (eg, an oracle maximizing the probability that the answer is true), and b) instrumental goals (eg, pretty much every AI system includes sub-functions that have "maximize X" as their explicit goal).
>
> more generally, it seems to me that any powerful optimizer (which AGI is) must, in order to be safe, be a) parameter-sufficing (both in entirety and subsystem-wise) and b) have limited interactions with environment. furthermore, both criteria can be hard to meet and seem to require developers who are safety-aware (parameter-maximizer is algorithmically easier than parameter-sufficier and interaction control is effectively an exercise in AI-boxing).

This argument has repeatedly been made to me by SIAI affiliates, but never in a way that has made sense to me, at least outside the context of a program specifically designed to rewrite its own souce code and/or collect its own data.

Here's how I picture the Google Maps AGI ...

```
utility_function = construct_utility_function(process_user_input());
foreach $action in $all_possible_actions {
  $action_outcome = prediction_function($action,$data);
  $utility = utility_function($action_outcome);
  if ($utility > $leading_utility) { $leading_utility = $utility; $leading_action = $action; }
}
report($leading_action);
```

construct_utility_function(process_user_input()) is just a human-quality function for understanding what the speaker wants. prediction_function is an implementation of a human-quality data->prediction function in superior hardware. $data is fixed (it's a dataset larger than any human can process); same with $all_possible_actions. report($leading_action) calls a Google Maps-like interface for understanding the consequences of $leading_action; it basically breaks the action into component parts and displays predictions for different times and conditional on different parameters.

In this framework, the only function that really even needs to do something beyond the capabilities of current humans and computers is prediction_function. Which function(s) would you be concerned about here?

> 3. re the "google maps AGI" there are 2 pending questions: 1) is it safe? 2) is the work (the likes of) SIAI are doing relevant for such scenario? we seem to be in rough agreement about the former (it's *not* safe because the risk of misuse - either due some bug or malice - is too high). about the latter, i would argue that yes it is, because a) such oracle is just another specimen in the zoo of potential AGI-s whose likelihood and properties must be researched before it's too late, b) SIAI's work would increase the chances that such system would be developed by safety-aware developers, and c) such developers (and by extension entire world) would benefit from safety-related tools and techniques.

I think the "safety" that SIAI is working on is very different from the "safety" that would be needed for the Google Maps AGI. GMAGI would be dangerous if the user were greedy or malicious, in which case they wouldn't make use of Friendliness research; if the user had good intentions (particularly a willingness to allow other humans to put checks and balances on him/her), I think GMAGI would be safe.

> (for similar reasons i think SIAI's work can be relevant for whole brain emulation scenarios: even though they don't involve recursive improvements by default, there's still the question of how to control/contain/align a mind that's more capable in reaching its goals than we are in preventing it from doing so)

Perhaps.

> 4. btw, just thought of a "test question for seemingly safe oracle designers" (this email exchange is indeed really useful -- it's just unfortunate how badly it eats into the time i need for my presentation): what would the oracle do if asked a *really* hard question (such as "is p=np?" or "does this program halt?")? in particular, given that it a) obviously prefers

> a future where that question is answered, b) has to have powerful sub-systems that work towards that goal, and c) possibly knows the laws of physics better than we do, explain the mechanism that guarantees the lack of detrimental side-effects to that computation.

Sounds reasonable for one who is worried that their Oracle is unsafe.

> 5. re the probabilities of scenarios #2 and #3 in my previous email: the reason why i think the recursive improvement scenario is more likely is similar to your (or was it dario's?) observation of teams-of-men-and-machines being the most powerful optimizers we have today. i think the main reason behind that is that the machines are *different* than humans. therefore, i think it's valuable to use early iterations of a would-be-AGI as a tool in its own development process -- not unlike compilers are used to compile themselves today.

I agree that developers would likely develop tools as they go and work with those tools. But you're describing humans working in tandem with computers to design tools and use them on their way to building an AGI - not designing a dumber-than-humans computer to modify its source code all on its own until it becomes smarter than humans. I don't see how the latter would be possible for a general intelligence (for a specialized intelligence it could be done via trial-and-error in a simulated environment). The latter has a clear need for the kind of research SIAI is doing, but I don't think the former does, for the reasons sketched above.

> finally, let me make the meta-point that this email exchange really is an instance of the class of discussions the likes of SIAI are engaging in -- so if you see value in this particular instance, i hope you see value in the whole class :)

I see value in the whole class. I find SIAI to be focusing on a particularly unhelpful subset. I'd change my mind if someone persuaded me at the object level (that's why I'm having this conversation with you) or if I saw a broad enough consensus of people who ought to know better than I.

Email from Jaan Tallinn, 5/25/11

hi again -- just had a free afternoon to think about this topic.

to answer your question: in that particular system i'm definitely concerned about the prediction_function().

i think it would be worthwhile to concentrate on the GMAGI scenario a bit more, since it seems that we assess it differently. hence, let me ask you a few questions. nick bostrom has a useful classification of AGI-s in his upcoming book: 1) oracles (systems that answer questions), 2) genies (systems that fulfill wishes) and 3) sovereigns (systems that have an open-ended mandate to operate in the world).

with that in mind:

1. would you agree that GMAGI has to include a full blown oracle under its hood in order to be powerful enough to qualify as an AGI? i'm thinking along the lines that if there are questions that an oracle could answer but GMAGI (implicitly) can't, then it would not be able to make very good predictions in domains that involve such questions.

2. would you agree that it is rather trivial to extend such GMAGI to be a genie or sovereign by adding a main loop (the "master function") that simply uses the predictor to maximise an explicit utility function?

3. would you agree that there's the obvious way to make the GMAGI more powerful by asking it for ways to improve its own source code? and, furthermore, such improvement would seem by far the lowest hanging fruit to the developer who implemented the

GMAGI in the first place (assuming he's not concerned with safety)?

4. in the world where there are multiple competing GMAGI-s, there's enormous strategic pressure to groups running them to make their GMAGI the most powerful one by 1) maximising its prediction ability, 2) minimising the time the GMAGI is not utilised (eg, the time it is waiting for next human interaction), 3) trying to sabotage the competing GMAGI-s?

- jaan

Email from Holden Karnofsky, 5/26/11

Thanks for the thoughts, some responses:

> to answer your question: in that particular system i'm definitely concerned about the prediction_function().

Can you spell that out a bit, or is that what you're doing with the below questions?

> 1. would you agree that GMAGI has to include a full blown oracle under its hood in order to be powerful enough to qualify as an AGI? i'm thinking along the lines that if there are questions that an oracle could answer but GMAGI (implicitly) can't, then it would not be able to make very good predictions in domains that involve such questions.

I was thinking of GMAGI as a form of Oracle. Not sure what you mean by "full blown oracle under its hood."

> 2. would you agree that it is rather trivial to extend such GMAGI to be a genie or sovereign by adding a main loop (the "master function") that simply uses the predictor to maximise an explicit utility function?

Probably though not definitely. If the "master function" is just sending packets of data, as you proposed, it won't necessarily have the ability to accomplish as much as a well-funded, able-bodied human would. I'm aware of the arguments along the lines of "humans figured out how to kill elephants ... this thing would figure out how to overpower us" and I think they're probably, though not definitely, correct.

> 3. would you agree that there's the obvious way to make the GMAGI more powerful by asking it for ways to improve its own source code?

Maybe. It seems easy for a GMAGI to be intelligent enough to create all-powerful weapons, cure every disease, etc. while still not intelligent enough to make improvements to its own predictive algorithm *that it knows are improvements for general intelligence, i.e., predictive intelligence in every domain.*

I think it's likely that we will ultimately arrive at prediction_function() by imitating the human one and implementing it in superior hardware. The human one has been developed by trial-and-error over millions of years in the real world, a method that won't be available to the GMAGI. So there's no guarantee that a greater intelligence could find a way to improve this algorithm without such extended trial-and-error. It depends how much greater its intelligence is.

and, furthermore, such improvement would seem by far the lowest hanging fruit to the developer who implemented the GMAGI in the first place (assuming he's not concerned with safety)?

Probably not. If I were a possessor of a GMAGI, I'd want to develop superweapons, medical advances, etc. ASAP. So first I'd see whether it could do those without modifying itself. And I think a GMAGI capable of writing a superior general intelligence algorithm is probably capable of those other things as well.

4. in the world where there are multiple competing GMAGI-s, there's enormous strategic pressure to groups running them to make their GMAGI the most powerful one by 1) maximising its prediction ability, 2) minimising the time the GMAGI is not utilised (eg, the time it is waiting for next human interaction), 3) trying to sabotage the competing GMAGI-s?

Yes, especially 3), but doesn't seem very relevant to me. The GMAGI is, to me, the most likely "basement AI" scenario, i.e., the one where turning yours on a day before anyone else is a decisive advantage. That's because the GMAGI could be used to quickly develop superweapons and gain control of the world, including others trying to build their own AGIs. While I see a "multiple GMAGIs" scenario as possible, I see a "multiple GMAGIs each run by reckless and unethical people" very unlikely. The first reckless and unethical person to turn theirs on probably wins, with or without self-modification.

If I built a GMAGI and I were greedy/unethical,

- I'd get right down to taking control of the world. As stated above, I think it's unlikely that an AGI would be capable of making helpful modifications to its own source code unless it were already capable of designing the technologies that could take over the world.
- If I did try self-modification, it would be along the lines of "modify prediction_function; calculate action that takes over the world; if action is fast enough and high-confidence enough, end loop and display action to me, otherwise modify prediction_function again and repeat."
- The last thing I'd want to do would be to "let the AGI loose" to self-modify beyond recognition and take its own actions in the world - especially if I were familiar with the very arguments SIAI is advancing for why this is so dangerous.
- If useful research for assuring "friendliness" were available to me, I wouldn't trust it anyway. I don't find it realistic that one could have *high enough* confidence that an unrecognizably self-modified AGI will be Friendly, *even if* the algorithm for making it so has been designed and thoroughly checked.

Email from Jaan Tallinn, 5/26/11

thanks, i'll think about this when i have a moment and write a longer reply, but here's a quick meta-comment. what i'm getting at with my questions/arguments is to demonstrate that GMAGI is either:

1) a narrow AI that's safe only due to being significantly below human level in some crucial aspect (such as programming ability or the ability to model/predict humans) OR

2) an oracle in disguise that *seems* safe because we're only considering the case where it's applied to a narrow domain.

right now, based on your responses, it seems to me that (1) is the case. but i'll do more thinking and get back to you (probably after my talk next week).

- jaan

Email from Holden Karnofsky, 5/27/11

Sounds good.

Email from Jaan Tallinn, 6/13/11

…it seems to me that the key to our differences in opinion is here:

> It seems easy for a GMAGI to be intelligent enough to create all-powerful weapons, cure every disease, etc. while still not intelligent enough to make improvements to its own predictive algorithm

so GMAGI would -- effectively -- still be a narrow AI that's designed to augment human capabilities in particularly strategic domains, while *not* being able to perform tasks such as programming. also, importantly, such GMAGI would *not* be able to make non-statistical (ie, individual) predictions about the behaviour of human beings, since it is unable to predict their actions in domains where it is inferior.

would you agree to that?

if so, then we have the following questions: 1. how powerful could such narrow-AI augmented human teams get? 2. would they become powerful enough to delay or block the arrival of "true" AGI (eg, by enabling an all-controlling world government -- a singleton)? 3. is it worthwhile to think about those questions, and are the likes of SIAI qualified to do so?

here are my own answers/opinions:
1. i think it's a very valid question (and i've heard the assertion of narrow AI being an imminent danger before from an SIAI critic). i don't think the answer is trivial though, because from one hand, narrow AI-s will indeed get much more powerful as the algorithms and hardware improve, but on the other hand, being narrow leaves them (and their sponsors) vulnerable to attacks that utilise the domains the AI-s are incompetent in. also, predicting the world that's full of intelligent agents does not strike me as simpler activity as programming (for example, i know lots of people - including myself - who are good at latter yet suck at former)
2. given the above considerations, i would assign nontrivial but significantly below 50% probability to such "narrow AI will prevent AGI from happening" scenario
3. yes & yes!

- jaan

Email from Holden Karnofsky, 6/15/11

…thanks for the further thoughts .

> It seems easy for a GMAGI to be intelligent enough to create all-powerful weapons, cure every disease, etc. while still not intelligent enough to make improvements to its own predictive algorithm

> so GMAGI would -- effectively -- still be a narrow AI that's designed to augment human capabilities in particularly strategic domains, while *not* being able to perform tasks such as programming. also, importantly, such GMAGI would *not* be able to make non-statistical (ie,

individual) predictions about the behaviour of human beings, since it is unable to predict their actions in domains where it is inferior.

would you agree to that?

I don't think so, at least not fully. I feel like once we basically understand how the human predictive algorithm works, it may not be possible to improve on that algorithm (without massive and time-costly experimentation) no matter what the level of intelligence of the entity trying to improve on it. (The reason I gave: The human one has been developed by trial-and-error over millions of years in the real world, a method that won't be available to the GMAGI. So there's no guarantee that a greater intelligence could find a way to improve this algorithm without such extended trial-and-error)

Similarly, it may be impossible to predict the actions of humans with too much precision. We can't predict the actions of a given insect with too much precision despite being much smarter than it.

If I'm wrong about both of these things, I still think it is possible to build a GMAGI capable of reprogramming itself usefully, and still choose not to have it do so. If I had just flipped on such a GMAGI, I might be afraid to have it reprogram itself due to safety concerns, and choose instead to have it develop superweapons or other insights directly, feeling pretty confident that I had at least a week head start on others and that this would be plenty of time to build an insurmountable advantage without source code rewriting.

1. how powerful could such narrow-AI augmented human teams get?

1. i think it's a very valid question (and i've heard the assertion of narrow AI being an imminent danger before from an SIAI critic). i don't think the answer is trivial though, because from one hand, narrow AI-s will indeed get much more powerful as the algorithms and hardware improve, but on the other hand, being narrow leaves them (and their sponsors) vulnerable to attacks that utilise the domains the AI-s are incompetent in. also, predicting the world that's full of intelligent agents does not strike me as simpler activity as programming (for example, i know lots of people - including myself - who are good at latter yet suck at former)

- As stated above, I don't think of the GMAGI I'm describing as necessarily narrow - just as being such that assigning it to improve its own prediction algorithm is less productive than assigning it directly to figuring out the questions the programmer wants (like "how do I develop superweapons"). There are many ways this could be the case.
- I don't think "programming" is the main challenge in improving one's own source code. As stated above, I think the main challenge is improving on a prediction algorithm that was formed using massive trial-and-error, without having the benefit of the same trial-and-error process.
- When talking either about a narrow-AI-and-human team, or a GMAGI-and-human team: it seems pretty clear to me that one of these teams could be powerful enough *to render all previous work by the likes of SIAI moot*, which isn't necessarily the same as (though could involve) conquering the world. If such a team develops superweapons, gains control of the world's resources, and stops the development of other AI's, then SIAI's work is moot. But also, such a team could turn its powers on studying the same questions SIAI is currently studying, and very quickly come up with everything the likes of SIAI has come up with and far more.
- It is actually pretty hard for me to imagine that this *won't* happen, i.e., that a self-improving AND all-domain AND autonomous/acting (as opposed to oracle) AI will be developed before a team exists that has the ability to moot SIAI's work in one way or another.

2. would they become powerful enough to delay or block the arrival of "true" AGI (eg, by

enabling an all-controlling world government -- a singleton)?
2. given the above considerations, i would assign nontrivial but significantly below 50% probability to such "narrow AI will prevent AGI from happening" scenario

See above.

3. is it worthwhile to think about those questions, and are the likes of SIAI qualified to do so?
3. yes & yes!

Agreed, but instead I see them focused on a goal that seems to presuppose answers to these questions.

Email from Holden Karnofsky, 6/15/11

This thought is in the below email, but I wish to highlight it:

I think that if you're aiming to develop knowledge that won't be useful until very very far in the future, you're probably wasting your time, if for no other reason than this: by the time your knowledge is relevant, someone will probably have developed a tool (such as a narrow AI) so much more efficient in generating this knowledge that it renders your work moot.

If you're aiming to develop knowledge that won't be useful *until society is capable of creating an AGI*, you seem to be almost guaranteeing this situation.

Email from Jaan Tallinn, 6/21/11

hello!

a few metapoints first:

1. i don't think we should talk about guarantees -- clearly there is no guarantee that earth will not be hit by asteroid today afternoon, but that does not invalidate what SIAI is doing. ie, we should talk about *probabilities* instead.

2. i stand corrected re the GMAGI definition -- from now on let's assume that it *is* a full blown AGI in the sense that it can perform every intellectual task better than the best of human teams, including programming itself.

3. looks like i need to better define what i mean by algorithm, program and programming (a.k.a. coding). i hereby define "algorithm" as a sequence of instructions that transform input (data) to output (data or actions). "program" is an implementation of an algorithm that performs the transformation with the help of resources such as time, memory and energy. "programming" is the activity of constructing or modifying a program (thus potentially modifying the algorithm that the program implements). importantly, i categorise "algorithm design" as a sub-task of "programming".

4. i think we should make an effort to try to converge the current discussion on a set of differing intuitions and/or probability estimates.

anyway - in the light of the above - here are my responses to your points:

5. re the "human cognitive algorithm cannot be improved" argument: i don't think that's true. reasons: a) while it's true that evolution had a massive amount of tries at its disposal, human cognition is very recent and thus likely sub-optimal "invention" in evolutionary context, b) evolution had a ridiculous amount of constraints that it had to work with (basically constructing computers from what seem very unsuitable materials while making sure that they perform a ton of functions unrelated to cognition), c) evolution optimised for gene survival not necessarily for cognitive performance (this fact is reflected in the plethora of cognitive biases that we exhibit), and d) i think it's very likely that the first AGI will find lots of low hanging fruits in terms of algorithm/program improvements just by the virtue of being non-human mind, and hence having different abilities (such as being able to follow extremely long chains of inference, having different modalities, etc).

6. re AGI not choosing to reprogram itself: there are 2 problems with assuming this: a) since AGI is (by definition) at or above the level of humans, reprogramming itself will clearly occur to it as a possible action (instrumental goal) to take, b) for an intelligence explosion it might suffice if it would program another AGI, not necessarily improve itself. so there has to be a clear reason/obstacle why it will refrain from programming, while clearly being capable of doing that.

7. moreover, even if we had a proof that AGI would be unlikely to engage in programming, we should be confident that the creators of AGI would not tap into the fledgling AGI as programmer resource (see point (5) above for reasons why they might consider AGI a resource). since software development teams in general are always in lookout for programming talent, i find it rather plausible that AGI teams would try to devise ways how to do that. perhaps our different opinions here are caused by our different intuitions about how an AGI development process might look like -- in my view it's unlikely to be a linear "design an algorithm, enter it into machine, press f5 to run" process. instead, i envision a drawn out and iterative process that's typical for complex software projects today.

8. re a narrow-AI equipped team being better able to think about AGI than SIAI: yes, that's possible. however, in order to be confident (see my point (1) above) we should have a very plausible scenario that causes such team to materialise and supercede SIAI's work -- ie, we can't just postulate it, just like we should not postulate an asteroid hitting the earth this afternoon. also, let me point out that my experience with talking to narrow-AI developers indicates that they are usually not aware of safety/AGI issues (and when they are, the awareness can often be traced to SIAI's work).

9. re the creation of AGI being "very very far in the future": can you quantify that a bit more? myself i would give at least 10% of probability to that happening within the next

2 decades (of course i will update this as our knowledge about brain and AI improves). since i believe this to be irreversible event of enormous magnitude, i don't think we can afford to sit and wait for our capabilities to improve.

- jaan

Email from Holden Karnofsky, 6/21/11

Hello,

I think the biggest disagreement between us still comes down to whether it is possible to build a GMAGI as I define it. I earlier sketched out what I think the code of this GMAGI would look like, and said that I couldn't see why such a thing should be expected to do anything other than think/report/discuss - I saw no function that seemed it could involve trying to obtain computronium, rewriting itself, building another AGI, etc. (Sure it could *suggest* these actions, but not execute them.) Just as Google Maps (no matter how much better it is than you at navigation) has no desire, ability or possibility of driving your car.

It seems to me from #6 below email that you still don't think this is possible, but I don't see why. All I know is that you think prediction_function() is where the risk is. I don't understand this position, because prediction_function() need not be a self-improving function or a "maximizing" function; it can just be an implementation of the human version with fewer distractions and better hardware.

On other points:

> 1. i don't think we should talk about guarantees -- clearly there is no guarantee that earth will not be hit by asteroid today afternoon, but that does not invalidate what SIAI is doing. ie, we should talk about *probabilities* instead.

Agreed, but I think that sometimes attempting to quantify a probability that I have very little sense of damages rather than helps the clarity of the message. When I say "almost guaranteeing" I'm trying to communicate that the probability seems about as high as I'm willing to assign on questions like these (and I'm not sure exactly what that number ought to be, which is a separate discussion).

> 4. i think we should make an effort to try to converge the current discussion on a set of differing intuitions and/or probability estimates.

I think we are getting there, but the point at the beginning of the email seems like the big one to me. In most of my answers on this email, I try to give the conclusion for where I stand on various key probabilities.

> 5. re the "human cognitive algorithm cannot be improved" argument: i don't think that's true. reasons: a) while it's true that evolution had a massive amount of tries at its disposal, human cognition is very recent and thus likely sub-optimal "invention" in evolutionary context, b) evolution had a ridiculous amount of constraints that it had to work with (basically constructing computers from what seem very unsuitable materials while making sure that they perform a ton of functions unrelated to cognition), c) evolution optimised for gene survival not necessarily for cognitive performance (this fact is reflected in the plethora of cognitive biases that we exhibit), and d) i think it's very likely that the first AGI will find lots of low hanging fruits in terms of algorithm/program improvements just by the virtue of being non-human mind, and hence having different abilities (such as being able to follow extremely long chains of inference, having different modalities, etc).

The scenario I'm picturing, in which a GMAGI appears, is where humans themselves get around (c) and (d) simply by figuring out what prediction algorithm humans implement, and implementing a "pure" form of it in a computer. This computer would then be orders of magnitude more "intelligent" than humans in all ways. But it doesn't necessarily follow that it would be able to improve on its own prediction algorithm. Your points (a) and (b) are valid and applicable; it is also true that the computer would be smarter than humans (this pertains to (c) and (d)) and may therefore think of ways to improve its algorithm that we haven't thought of. But it is unclear whether, and by how much, these points outweigh the huge head start that evolution had in designing the prediction algorithm.

All I'm saying here is that if we did build an AGI, there would be a substantial probability that it could improve its own prediction algorithm, and a substantial probability that it could not (at least not efficiently).

> 6. re AGI not choosing to reprogram itself: there are 2 problems with assuming this: a) since AGI is (by definition) at or above the level of humans, reprogramming itself will clearly occur to it as a possible action (instrumental goal) to take, b) for an intelligence explosion it might suffice if it would program another AGI, not necessarily improve itself. so there has to be a clear reason/obstacle why it will refrain from programming, while clearly being capable of doing that.

This relates to the point at the beginning of the email.

> 7. moreover, even if we had a proof that AGI would be unlikely to engage in programming, we should be confident that the creators of AGI would not tap into the fledgling AGI as programmer resource (see point (5) above for reasons why they might consider AGI a resource). since software development teams in general are always in lookout for programming talent, i find it rather plausible that AGI teams would try to devise ways how to do that. perhaps our different opinions here are caused by our different intuitions about how an AGI development process might look like -- in my view it's unlikely to be a linear "design an algorithm, enter it into machine, press f5 to run" process. instead, i envision a drawn out and iterative process that's typical for complex software projects today.

I would bet that development teams would tap into budding AGI in whatever ways they could that were clearly "safe," such as asking it for ways to improve itself and considering them (ala Google Maps), or building specialized sub-intelligences to execute iterative loops to improve particular parts of the process. I do not think they would be likely to set the AGI on the sort of open-ended, fully-authorized-to-intervene-in-the-world task for which SIAI's work would be necessary to ensure safety. Here, again, I am giving implicit probability estimates; I understand that yours are different but see no arguments I hadn't previously considered, so we may have to agree to disagree here.

> 8. re a narrow-AI equipped team being better able to think about AGI than SIAI: yes, that's possible. however, in order to be confident (see my point (1) above) we should have a very plausible scenario that causes such team to materialise and supercede SIAI's work -- ie, we can't just postulate it, just like we should not postulate an asteroid hitting the earth this afternoon. also, let me point out that my experience with talking to narrow-AI developers indicates that they are usually not aware of safety/AGI issues (and when they are, the awareness can often be traced to SIAI's work).

I think there are many such plausible scenarios. For example, a specialized "philosopher AI," capable only of doing philosophy at a superhuman level, could quickly render SIAI's existing work moot. Or, a specialized "psychology/social sciences" A.I. may enter the sort of debate we're having and demonstrate that the existence of Friendliness theory would have practically nil impact on the actions of those who eventually develop AGI. Any tool that is very helpful in either of those two fields could do the trick.

> 9. re the creation of AGI being "very very far in the future": can you quantify that a bit more? myself i would give at least 10% of probability to that happening within the next 2 decades (of course i will update this as our knowledge about brain and AI improves). since i believe this to be irreversible event of enormous magnitude, i don't think we can afford to sit and wait for our capabilities to improve.

I don't think this is a key point of disagreement. Your estimate doesn't sound unreasonable to me. What I meant by "far" was more in technology terms than in time terms. I believe that we are likely to create a lot of useful, revolutionary narrow A.I.s before we create AGI, even if the latter happens within 20 years.

Email from Jaan Tallinn, 6/22/11

hi!

ok, let's stash the auxiliary points (programming ability of AGI, anatomy of AGI projects, role of narrow AI in near future) for now and focus on the main point of difference (GMAGI).

i do agree that GMAGI is *possible* (just like an asteroid strike is). however, in order to use it as an argument against the value of SIAI's (current) work, you'd also need to show the following 4 points:

1. is overwhelmingly likely that GMAGI will be the first realised AGI project (if it's not, then it does not relieve us from the need to consider the entire class of AGI-s. i don't

remember if you've addressed this point);

2. it is very unlikely that GMAGI will be used (accidentally or intentionally) to create another AGI (this we started to discuss, but i don't think we reached an agreement);

3. GMAGI's predict() function would be free of side-effects (we also started discussing that, but i expect that this is actually very hard to show, since we have defined GMAGI to be an instance of AGI, making predict() effectively "a rug" that all potential issues are swept under);

4. there aren't any significant aspects of SIAI-s (current) work that would be applicable to GMAGI (if the above points (1-3) are demonstrated, then this is probably true, but someone more familiar with SIAI's work might still point out something).

- jaan

Email from Holden Karnofsky, 6/22/11

Hi Jaan,

I do think we have a big disagreement on where the "burden of proof" lies and how to model uncertainty about the relevance of SIAI's work. But I think we can table that - my goal is to figure out which of our disagreements are based on (a) differing intuitions vs. (b) something else such as differing info, insights or understanding of computers. I am now thinking that (a) is the main factor behind most of our disagreements.

> 1. is overwhelmingly likely that GMAGI will be the first realised AGI project (if it's not, then it does not relieve us from the need to consider the entire class of AGI-s. i don't remember if you've addressed this point);

I do think this. The "google maps" framework is conceptually very broad. In fact, I think that pretty much all software developed to date fits the "google maps" description as I intend it, conceptually: it seems to me that pretty much all software to date is designed as a partner to humans that makes predictions & gives suggestions, and provides tools for digging on and understanding these predictions & suggestions. It doesn't present inherent "boxing" challenges: it acts in the way I described until and unless humans decide to hook it up to an acting agent, which they don't do until they are very confident in the safety of doing so.

Because software that behaves in this general way is most useful to and safe for humans, and because most software to date has been easy to make behave this way, I feel it is the right default expectation for what an AGI will look like, and I expect others to explain why AGI inherently must work differently. The only explanation I've heard so far is that an AGI may come about by rewriting its own source code, a process that is unstable and unpredictable in a way that all past "using software to help one write better software" endeavors have not been.

I argued that any program capable of improving its own source code for general predictive purposes - without human help or time-consuming real-world experimentation - is likely already smart enough to qualify as an AGI and thus to be useful enough in the GMAGI framework to conquer the world, study Friendliness, or otherwise moot the work of SIAI.

> 2. it is very unlikely that GMAGI will be used (accidentally or intentionally) to create another AGI (this we started to discuss, but i don't think we reached an agreement);

My sense is that we just have differing intuitions here.

- You argued that humans are likely to set an AGI to improve itself in order to make sure that they have the best AGI before anyone else does.
- I argued that humans are unlikely to do this if it is unsafe, and that an AGI capable of self-improvement is probably already powerful enough to be useful enough without this measure. I also questioned whether an AGI would necessarily be able to self-improve, but that was a more minor point (I concede it might be).

> 3. GMAGI's predict() function would be free of side-effects (we also started discussing that, but i expect that this is actually very hard to show, since we have defined GMAGI to be an instance of AGI, making predict() effectively "a rug"

that all potential issues are swept under);

I don't see it this way. In order to consider several different paths of action, it seems that a computer must be able to estimate the consequences of an action without actually performing the action. Certainly the human predictive algorithm works like this and is therefore "safe" in this sense.

I see prediction($action,$data) as a straightforward calculation function modeled on the human version, but far more powerful when implemented in pure form in superior hardware. $action and $data are given digital representations and there is some mathematical formula that produces a prediction from them; we currently use computers to compute complex formulas on complex data all the time, and I don't see where the risk of the computer's directly interfering in the world comes in.

4. there aren't any significant aspects of SIAI-s (current) work that would be applicable to GMAGI (if the above points (1-3) are demonstrated, then this is probably true, but someone more familiar with SIAI's work might still point out something).

Agreed.

To be clear, my stance is:

- The overwhelmingly likely case is that a team of humans and narrow AIs moots SIAI's work before AGI is ever developed. (I think we've gotten to the point here where it's clear that we just have differing intuitions.)
- If AGI appears suddenly enough to pre-empt the above possibility, I think it will take the form of implementing a human-like prediction algorithm in superior hardware, which will lead to a GMAGI, which will moot SIAI's work. (This is the main topic of this email.)
- It's possible that things will go differently, and the two of us have differing intuitions about how much weight to give this possibility and how to incorporate it into our actions.

Email from Jaan Tallinn, 6/24/11

okay, let me try to rephrase your position to distill out the essence of it:

you believe that
- barring disruptive events (asteroids, nuclear or biowarfare, narrow ai disasters, etc),
- there's a very high (99%+) probability that
- the first AGI created will be GMAGI (smarter than human oracle for devising strategical paths)
- AND the GMAGI will be free of side effects (such as tapping into additional resources, giving manipulative answers, etc)
- AND the GMAGI will not be used to create an AGI of another class (even though GMAGI is powerful enough to allow that).

would you say that's accurate?

i would agree that for a person having such belief the work of SIAI would appear to have very low value (but like i said, i'm not familiar with all the work SIAI is doing). myself i would assign less than 10% probability to such scenario based on the "too many ANDs" heuristic alone, and probably less than 1% after weighing in additional evidence.

- jaan

Email from Holden Karnofsky, 6/24/11

That isn't exactly my position, and of course the way I would put it would sound more reasonable :), but you have hit my major intuitions and what you wrote may be close enough for our purposes for now.

It seems to me that all the ways in which we disagree have more to do with philosophy (how to quantify uncertainty; how to deal with conjunctions; how to act in consideration of low probabilities) and with social science-type intuitions (how would people likely use a particular sort of AI) than with computer science or programming (what properties has software usually had historically; which of these properties become incoherent/hard to imagine when applied to AGI). Does that sound fair?

If so, perhaps we ought to publish this email exchange with a summary from each of us? Of course I am not trying to end the discussion, but if we've reached the point where we're more discussing philosophy and social science-type topics, that might be a good time to put the conversation out in the open and let others weigh in as they wish.

Best,
Holden

Email from Jaan Tallinn, 6/24/11

well, i don't think the analysis into the nature of our disagreements would be very valuable at this point, because 1) i have deliberately not responded to many of your points in order to keep the discussion focused, and - more importantly - 2) the result of that analysis does not really matter much as we have dug up a rather concrete disagreement (about the probability the GMAGI scenario) that alone suffices to explain our different valuation of SIAI's work.

(oh, and since there's really one mathematically correct way of dealing with uncertainty and conjunctions, i certainly hope that i haven't said anything that puts me on the wrong side there :) let me know if you think i have).

anyway, you have my permission to publish the exchange at this point -- i think it was very worthwhile. should we continue it sometime, a natural next step would be to list all the arguments in favor/against the components of the GMAGI scenario and try to assess how they influence the probabilities.

also, btw, unless something comes up, i should be coming over to new york this october in order to present at the singularity summit -- perhaps we could meet in person then!

- jaan

Email from Holden Karnofsky, 6/24/11

Great, thanks, and it would be great to meet you so let me know when you're around.

I think I will publish this exchange plus your version (i.e., the version with your edits) of the phone conversation. I think that will be followable by motivated enough people. I may later write up a summary and if you wrote up one of your own I'd be happy to include it. Sound OK?

A couple of points of clarification:

- I think the most likely scenario is that narrow AI moots SIAI's work (not necessarily through "catastrophe," could be by contributing to friendliness research or doing something else that we haven't thought of). I think another very strong possibility is "We just have the whole framework wrong and all of these conversations will seem silly in 50 years for reasons that we can't foresee now." I do see GMAGI as taking up most of the probability-space of AGI scenarios, but I think that's still a relatively small space compared to the other two.
- I very much doubt that we disagree on the *mathematical* aspects of how to handle conjunctions and uncertainty. But we are not dealing with well-defined or -quantified probabilities. Any prediction can be rephrased so that it sounds like the product of indefinitely many conjunctions. It seems that I see the "SIAI's work is useful scenario" as requiring the conjunction of a large number of questionable things, whereas you see the "GMAGI" scenario that way.

Best,
Holden

Email from Jaan Tallinn, 6/24/11

Sure, go ahead.