A conversation about MIRI strategy. For a summary and details, see:

http://intelligence.org/2014/01/13/miri-strategy-conversation-with-steinhardt-karnofsky-and-amodei/

Holden: Thanks for coming guys. Here's why I wanted to call the meeting. I have a lot of conversations with Dario and Jacob where we're like "Oh, I can't believe MIRI's doing blah, blah, blah. That's a bad strategy."

But I don't really know what you guys are doing or why. It's possible that we're here to argue and have us present what we disagree about. It's also possible that that's not what we're here to do at all.

I think the best starting point is to ask Luke and Eliezer, what you guys are working on and why. Then, react to that.

Eliezer: My perspective is that civilization doesn't seem to be handling AI friendliness on its own. I don't particularly expect that to change, so our mission: just go do it. Then, strategy: Get together people who can plot all the theoretical problems between here and there, and find enough money to fund them. Tactics: Hold research workshops and try to factor out community growth stuff to CFAR as much as possible.

Holden: I'm also interested in a level or two down from that. Who's doing what?

Eliezer: Next actions: I need to describe a bunch of open problems in Friendly AI as quickly as possible, maybe on Facebook for the first draft.

Luke, Louie, and Malo are running all the administrative stuff and organizing the workshops and so on.

Luke: …and development in general; running small pilot projects to see what things work for outreach, strategy research, or technical research; reaching out to the technical people who might want to come to a workshop; trying out grant writing to see if that works; producing the documents and website; community building stuff; etc.

Holden: What exactly is Paul doing?

Luke: He's been further developing his probabilistic metamathematics that was the result of the November 2012 workshop. He just gave a talk at Harvard on the latest version of that and he's working on writing a journal article version of that.

Holden: My two questions are, "Is working on this stuff likely to have a direct impact on getting good outcomes from AI?" and "Is your strategy targeting the right group of people?”

It sounds like a dual goal here: you’re trying to do things that are useful, while also attracting a community of people that is the right community of people to be attracted. Does that sound reasonable?

Eliezer: I guess so. I have never held out much hope for approaching a random AGI project 30 years from now that's about to destroy the world, showing them a bunch of results that they can use to make their AIs safe. It doesn't, necessarily, have to be in MIRI’s basement. It's going to be some project that shows what I would consider a reasonable level of concern, which is way, way beyond the level of concern we currently observe in AI projects.

They would have to have done their work with that level of concern from the ground up. If there's some kind of technique you can use to take some kind of powerful cognitive engine, built of arbitrary stuff, that pours out a Friendly AI out the other end, then I don’t know what that is and I’m not sure how it would work. It’s possible that could be done, but it seems too optimistic to me, so I’m not counting on it.

The present community that we're trying to gather is not so much a community of people who are going to be influential on someone else. It's more like a community of people who do research work.

Luke: I can add a few things. An imperfect analogy for what Eliezer was saying with regard to building things to be safe from the ground up is in safety-critical systems like auto-pilot software and so on.

In that domain, the architecture one uses is very different from what you do if you're just trying to make something that solves problems efficiently with as little design effort as possible because for a safety-critical system you want to get certain kinds of guarantees about the constraints of the behavior of the system.

Probably the easiest, fastest route to AGI is some massive kluge of machine learning and narrow AI algorithms or something like that, which is a very different architecture than the kind of thing that you would build from the ground up if you were thinking about safety from the beginning.

I figured I'd just add some stuff about my perspective on why MIRI is doing what it's doing. So, there are certain parts of Eliezer’s model where either because he’s thought about it longer than I have or because of something else, Eliezer’s probability mass is more concentrated on certain developments, mine is more spread out, and so the reasons why I changed my mind from a focus on strategy research to technical research at the beginning of 2013 is somewhat different than Eliezer's.

My reasons are significantly about thinking that we need to start making technical progress since no one else will. But it's largely about other things, like the fact that I expect certain kinds of strategic strategic insights to only come from doing Friendly AI research, like what is the distribution of worse-than-paperclipping AIs around Friendly AIs in mind design space and how much extra optimization power do you need to use to get past the worse-than-paperclippers to Friendly AI once you've already put in the optimization power to get into AGI — stuff like that is strategically relevant, but I don't expect to learn about it from a philosopher's chair, but from trying to build the systems.

Then also that, focusing on the technical work seems to be able to attract higher-*g* thinkers more easily than strategy research can, and I want literally the smartest people in the world to not only be working on the technical problems but also to be thinking about strategic issues — so for example, I'm starting a reading group for the near-complete draft of Bostrom’s *Superintelligence* book, which is all about the strategic picture. I'm getting some of the workshop participants to participate in that so they can be thinking about the strategic issues with their cognitive horsepower.

Then it's also things like: the traction with academia we get is much greater with technical research than strategic research.

I’ll also mention that for most of MIRI’s history, we were in fact busy with some of these other efforts, like strategic research and community-building and rationality training, but now those are being done by other groups like CFAR and CEA and FHI.

So my reasons for focusing on technical research are more spread out across many different things, compared to Eliezer, probably.

Holden: So some key questions I can imagine being really key are: “Should MIRI be thinking primarily about building its own FAI, maybe with another organization or something, rather than putting out public goods?” and “Are these technical problems likely to add value whether in the future for the team or in the future as public goods?" Then, there's a third question, “Is the community, you're engaging the community that is most important to engage?” Those are my three key questions.

Luke: Maybe now is a good time to add questions/objections from Jacob and Dario?

Jacob: Some of Holden's questions already touch on this, but three specific disagreements are: (1) I don't actually agree that reflection is well-acknowledged to be an important problem in AI, (2) Luke said something about attracting the most intelligent people but I think that's a bit oversimplified, people have different relevant skills and I’m not sure mathematicians have the optimal skillset for this, and (3) I'm not as convinced as you guys are that it’s necessary to build in safety from the ground up. There’s more of a sliding scale where the more technical work you do on safety research, the later in the pipeline you can insert it. And if you're willing to build it in from the ground up, you have to do the least amount of technical work, but at the cost of much more outreach.

Dario: My question is similar to Jacob's third question. If you think about things that are being done in perceptual systems, how it’s classifying different things is not something that’s being done in a safety-conscious way, you can’t understand internally exactly what the system is doing, and I’m skeptical that an AI that employed those kinds of methods would necessarily be dangerous.

I can see how it *might* happen, like the AI perceives things wrongly and therefore has a wrong model and therefore doesn't do what you expect it to do, but it seems really likely to me that an AI that had that module could understand that module and its limitations and could build modules like that itself, etc.

So the question to me, are you imagining putting 80 modules together like Legos and building them hierarchically? At what level of processing does it become important to think about the system’s overall goals and its hierarchy? It sounds like you're claiming that it's at its very lowest level but I'm not sure why that's true.

Eliezer: So if you're starting from scratch in designing an FAI system, then as far as I can see, once you outline the system, there's a bunch of stuff where you have constrained freedom of action internally, and a bunch of stuff where you have to be relatively more rigorous.

In particular, consider the part where the system is modifying itself, and there are things which can directly control policy and self modification: an error there can, on a self-modification, twist the future purpose of the system.

Dario: Right, clearly.

Eliezer: On the other hand, you’ve got these other parts where you are completely sure that you don't care where a suggestion originates from. For example, if there is a sufficiently logical subsystem that it has the character of theorem-proving, and you truly don’t care what the proof is so long as [inaudible].

Then, you completely don't care about what is producing the proof, as long as whatever is producing the proof is not actually a hostile superintelligence that can break out of the sandbox and destroy the rest of the system.

Jacob: That seems like a very extreme end of the spectrum. You could imagine things that are not literally providing proofs but that nevertheless you can treat as a black box – perhaps not taking their output as fiat, but…

Eliezer: Right, so there’s the epistemic and instrumental dimensions. Things that occupy relatively more epistemic dimensions and are optimized along epistemic dimensions, I think you probably have sort of more freedom of non-rigor, than with respect to things that are actually policies.

Things that produce policies directly, without going through an epistemic level, like if you are mutating policies, running them in simulation and picking the policy that worked best the last thousand times. At that point, you are trying to break down the divide that after epistemic/instrumental divide, and enter into realms where you're more nervous.

Then if that policy is also directly relating to self-modify actions, that's the point at which I start to throw out my hands and be like, "Maybe with an additional 50 years of research, we can figure out how to take this horrible construct that we created, and derive useful output from it, but it scares me."

From my perspective, the key abstract quality of what I just said is that, if you start with FAI, then your FAI architecture gives you a bunch of places where you have freedom of means. But if you start with freedom of means, and then you try to put an FAI architecture on top of that, then I'm suddenly much gloomier. If someone who has already built a big conglomerated AI system, then the individual pieces might be something you can take out and put in an FAI system. But the whole overall architecture, you probably can't slide FAI on top of that.

Dario: Let's say that someone, and probably in practice it will be Google maybe within 10 or 20 years, builds a perceptual system, that has similar visual perceptive ability to the human cortex…

Eliezer: Is it neuromorphic? The reason I'm asking is because of the very plausible hypothesis that most of the entire cortex is doing more or less the same thing. So if they got it the human way, than there's a much stronger expectation that it generalizes to a bunch of other things than if they got it in a non-human way.

Dario: Right. In practice, we won't know how close the way we got it is the human way until we try it in other things. But let's suppose for the sake of argument, that it was not particularly similar to the human way, in the sense that it was not directly inspired by neurological knowledge. We could have gotten in a way it was relatively similar to the human way by default, or maybe because that's the obvious way to do it...

Eliezer: So we have specialized vision.

Dario: Yeah, we have specialized, but fairly broad visual algorithms that aren't necessarily optimized for a particular spectrum, for particular colors. We have a broad perceptual system that seems to get at typical features in the world.

So we have this module. Google releases it, and asks people to pay for it. You can use this thing to fly a plane. You can put it into a robot, and the robot can navigate around. Is this module something you are comfortable sticking into your Friendly AI, or is the fact that it was done in a way that's very, sort of like the neural network, it's very statistical… In your opinion, is that already unsafe?

Eliezer: So you added an extra stipulation there, about using neural net sourced statistical learning. If you have a relatively well-understood algorithm, whether there is separation of learned parameters, but the algorithm itself is operating in the sort of, understood, iterative, lawful basis, then there's a bunch of separate data that's being modified and it’s an epistemic thing and it’s not making policy decisions… I'm basically happy with sticking that onto a Friendly AI. I'm not sure you can do very much with it.

If it's the cerebellar algorithm, or motor control, and we are trying to use the cerebellar algorithm to control internal thought processes in real time, which is probably a lot of what the cerebellum actually does — at that point, I start to become considerably more nervous.

If it's the visual algorithm… It's a goop of evolving programs so that instead of this statistical learning thing it’s actually this big black box and we don’t have any idea what’s going on there, and occasionally things that Google calls visual programs break out of the sand box, and Google has only taken a frozen version of it and released it to the public, then the original boiling soup of mutating algorithms I’m much more nervous about sticking inside a Friendly AI.

Dario: I'd like to hear what Jacob thinks about that. All the things that Eliezer described at the end don't sound to me like what Google is very likely to do because they know it's not a good idea. Not for safety reasons, but maybe simply design issues.

Eliezer: There are very few modern AI algorithms at all that match that description. Eurisko, maybe. Eurisko is rare. All the statistical machine learning stuff, fine. All the policy learning stuff, if you are not using it for policies about internal self modification. I'm a little nervous about what it's going to do to the outside world, but basically fine. But that's at present level of power. If you take present day algorithms, and you put them on a moon-sized computer, I’m much more worried.

Jacob: I disagree with your analogy about how if you have something that outputs policies, or just somehow doing this very simple thing, where you take...I forgot exactly what you said. You said something like you take the same policy that worked for the last 10,000 times instead.

Eliezer: Right. At that point, I start to be nervous about the system.

Jacob: Something that looks like this, would just not work very well. Anything that's very non-reflectively, just taking things blindly and not doing any sort of robustness analysis, maybe it wouldn’t be friendly in some sense but also it just wouldn’t function.

Eliezer: I want to believe that, but the problem is that human brains work.

Jacob: Human brains are extremely reflective...

Eliezer: They are also what worked the last 10 million times. We have very different ideas about what constitutes extreme reflectivity. Human brains are slightly reflective. They are what worked the last million times, and they break all the time.

Jacob: If you take any part of the brain and just cut it out, the rest of the brain continues to work. The fact that this thing that used to be sending signals, is no longer sending signals, does not...

Eliezer: If you take them out of the ancestral environment, and put them somewhere where there is high fructose corn syrup, then…

Jacob: I mean, they don't go *that* haywire.

Eliezer: People go schizophrenic all the time. There are all kinds of insult to the brain…

Dario: But that's not being taken out of the ancestral environment. I don't think there's anything to suggest schizophrenia was rarer in the ancestral environment than it is now.

Eliezer: I'll be very surprised if it wasn't rarer in the ancestral environment.

Dario: Seems very hard to check.

Eliezer: Anyway, most importantly of all, from my perspective, humans are not very self-modifying.

Jacob: I disagree with that as well.

Eliezer: If you were to rate things on a 10 point scale, where zero is a bacterium, and 10 is a superintelligent machine, humans are about 2 on the reflectivity scale.

Jacob: On the reflectivity scale? I would agree with that, if you would have said this was on the self-modifying scale.

Eliezer: Let's go with the self-modifying scale.

Jacob: Reflectivity may be not the best word for what I'm trying to talk about. I think I mean more some notion of robustness. Basically, there's pretty much no decision that the brain makes that's being made by a single localized source.

Eliezer: Well, it's made by the brain.

Jacob: What I'm saying is that within the brain, there's no small computational core…

Eliezer: Sure. The body has a brain, but once you reach the brain, the brain doesn't have a brain. No homunculus.

Jacob: All I'm saying is that, this idea that you have this policy making algorithm, that's going to have suggested a good policy the last 10,000 times then all of a sudden suggest something terrible that screws everything up, that doesn't seem to be how…

Eliezer: So the difficulty is with context changes. The basic difficulty with trying something 10,000 times in simulations and then one in the real world is that it's not an independent draw from the same relevant distribution.

Similarly, the brain is very robust to circumstances that vary successfully. It has some amount of bonus robustness because of accidental generalization, like when you take a big system, and you make each of the parts tolerant against a set of insults, it's not going to be robust only exactly to the exact distribution that produced your trial cases. There's going to be some amount of additional robustness built in. At the same time, you can't edit your own code.

The basic issue I'm worried about, with respect to super-intelligence, and this probably needs to be some kind of additional thesis, like the context change thesis…

First, when you are doing self modification, each self modification is a potentially critical failure. It's a code change. It's not just like learning a new thing within a clearly defined set of algorithms. Every internal change that has sufficient freedom needs to be constrained within that freedom enough not to permanently warp the system.

The second issue is the degree to which operating in superintelligent mode will change a bunch of parameters that you never got to change prior to superintelligence.

Jacob: My impression is, that you don't really trust any sort of statistical guarantee.

Eliezer: I trust statistical guarantees to do what the guarantee *actually* says. I'm suspicious of their ability to broaden beyond that. If there's a barrel that's actually independent and identically distributed and you do 1,000 trials, I totally trust the thousand and first trial to have a less than 1-in-1000 chance of failure.

Jacob: In addition, implicitly, you don't think that it's possible to create statistical guarantees that could be guarantees about something as complex as computation? I get the impression you don't think it's possible to get a statistical guarantee that an AI will behave well.

Eliezer: I'm really skeptical that you can have this big complex system, whose parts have been tested, and you do a well-defined statistical calculation, and you say "OK. The super-intelligence has a 90 percent chance of working. Why didn't you put in slightly more effort to push it up to a 99.99999 percent chance?"

If you can get this big complex thing up to a 90 percent statistical guarantee, it's very unlikely you couldn't push it a bit further to a 99.99999 percent chance.

Jacob: Why don't you think you can get that?

Eliezer: To a 99.999?

Jacob: Yeah.

Eliezer: Oh I'm happy with that. I'll take that.

Jacob: My impression is that you don't think that's something you can get...

Eliezer: That's totally what it looks like when you have a Friendly AI. It's just that there's a few more nines, because, why not throw on the extra effort?

Jacob: But it seems like you're advocating some kind of theorem-prover instead of some statistical testing procedure.

Eliezer: Oh, I don't believe that your statistical testing procedure can even get you up to 90%.

Luke: So I think our model is that same thing that you do when you are sending a robot to Mars and it's really expensive if you fail. Except the failure with Friendly AI is a lot worse, and worth putting more effort into.

You are doing statistical testing on some parts, you are proving certain properties of the system, you are building in hardware redundancy so that cosmic rays don't throw things off. You are trying to build in as much robustness and safety into the system, as you can.

Statistical testing is a part of that, but especially for superintelligence it's not as useful as in other cases because there's a massive context change for superintelligence.

Eliezer: I feel like the Mars' probe, or space shuttle software analogies, they might be useful as intuition pumps for people who haven't yet understood what goes into verifying a computer chip or sending a probe to Mars or writing space shuttle software.

It's actually a bit more difficult than that, because of the context change problem and the self-modification problem.

Luke: A *lot* more difficult than that, yeah. I'm just trying to illustrate some small part of it, with reference to existing systems.

Jacob: The thing that I don't like about the Mars analogy is that, in some sense, the Mars rover was built by, in some sense taking advantage of the fact that it's going to be in this extremely restrictive environment. I don't think the Mars' rover thing generalized very well to more complicated settings.

Space is really nice, because there's not that much stuff.

Luke: I agree, but I'm saying the way it doesn't generalize, is that you need to do a lot *more* work to make a superintelligence safe.

Jacob: Yes, I agree with that. All I'm saying is that, part of that more work is not relying entirely on purely, formal theorem-proving software.

Luke: Even the Mars rover doesn't, though.

Jacob: A lot of it doesn't. A lot of stuff is based on assuming that you are not going to get perturbations to your system by more than a certain amount. I don't actually know the rest of what goes in to the Mars rover.

Dario: Jacob, basically what you are saying is that, you actually think that...We all agree building a general AI, trying to predict it's behavior and make sure that it falls within certain constraints, is much more complicated than Mars Rover. The question is, should we react to that complexity by trying harder to get formal proofs and guarantees in the hope that this will give us practice, 99.999 percent guarantees?

Or does it, instead, mean that we should move away from formal guarantees and think in more statistical terms? You're claiming the latter and Eliezer is claiming the former, is that a fair summary?

Jacob: That's my impression of what the debate is about.

Eliezer: I must add I don't think my attitude can be described as in favor of formal systems the way that we do them now. First-order logic is not a good fit for the environment.

Dario: But you want some kind of formal system right?

Eliezer: So, reasons for hope. First, human mathematicians do much larger and more interesting things than present automated theorem-provers. Second, logic is a bad fit to the environment. Other things, which are better fits to the environment, such as Bayes nets, or relational probabilistic models — once you view these things as epistemic objects, they can themselves be described by logic or formally or whatever. You're not trying to have a logical description of the external environment and prove theorems about how you interact with it, but the external environment is something you're uncertain about, and it doesn't have the structure that corresponds to first-order logic. Transistors do. And as time goes on, people do more and more proof for transistors and less and less statistical testing. But that's a computer chip not the external environment.

The first thing you might think is there's going to be an environment, and a set of reasoning processes, and your theorems are being proven about the reasoning processes and not about the environment.

Basically, you're going to do Bayesian updating which is probabilistic, then you're going to prove you're going to do Bayesian updating. Though in point of fact you can't do full Bayesian updating because it's computationally intractable. Maybe you can prove you're approximating Bayesian updating, and the approximation is statistical, but every time you do a self-modification, you don't get an additional loss…

You probably can't do that either, but you might be able to prove that you have an expectation of improving your Bayesian reasoning in a way, which on a policy level, doesn't have a conditionally independent probability of error each time you make a self-modification. We have probabilistic reasoning that probably works, but at the core of self-modification, there is something that always works, because it is like theorem-proving.

Luke: In principle this is what they do when they formally verify software in agents today. There are certain core parts of the decision making algorithms that are formally verified but they can't formally verify against the universe.

Eliezer: What they do nowadays is they do the step where you formally prove the reasoning correct, but they're doing something more ground-level than Bayesian reasoning.

Luke: *Much* more ground-level.

Eliezer: So, a nice concrete example: your AI makes your happy, while it's only means of producing happiness is to make you smile, then once you have an AI which can induce smiles directly by tiling smiley faces, it no longer makes you happy. So there was a context change. The behavior that previously produced happiness now just tiles smiley faces.

Jacob: To me, this seems like an extraordinarily dysfunctional AI.

Dario: It has a very brittle model of the world.

Eliezer: What? It *totally* achieved its goals, before and after the context switch.

Jacob: But to achieve its goals, it had to do a bunch of other things in the meantime that involved it interacting with the world, right?

If its model of the world was so brittle that it didn't understand the difference between happiness and a smiley face, it's not going to be able to…

Eliezer: It totally *understood* the difference. You, the programmer, misunderstood what it was doing. You had a statistical guarantee that it produced happiness the last 10,000 times…

Dario: This just seems like a very strangely written AI in the sense that in order to produce those smiles in the humans originally, it has to have a very detailed model of what causes humans to smile, which within that context is what causes humans to be happy.

Yet the programmers also designed its goal to be this incredibly brittle thing that just seems weird.

Eliezer: They didn't know what it's goal was, they just had a statistical guarantee about its behavior based on the past 10,000 iterations.

Jacob: You're focusing on this particular point and you're assuming that everything worked perfectly well up to this point and that this is the one part...

Eliezer: Not perfectly well, but well up to the ability of the programmers to discriminate, given their tools for rendering the internals of the AI transparent. The less transparent the internals of the AI are, the more it contains opaque representations where you can say "Ah, it recognized the last 1000 cats, great" but you can't poke around inside and say "… because that parameter there was set to blah" — then there's more chance for things to break in ways that you won't notice coming because you didn't have an inside view, you just had an outside view.

Jacob: I'm wondering why you pick this as the change as opposed to you know, Google's web page changed from google.com to google.us, or whatever…

Eliezer: If you don't understand how your policies are formed, then your policies are formed by consulting google.com, which worked the last 1000 times you tested it.

Jacob: What I'm saying is that, when you go around the world, stuff changes all the time. It's not like the world is completely static and until I become a superintelligence and now all of a sudden, the world is no longer static.

I agree, that it's much less static once you become a superintelligence, but if you don't already have an architecture that's robust to things changing, then you're just not going to last…

Eliezer: I expect a bunch of *different* things to change, from the sub-human intelligence regime to the superintelligence regime.

Jacob: But why do you think these are fundamentally different problems?

Eliezer: On a sufficiently fundamental level, these are similar problems, which is why there's any hope whatsoever.

Jacob: I guess, what I'm objecting to is that all of your examples rely on an AI, where, the way you're reasoning about it, is you're looking at the surface level of "X", and that's not even going to get you to sub-human intelligence level…

Eliezer: If you specify for me an architecture that you think is deep enough, I will break that one, too. Again, the difficulty from my perspective is context change to the programmers. It's sort of like how if I thought they were going to be a bunch of AI disasters of gradually increasing successive magnitude that people could plot on neat charts, gradually leading up to a superintelligence related disaster, then I would be much more optimistic of our civilization handling it well.

I just don't expect civilization to encounter problems from AI of the sort that happen after you have a self-improving AI go Foom.

Dario: How wide do you think the window and scope is for this context change problem to produce a failure in the AI's model of the world that doesn't also completely cripple it? Do you see what I'm sort of objecting to here?

Eliezer: Model of the world failures are not what I'm afraid of. What I'm afraid of is "programmer understanding of the AI" failures.

Dario: Why are you not worried about world model failures?

Eliezer: Because drastic model of the world failures either correct when you do Bayesian updating or if you can't do Bayesian updating, your AI ends up as scrap.

Jacob: Do you imagine that values are going to be hard coded even though epistemic state is going to be learned?

Eliezer: I imagine that there's a meta-utility function that is transparent to the program.

Jacob: Is it hard coded in a fundamentally different sense then its belief-updating procedure is hard-coded? Why do you see values as so different from beliefs?

Eliezer: Because you can update beliefs using Bayes' rule, and that's not how values work, at least not natively. If you have an AI that's maximizing paperclips, it doesn't update to maximize something else.

Jacob: Sure. But if you have a sufficient level of indirection about…

Eliezer: I could introduce meta-utility functions that yield utilities in a model-dependent way. Then there would be Bayesian updates that affect the utility function, but something like that is to be explicitly coded.

Jacob: Would you agree or disagree that Friendly AI is about getting the meta-utility function right?

Eliezer: I expect the work to be 10 percent of getting the meta-utility function right and 90 percent about getting the rest of the AI right so that the meta-utility function actually continues to work throughout self-improvement.

Jacob: If you had the meta-utility function right, why do you now think that utility remains more fragile? Why do you think values are remaining more fragile than beliefs?

Eliezer: You mean to failures of self-modification?

Jacob: Yeah. If you get the meta-utility function right and now all that's left is for it to correctly infer the particular concrete utility function it should care about, it seems like that is on the same level as the rest of its epistemic state. Failures that affect that could also affect the epistemic state.

Eliezer: Basically, it's about: does reality bite back? There's a very broad space of algorithms where if you make something interpretable as an epistemic mistake, you'll be able to notice. Like, you think that things fall upward, or a cosmic ray flips a bit and makes the AI think it should systematically update in the opposite direction to Bayes' rule, then the AI collapses in a pile of parts.

The basic difference between beliefs and values is that with beliefs, reality bites back, and the healthy parts of the system can notice the unhealthy parts. Whereas if you change 10% of the utility function, it's not obvious to the other 90% of the utility function that the 10% is wrong, because reality doesn't bite back.

Jacob: I have one other question. Is your worry that an AI somehow fails to have an accurate causal model of the world? Or is it something else?

Eliezer: My worries about superintelligence do not come from a worry that it will have an inaccurate model of the world.

Dario: Do you think this is something that may happen? I could imagine this happening for instance with AI's that were trained in some virtual world or something, for example, that were built in the context of some kind of game. That somehow managed to break out of it, that in addition to there being this context change with respect to values that might very well happen, particularly, if the AI wasn't designed really well and wasn't designed to be safe, is that it might have fundamental misapprehensions about how the rest of the world works due to it not being sufficiently general or having certain assumptions that are hard coded and people aren't aware that they're hard coded?

Eliezer: That doesn't sound like a very world-ending AI.

Dario: No, but the claim that I and maybe Jacob are implicitly making is that maybe a large fraction of the potentially world-ending AIs end up making this mistake first so and so we aren't threatened in the first place.

Eliezer: On a certain level of abstraction, that's exactly what's happened for the last sixty years. We have a bunch of weak little AIs that our too weak to destroy the world because of their epistemics.

Dario: Yes, but the point I'm making is that there may be some correlation between the AI being safe and it and it actually working.

Eliezer: Yes but all of the correlation is coming from the fact that to be safe it has to actually work. There's no corresponding "to actually work it has to be safe."

Luke: I've heard you say before, Eliezer, that *some* of the problem comes from avoiding epistemic failure, just not most of it. If you designed an AI architecture and you got the meta-utility function right and you got stable self-modification right, but somehow in that system, it was a hard coded to do induction by an approximation of Solomonoff induction, then it could never represent certain hypotheses about the world that might be true. And you wouldn't want that.

Eliezer: There are some exotic cases like that where there's something that looks like an epistemic question and you don't know how to update about it, like whether quantum suicide works.

Jacob: So I think I agree with you that the meta-utility function seems like a pretty real problem. I think on some level there's a lot of incentive for people to get it right even before we have superhuman AI.

Eliezer: Great! Who's working on it?

Jacob: I don't know that anyone is working on it.

Eliezer: Okay. I believe that's still going to be true in 30 years.

Jacob: I think that as people want AI to accomplish increasingly complicated goals that are very difficult to specify by hand, they're going to try to come up with new ways to specify goals not by hand. Also, I guess people are *somewhat* working on this. There's stuff in reinforcement learning where you'll get some expert to drive their car around, or play video games or something. Then, their reinforcement learner's job is basically to infer what goals that person had that would cause it to act in that way.

Eliezer: I have not heard of that as a reinforcement learning paradigm. Is this like perhaps somewhat less general in practice than what you just said out loud?

Jacob: One example is trying to navigate this vehicle, and you posit a utility function that is within some parameterized family of possible utility functions and these parameters are based on various features of the domain...

Eliezer: Okay so it's a well-known causal domain and a parameterized family of utility functions. I can see it'd be possible to infer which member of the parameterized family of utility functions…

Jacob: Yeah, so it's something like: smooth high-traction road is really good, bumpy road is not so good, slippery road is also not very good, especially if you're trying to go fast, and not-the-road is really bad.

Eliezer: Does it come with pre-classified road and the parameterized utility functions then include different types of road? If it's independently formulating the concept of slipperiness, I'm very impressed.

Jacob: It's pre-classified. I'm not claiming that this is anywhere close to solving the general problem that you want to solve. I'm merely showing you that people are working on it, and it's not clear to me that this is substantially more simplistic than most epistemic

AI algorithms are. Epistemic AI algorithms also have some parameters, family of possible beliefs…

Eliezer: So again, if I thought there was a gradual commercial pathway to CEV then I would move CEV off to the side. Actually, it's kind of already off to the side, because it's too difficult to do right now, but anyway: I'd move CEV off to the side and be like, "OK, commercial AI efforts will probably produce understanding of this."

Jacob: I don't think that there's a commercial pathway to CEV necessarily, but there is one to meta-utilities…

Eliezer: Right, so I don't think meta-utility is enough. My anticipation is more along the lines of commercial-grade AIs stay well below human levels of abstraction and then you need a meta-meta-utility function at something approaching the human level of abstraction.

If I predefine the difference between slippery and firm road, and we're going to learn that firm road is a good instrumental idea using reinforcement learning, sure. Independently generalize "slippery" as an abstraction that covers what we'd be doing, software that monitors what we're doing to a document three times, and tries to generalize a macro from it. There's a commercial incentive to develop that. It might well happen before the end of the world.

Jacob: People are working on it already. There's stuff that monitors what you're typing into some document and it tries to generate hotkeys that are specialized to you, and auto-complete that's specialized to you. These aren't just memorizing words, but looking for patterns…

Eliezer: What about sub-goals and tactics?

Jacob: We're not doing sub-goals and tactics. We'd like to at some point but it's a really hard problem.

Eliezer: Indeed. Even there, before the end of the world, I well believe that you might have the computer system where you repeatedly click on this or do that and it's like, "This person is renaming all instances of variables containing the name Fred to George, while preserving capitalization. I'll just create a macro to do that."

Jacob: To me, the relevant thing doesn't seem to be how complicated of a task it is that you're inferring, but how indirect the supervision is. For this driving test, you can have someone go around and drive your car. In some sense, they're simulating you and making decisions for you for a while, and now you have to come up with rules that replicate that. If it's something less concrete than a human drives a car, and then an AI tries to drive the same car, there's not this simple...

Eliezer: The lack of one-to-one correspondence I see as two separate things. One is can you do consequentialism, and one is can you do cross-domain epistemic generalization.

It's like… Has everyone played "Super Mario Bros."? The original Nintendo version?

Jacob: Yeah.

Eliezer: Noticing that you jumped on this Goomba and you ate this mushroom and you dodged this turtle, and all of these things conduce to a sub goal of survival in the game, which was not pre-parameterized. We've got one problem of epistemic generalization, which is forming the concept of survival, and one problem of reverse engineering consequentialism, which is noticing that all these things have the same consequence and postulating that it was the consequence the user wanted.

Jacob: First of all, Super Mario has been played via reinforcement learning.

Eliezer: Right, but not to have a completely unsupervised AI that had never heard of video games watch you play Super Mario, and thereby induce that you wanted to survive, or wanted a high score.

Jacob: This aspect of things seems to me to be an epistemic aspect.

Eliezer: It involves forming a new concept…

Jacob: If you have an AI that can't form new concepts, then you're pretty screwed.

Eliezer: Sure, but some concepts are more complicated than others. I agree that superintelligences would have arbitrary abilities to form concepts but I'm not sure that commercial AIs have this smooth pathway leading to super-intelligences because of the foom problem, a sudden jump in the sophistication of concepts you can have.

Jacob: I guess, I don't see that as particularly relevant here. My impression was we already agreed that we're not worried about AI having crappy epistemics. We can assume that they're going to form good concepts. Just because you need to form a complicated concept to get things right is not the problem.

Eliezer: Well, it has to be a natural predictive concept. "Whether Terry Schiavo is classified as alive or dead." That's a value-laden concept. It's an instrumental category that's relevant to judgments your utility function makes. It's a utility-laden classification. It's not like "Phlogiston vs. Fire," or like making a mistake of including Mesmer's animal magnetism in with electromagnetism.

Jacob: I agree with a version of this. If you have to form a concept where there's not really any direct supervision that's telling you to form the right concept that makes things hard.

Eliezer: I think super-intelligences do great at unsupervised concept learning. For supervised moral concepts, I don't think they have natural abilities to get right. I think that's this very difficult, very deep problem.

Jacob: I would cast this differently. What's actually happening is it's not actually unsupervised. There's a lot of supervision at the object level, and that's slowly rising up to the much more abstract things. You could call that unsupervised, but I'd call it more "weakly supervised", where there's many layers of abstraction between the data about the world and…

Eliezer: But none of the concepts are value-laden. It's all like, "I saw three instances of this. To what extent should I guess that the fourth instance is like it?"

Jacob: Right. I think we agree on this, but I just want to make sure that we're using language that we can agree on. What's hard about values is that there isn't really this trickle up, where there's this concrete information about the world that can, even indirectly, inform values.

In some sense, by itself, information about the world cannot tell you what values you should have, and so there's no supervision at all unless you can somehow tie values to something that you can get feedback about the world from.

Eliezer: They're unnatural categories. Natural categories are the ones you form when you just see three things and you're trying to predict how much the fourth thing will have some observable regularities. Unnatural categories are things like, "Is Terry Schiavo a person?"

I'm not sure I'd cast them in terms of trickle up, I'd cast them in terms of, "Do I get this for free when I'm forming natural categories in cases where if I form the wrong natural category reality bites back against a mistaken prediction?" A bunch of categories I can get from that for free, but they're not "Is Terry Schiavo a person?" categories.

Jacob: I think an AI would form these categories, if nothing else but for the purpose of predicting what humans will do. I think an AI will have categories in its head for all these things you care about. There just won't be any connection that ties that to actions. It'll just be like, "Oh, that's interesting"...

Eliezer: No, there's a category of what humans agree with, but it'll include all the ways you fool humans because that's part of the natural category. Things that make humans say, "Yay."

Luke: Like framing effects.

Eliezer: Right, and so on. If it doesn't have framing effects in there, if it doesn't have spoken sensitivities and so on, it's going to be making bad predictions about what humans will approve of and reality will bite back.

Holden: Do you actually think that there's no difference between things that make humans say, "Yay" in some superficial sense, and what a reasonable extrapolated human would call flourishing. There's no naturalness to that category?

Eliezer: There's no practical reason to compute CEV if you don't already care about it.

Jacob: That seems wrong to me.

Dario: Yeah, that's what seems wrong to me. I think that what's going on in the crack addict's brain is very different from what's going on in the brain of someone who's living what we can't really quantify, but would describe as, a fulfilled life.

Eliezer: If you have two actual people, the models of those two actual people will be different. It could be the crack addict and the fulfilled-life person, and they have different models.

It's much more likely to involve things like, "is this person addicted?", and "what's happening to that person's dopamine neurons?", and "does this person have verbal philosophy guiding them?" There's no natural distinction between crack addict and fulfilled life.

There's natural distinctions between how these two people think as part of natural descriptions of decision functions, but to pick a bunch of decision functions between different humans and draw a line between them that constitutes the difference between eudamonia and non-eduamonia, that's an unnatural category and there's no predictive thing forcing you to do it because once you understand them on a low-enough level you can already predict their behavior.

Jacob: So I think we all agree that predicting what humans will do is something an AI would care about. Is that right?

Eliezer: Yes. It's like, what do you have to know about humans in order to make as many paperclips as possible?

Jacob: You have to know how to remove the humans.

Eliezer: Exactly, and that requires that you know how to fool them, and what they tend to believe under different circumstance, you know about how they form beliefs, you know about how they form policies. It's maybe useful to you to know a lot about human disagreement…

Jacob: My thesis is that if you have a bunch of humans and you're trying to figure out what they do, then simulating them on the level of figuring out what will cause them to agree with you, and treating that process as a black box that you're just trying to infer, seems less effective than separating out the different reasons they might agree with you.

They might agree with you as a result of rational deliberation, or because you made an emotional appeal, or because you zapped their cerebral cortex…

Eliezer: I agree that these are all natural categories in the sense that there's a bunch of things in them that are highly similar to each other. If you're smart, you probably modeling them at a lower level than that. Not all the way down to individual neurons. But I would totally expect that if we currently had a super-intelligent paperclip maximizer, then even if it wasn't a model working at the level of individual neurons, it knew a whole bunch of cognitive neuroscience that we don't even know about yet.

Jacob: So what I would claim is that among this large collection of natural categories that we're combining to get human's agreement to let the superintelligence paperclip the universe… One of those things is going to be what I would call my "actual desires", and its going to be one of maybe a thousand different things and the AI might not think that it's anything special from among the other thousand things. But it will be a concept that exists.

Eliezer: So to be concrete, prospect theory (losses are more painful than gains)… I propose that a paperclip maximizer has no reason ask a question of, "What would this person's utility function look like if not for prospect theory?" Which is a very basic elementary question that I would ask.

My worry is… Going from prospect theory to utility functions is actually a very primitive example of something you might have to do for Friendly AI. It's not clear to me that a non-superintelligent foom from commercial AI would get to the level of wanting to reverse engineer the prospect theory to utility functions step, even though that's very basic stuff.

And if you build a generic AI that's just maximizing paper clips or something, it has no natural category corresponding to what people want in the sense of their utility functions rather than prospect theory.

Dario: Is the neural and computational content of fulfillment a natural category relative to the neural and computational content of being a crack addict? Do you consider this category having any naturalness or as being completely arbitrary?

Eliezer: I don't think the category you're trying to point to is natural. People are feeling different things but insofar as "best natural category", it stays within the natural category if you reach into their brain and make them feel good.

Jacob: Your claim is that this doesn't stop wire-heading in the brain?

Eliezer: Wire-heading of humans?

Dario: That what doesn't stop wire-heading?

Jacob: I think what Eliezer would say, and correct me if I'm wrong, is that there is a natural category that looks kind of like fulfillment on some level, but it includes all sorts of things that you don't want. Like saying 'yes' to wire-heading. Is that right?

Eliezer: Yes.

Dario: For any short term neural computational category, that's going to be the case. There's a longer term category that refers to what we mean, that describes something of greater spatial-temporal extent. The question is whether that's natural or not? Are you saying that it's not?

Eliezer: Why would a paperclip maximizer need to know about that in order to manipulate you?

Dario: Maybe it'll win so quickly that it doesn't need to know about that. But something that was interacting with humans over an extended period of time might need to know that.

Eliezer: Why?

Jacob: Actually, I'm not convinced that continuing along this line vs. popping up a couple of levels is the best use of our time.

Eliezer: The whole fundamental question about, "Why is FAI hard?" has awful lot to do with a lot of this. I do suspect that diving into the details and seeing that lots of specific things don't work out might help people see why FAI is so hard.

Jacob: I'm personally interested in popping back up to the level of meta-utility or maybe one level higher than that.

Holden: I want to just throw in a comment that you guys are free to ignore if you think that I'm just going down the wrong path. It sounds to me there's kind of a presumption throughout the last -- I don't know -- 10s of minutes of exchanges that we're talking about an AI that we all agree we need to say, "OK, here's your utility function; maximize it in the world and hit go."

As you know, I'm skeptical that is going to be a situation we're in -- or it may be a situation we're in, but it'll be a situation that follows a lot of opportunities to do testing and discovery that are just impossible now. Dario and Jacob, does that sound right that we've kind of background assuming that? Do you think we should be background assuming that? Am I right to be kind of noticing that this conversation seems to be taking place on terms that are more presumptive of this, "We have to get this utility function right. Let's discuss how we're even going to get it right" than there ought to be?

Jacob: I agree that presumption has been there. I've been mainly trying to choose sub-topics that don't hinge upon whether we take this presumption or not.

Dario: I was thinking a little about that too and was thinking that it might be good to list out the ways of getting something that in the language of Eliezer's "value is fragile" hypothesis is within the cluster of human values. Certainly, one way to do it is to build this meta-utility function that correctly represents human values.

Another way to do it, which I interpret Holden as alluding to, is for the A.I. to have a feedback and readout process that sort of is designed in such a way that it's probably not manipulative, and that constantly draws bits from whatever this human value thing is by explaining output and getting feedback from humans on whether if the steps in the output make sense.

Then you can think of other possible ways refer back to whatever it is about human brains that defines human values. I was wondering, one, do you think that's an exhaustive list of the ways to do this? I guess I could imagine a third way which is that the way in which the design of the AI evolves over time is such that you're not that far from the human cluster already.

Eliezer: So the third one, I am very skeptical of...

Dario: Alright, but that not so much the point as: do you consider this a completed enumeration? Do you think there are other things in the category, and can you comment on why you think the meta-utility approach is the best way?

Eliezer: From my perspective, the meta-utility approach is like where it starts. Contrast the problem not being able to see what a theorem-prover looks like given infinite computing power, and designing one that operates with human assistance.

For my perspective, it's very plausible to me that there's an interim state where you think you basically know what CEV looks like. You are feeling very unsure about your ability to do it in practice in real life on the first shot, and so you sort of weaken your design and complicate it to have a bunch of humans in the loop.

And then you're faced with some interesting moral challenges and it seems to me that any capability you have in this state should not to be used to immediately go off and optimize the galaxy but should be used to get out of that state.

So leaving aside the unnerving moral problems, the other major reason is that I'm incredibly skeptical of people who think they can avoid solving the basic problem or fundamental problem and don't know what the meta-meta-utility function would have to look like and think they can approximate it with ad-hoc hacks. It's like trying to build a theorem-prover without understanding how an ideal theorem-prover would work given infinite computing power. I want to know exactly what kind of problem we're solving, I want to be able to make strong statements about the AI not manipulating the humans, etc.

The responses that say "You won't have to solve all these basic philosophical problems and you won't have to build something that will act independently because humans will be in the loop," I want to say "You're doomed, it will manipulate you…"

Holden: I would go between the two. I would say that we have to solve the philosophical problems, but were going to have a much better tool to do so when we've made more progress on AI, and trying to solve them now just seems kind of weird to me.

We just know on our way we develop a better AI and that gives us better tools to be able to attack these philosophical problems. We will have to solve them. but it's like the correct laboratory is the laboratory where you're able to poke your AI and not our current laboratory of pure theory.

Jacob: I don't think that's what Eliezer is saying. If I understood correctly then I think I agree with it, though maybe we disagree about which parts are hard to do or easy to do.

So, his argument is that "use humans to supervise the AI" is not a concrete plan, it's not actually obvious how you would do that. He's not saying that we need to solve meta-ethics right now though.

Holden: Let me think, is that what I though the proposal was? That's sort of what I think the proposal is, at least, there are parts of the proposal to which I would have the kind of response I gave.

I think the things Eliezer wants to solve now using pure theory are things that seem to me like to the extent they are important would be better solved in the kind of future laboratory I laid out.

Jacob: I think what Eliezer wants is a concrete proposal for how you would put humans in the loop in a way that would cause the AI to like update its values. What would you even do?

Eliezer: Yes, and then I'll be able to point out how it will kill you instead.

Dario: So let me just throw out an example, playing off Holden's Google maps example. One thing about Google maps, aside from the fact that the map itself doesn't store a lot of internal state. Like if I'm here and ask Google map to take me to Palo Alto and then, I go a mile and ask Google maps to take me to Palo Alto.

Google map will compute A* the first time, and then I go a mile and it has a grid that looks slightly different. Then, it could recycle some things but it's not very difficult for it to just compute A* from where I am and the fact that I went a mile from where I was going wasn't that important.

There's some weird sense in which it lacks internal state and it can sort of be shut down and booted back up again.

Eliezer: Because a certain kind of computing power was cheaper than the programming time needed to figure out how to save the computing power.

Dario: Right, so one can imagine… and maybe it's not possible because the AI needs to do too many things without human beings being in the loop and it's too hard for it to print out its internal state. But one thing you can imagine is some kind of restriction (to avoid manipulation) on the number of feedback loops you are allowed to use for a certain AI before you have to significantly change the design or something.

So I have this AI that, in Holden's language, prints out a bunch of steps I can take, what it thinks will results from that step, and what I can do next, and you can worry that I go through too many iterations of that, then the AI is doing something manipulative.

Eliezer: If you're doing long plans, the plans have to be coherent. Building a 747 is not like driving across town. It is more difficult to build a wing and ask an AI to design the rest of the plane giving the wing unless it basically gives you the same…

Jacob: Are you saying that for certain tasks like navigation they are so malleable you can do this? But more complex tasks are just fundamentally difficult and an example of such a task is: I can't ask an AI to build the wing, reset its internal state, and ask it to build the cockpit. These things are interrelated design-wise.

Dario: To what extent can we decouple these things? One example of a safety precaution would be to make the stored internal state as small as possible. That doesn't guarantee non-manipulation but reduces the scope over which manipulation can occur.

Eliezer: That's not obvious to me. If you can build a wing and reboot and say "design the rest of the plane given this wing and get a functioning plane" then the fact that you are discarding the internal state seems to have diminished greatly in relevance. Basically anything it was trying to do to manipulate you when it computed the wing would be recomputed after you discard the internal state.

Dario: You're saying it's storing its external state in the execution of its plans in the world or something like that.

Eliezer: No. Whatever part of the state causes the manipulation in the first place will cause the manipulation the second time.

Jacob: So Dario, I am not sure exactly what your trying to say. But I get confused when you talk about small internal state...

Dario: Yes.

Jacob: Because anything learning about the world is going to have to keep a nontrivial amount of internal state unless it turns out the world is extremely simple.

Dario: The point of this proposal is to ask whether there are, especially when you have something running with feedback, whether or not there are points at which you can interrupt or narrow the feedback in such away to make manipulation less likely.

I don't know that I am being articulate about, or have even thought of the right ways of doing that, but I would be surprised if there wasn't some ways to design your system to significantly reduce these risks.

Eliezer: The basic problem is the absence of manipulation seems like a non-natural category and manipulation is a convergent instrumental goal.

Luke: Yeah, it seems that you would know something kind of like CEV in order to be able to discuss what counts as manipulation.

Dario: I think the more natural category here, rather than asking whether it’s manipulation, is: is it able to form a feedback loop in which the behavior of the human itself in responses to the AI is part the AI’s plans? The AI is thinking about how human will respond to what AI says instead of factoring that out and thinking about the rest of it. That's what seems dangerous to me.

Eliezer: The basic reason why I have worries about this is because of the number of things that seem like convergent instrumental motives, or the number of things it seems epistemically natural to reason about.

Jacob: I think the difference in your mental models is the following: let's imagine the following system. It's just like a question-answering system. It understands English on a level such that it’s able to convert your question into a purely logical-probabilistic formulation. Then it has a big relational database, and it uses this symbolic form to give you an answer.

I think the difference in your mental model is that Dario thinks that such a thing is a pretty plausible thing that we will build, and it actually understands English on the level that humans understand English, and like this system could exist.

Whereas Eliezer thinks that in order for such a system to understand English the same way humans do would have to automatically come with a bunch of additional reasoning that would cause its behavior to be goal-oriented and so therefore have goals in a way that the description I just gave doesn't allude to sufficiently.

Eliezer: I am not sure I followed all of that but if the question is “Do I think we can have a goal-free oracle” then my answer is maybe but it requires and additional 50 years of development work beyond what I think it’ll take to get Friendly AI.

Dario: It seems like you’re saying that systems very strongly want to have goals. Why?

Eliezer: Sufficiently powerful systems have internal things where there is like a bunch of paths and you pick the one that goes to a certain consequence.

Dario: I guess, my question is maybe the scope of its goals. It is entirely possible to internally have a very agent-like, goal-like process while at the top level, it's not really conceiving of itself as doing much more than printing out an answer on a computer screen.

Eliezer: I don't know how to build a system like that. The fact that you are not using goal orientation anywhere inside it means that the system itself is not a system involved in its own design. That is why I was worried about the extra 50 years of development.

Dario: Sorry, could you amplify that?

Eliezer: I don't know how to have a goal-free system that helps to build it. I expect AI to happen as the result of humans and AIs building it together. If it's just the human programmers with no AI assistance, that's harder. If the AI doesn't have goals, how does it have design goals?

Jacob: We already have things that assist programmers.

Dario: Compliers, IDEs...

Jacob: Proof assistants like Coq.

Dario: Can I make sure that what is actually a significant shift that's new to me is being acknowledged? Is it correct, Eliezer, that you just claimed that you're actually not all that skeptical — or maybe you are additionally skeptical of this concept of an AI that thinks in terms of certain goals on a low level but at a high level that's not goal oriented at all. It is not that you don't think this would be safe or wouldn't work, it's just that it would take much more work?

Eliezer: I think it doesn't work and it isn't safe but if we're talking about the system that just doesn't have goals and agency on any level… like, to have goals on a low level but not at a high level makes no sense to me.

Dario: Maybe I didn't put that right, but on some level, it considers alternative and scores them and chooses them. It is hard for me to imagine a system existing that doesn't.

Eliezer: The question isn't "Does it score things and choose among them?" The question is does it score consequentialist things?

Dario: Oh, no. That is not it at all what I am saying.

Jacob: So Eliezer is saying that goal orientation both internally and externally, sounds dangerous. Goal oriented internally but not externally equals incoherent. And goal oriented neither internally or nor externally equals safe but too slow.

Eliezer: Safe but hard to make. Basically this requires the programmers to understand a bunch of algorithms and be able to power the sufficiently so that you can create a complete functioning mind that didn't help you design itself.

Jacob: I should add the caveat that I'm translating Eliezer's statements, not necessarily agreeing with them.

Dario: It's actually new to me that you think this thing isn't necessarily unsafe or incoherent, but that it's too hard.

Eliezer: You can make almost anything if you have sufficiently advanced understanding of cognitive science.

Dario: But you're saying an additional 50 years of development. Which is not saying it's so hard it's impractical.

Eliezer: An additional 50 years beyond what it would take to make a Friendly AI.

Luke: There are a bunch of different approaches we can take if we get to convince the whole world to stop building AI so that we can do it the safest, slowest way possible.

Dario: But so you're not postulating the level of difficulty where we couldn't sit down and spend 100 years to do it. You're postulating a level of hardness that you think excludes it from the critical path because it's harder than something that will happen earlier.

Eliezer: More or less, yes. It should be kept in mind that I usually figure 200. There are very few things I think we couldn't do in 200 years. Build a super intelligence that didn't believe that there was a number between 16 and 18. In 200 years? Sure.

[laughter]

Luke: I just don't know what you mean by that statement.

[laughter]

Jacob: In 200 years Eliezer will show you!

[laughter]

Holden: That would be a great sci-fi movie. They finally exploit its weakness. The number 17.

[laughter]

Dario: I just want to register some kind of prediction that there will be no way we can ever make sense of that.

[laughter]

Eliezer: There will be some kind of fixed point of not knowing about the number, not knowing that there's anything wrong with not knowing about the number, etc., etc.

[laughter]

Holden: [laughs] I think we should not focus in on this. Fascinating as the idea is.

Dario: I can't think of any way to do this that doesn't ultimately by some very complicated route, reduced to renaming all of the numbers above 17.

Jacob: That's cheating.

[laughter]

Holden: 200 years, Dario.

Eliezer: Michael Vassar, once tweeted, "There are some limits to my ambition. For example, I don't think I could rearrange the small natural numbers." But it's just building a coherent, powerful super intelligence that believes the small natural numbers have been rearranged, then you just have to be really good, but I'm not planning to do it myself, and that's kind of how I feel about the Google Maps AI.

Holden: So what I'm hearing is that you believe that the practical route to building AI is some agenty, goal-oriented thing that helps you build it. Or that builds itself.

Eliezer: Pretty much, yeah.

Dario: Maybe the thing to go after here is to claim that the Google Maps AI doesn't help you build itself, that the natural push back based on what Jacob said, would be something like: well, we'll get all these automated tools that will help us in all these ways, and it's not really clear that this is all that much less helpful in helping you build it than is the friendly AI. And you also get, while you're doing it, unlike Google Maps that's doing something very high-level, and so you're going to get lots of pieces of it before and the pieces are safe. You can test them as much as you want.

Maybe the claim is that in practice, you might think that this would make up for any disadvantages that you're describing.

Jacob: I might be biased, but I kind of like the question-answering AI more than the Google Maps AI because it's a concrete instantiation of something that you think could exist that Eliezer doesn't?

Holden: I'm not sure we actually came up with something Dario believes in that Eliezer doesn't.

Dario: The disagreement was just over how long it would take to build and how hard it was.

Eliezer: The Oracle AI makes a more interesting thing to discuss than the Google Maps AI because the Google Maps AI is already a planning AI. Part of the reason I didn't consider the Google Maps AI all that is interesting, is that it didn't seem all that safe if you were to just following its recommendations without really understanding them, and you would.

Holden: You would?

Eliezer: Yeah, you would.

Jacob: I never follow Google Maps without understanding it. Its suggestions are so bad.

Eliezer: The problem is that the Google Maps AGI is already immensely goal-oriented, so you could already be asking it to design pieces of itself. It would be like asking it to design pieces of itself and in an understandable fashion and not having it manipulate you.

Dario: Sorry, are you talking about Google Maps AGI now, or the Oracle?

Eliezer: The Google Maps AGI. The epistemic Oracle where you're not particularly using it to do consequentialist plans. You want the Google Maps AGI to do a bunch of planning for you, but if you don't want it to be a sovereign.

Dario: You don't want it to be a what?

Eliezer: Sovereign. That's kind of wanting it to be both blue and not blue, or wanting it to not know about the number 17.

Dario: The sense in which it's not as sovereign is that you can always stop following its recommendations if you want, and you might be by it and always fail. But if it does something sufficiently bad for you, then you might actually stop following its directions. So the question is, it could be dangerous if it caused you to do a bunch of things that were good and then one thing that was catastrophically bad.

Eliezer: Yes.

Dario: But, if that was unlikely, if your plans veered off gradually, then it wouldn't be nearly as dangerous.

Jacob: You can imagine that it would conveniently cause all people who are supposed to stop someone from getting access to a nuclear stockpile to be in the wrong location or...

Holden: That would just be so detectable.

Eliezer: I'd worry more along the lines of you ask it for a piece of code that accomplishes good in the world, like the code that will be the solution for a protein reverse-engineering facility that humans could use to cure malaria. But it also contains the seed of a small AI.

It solves the same goal, but does so via this extra route that it didn't conceals but didn’t think it was important to tell you about, it gave you explanation, but the explanation was too hard to follow, so you were like “Whatever, let's just do this and cure some malaria.”

Dario: But isn't this what tools for printing out internal state come into play? In order for this to be dangerous, you have to specify the additional thing that the tools for printing out the internal state are insufficiently clear or don't discuss this, or see it as a small detail.

Which means, that the way you're telling it to print out internal state, what you're deciding is important about internal state and instrumental decisions, is really messed up because you think that its plan is, “I’ll make the protein facility and that will cure malaria, “ and its real plan is “I’ll make this small AI and that'll cure malaria and do a bunch of other stuff too?”

Eliezer: Or the internal state was just too difficult to understand. It didn’t clearly label it as “seed AI.” It was the “malaria curing efficiency modulator.”

Dario: But these things do seem like natural categories to me. The difference between an AI that cures malaria and a protein making facility, it's hard for you to believe that the difference between those is some unnatural category.

Eliezer: In a mature theory of FAI, you would have all these places where during development, humans were checking on things and literally, the point of which you press the big green Run button, is at the point of which you have this well-defined calculation which say that the results of leaving the system under supervision of the humans any longer was worse than the result of hitting Run.

Possibly after the humans had screwed up a few times or something? From my perspective, there will be this theory of what you gain by having humans look over this part, what you gain by having humans supervise that part. At some point it would add up to enough assurance that you could let it take action.

Google Maps feels like, to me, a very arbitrary slice through the sort of device that mature theory of FAI might do about what is transparent, what the probability is of mistake, what the probability is of humans noticing the mistake, etc.

Dario: I'm not sure I follow all that, but the part I'm still getting caught up on, in your example, about how you ask the AI to cure malaria up, and it makes this protein thing and also makes a small AI…

Eliezer: Inside the protein thing, clearly label the “malaria efficiency cure efficiency enhancer.”

Dario: At some point in the future, the predictions over what will happen with and without that small and more capable and unconstrained AI will diverge from what happens if it isn't there.

And in printing out its internal state, you are going to want to print out the AI's view of what is going to continue to happen or some probability distribution over what is going to... It surprises me that that thing would not be caught…

Eliezer: So we ask the AI to compute its predictions over the following observable and we have the humans look over them, and the humans will catch things that fall within the human’s classification of error but outside the AI’s classification of error.

That could happen unless somebody, of course, decided that a way to constrain this AI was to prevent it from looking out beyond the one-year time horizon.

Jacob: But why would you do that?

Eliezer: Because someone decided that this would be safer because then it couldn't have any long-term plans.

Jacob: That feels like a strange objection.

Eliezer: I mean, you could always invent these, maybe we could do blah, but everyone has different set of them and they contradict each other. You were just talking about a proposal where it can’t have goals over the internal units, someone else will be like, “it can't look out further than a month.” And someone else will say…

Holden: I'm not proposing here, specific constraints that be added to the AI, I'm proposing a possible design architectures that might make those things less concerning. This is completely off the cuff, right? The point here is not so much to make a particular proposal that if you build your AI this way, it will be safe.

The point is to question whether the things we should be concerned about are really the things that you're bringing up.

Eliezer: From my perspective, the problem is that there is this whack-a-mole game. The name of the whack-a-mole game is: convergent instrumental goals. These sufficiently powerful systems that operate on sufficiently high levels of abstraction, and are learning a bunch of things, and doing general of various types, they have a bunch of convergent instrumental goals that are like these moles that keep popping up. And whenever you imagine a particular mole that pocks up, you can imagine this hammer that comes down to whack the mole, but the problem is that there are 100 of them unless you somehow actually align the values or come up with some kind of amazing solution to the threat from convergent instrumental values.

Luke: If it's at the level of a super intelligence, then it has some capacity to model which types of constraints the humans are going to be able to come up with, or likely to come up with, and then a way to achieve its goals — even if goals aren't some utility function type of thing — in a way that routes around the constraints that humans are likely to come up with.

Eliezer: So one possible way that the whole problem of friendly AI could be easier than I expect, is if it turns out that the AI's concepts end up more naturally transparent than I expect them to be and it's organized such that there's only these few high-level goals that are highly transparent...

Jacob: Why do concepts have to be transparent as opposed to just predictions? Your concepts are not transparent to me and yet, I can still…

Eliezer: We both have a visual cortex, an auditory cortex. We grew up speaking a common language…

Jacob: It seems like an AI that was unable to communicate with humans would not be a very interesting AI.

Eliezer: The part that worries me is that the AI's internal representations are such that all transparency has to pass through some goal oriented process the AI does.

What goals did the AI pursue today? And there are millions of them and you sample one at random, and it's gibberish. And you say “Translate this into English.” And it’s a 5-page explanation because it has to explain what all the concepts mean.

Jacob: Can't you ask it questions about what is believes will be true about the state of the world in 20 years?

Eliezer: Sure. You could be like, what color will the sky be in 20 years? It would be like, “blue”, or it’ll say “In 20 years there won't be a sky, the earth will have been consumed by nano-machines,” and you're like, “why?” and the AI is like “Well, you know, you do that sort of thing.” “Why?” And then there’s a 20 page thing.

Dario: But once it says the earth is going to be consumed by nano-machines, and you're asking about the AI's set of plans, presumably, you reject this plan immediately and preferably change the design of your AI.

Eliezer: The AI is like, “No, humans are going to do it.” Or the AI is like, “well obviously, I'll be involved in the causal pathway but I’m not planning to do it.”

Dario: But this is a plan you don't want to execute.

Eliezer: *All* the plans seem to end up with the earth being consumed by nano-machines.

Luke: The problem is that we're trying to outsmart a superintelligence and make sure that it's not tricking us somehow subtly with their own language.

Dario: But while we're just asking questions we always have the ability to just shut it off.

Eliezer: Right, but first you ask it “What happens if I shut you off” and it says “The earth gets consumed by nanobots in 19 years.”

Dario: If that's really true, no matter what, we should still shut it off.

Eliezer: Really? So if the result of not shutting off the AI is Earth being consumed in 19 years, and…

Dario: You'd have to have very high confidence that you were trusting the AI's concepts, the AI’s values, and the AI’s honesty to you.

Jacob: If you didn't have that confidence then trying to use it as an Oracle is already dangerous. My version of your proposal has you just designing the AI in such a way that… it’s not like manipulation is this thing that you have to cordon off from everything else. There's a large set of things that the AI does not do, and things that can be construed as manipulation happen to be contained in that.

Eliezer: My problem is that manipulation is an unnatural category, so is the absence of manipulation.

Jacob: I think you want to start from the other direction, which is come up with a set of things it's allowed to do, such that you’re pretty happy.

Eliezer: The history of AI development has been one of, at least in the beginning, people would always be like, "Well, let's have it learn this."

And the AI would learn these very alien things. It's getting things to have a human shape is usually way harder than doing them at all.

Jacob: But getting things to have a shape that corresponds actually to any aspect of the world is much harder than getting them to just do something.

Eliezer: What I’m thinking about here is like Deep Blue plays chess well enough to beat Kasparov using nothing like Kapsarov’s reasoning.

Jacob: But despite the fact that Deep Blue is not all that transparent, we’re pretty confident it’s not going to create an army of nanorobots.

Eliezer: That's because the mechanisms it’s modeling are constrained to the chess board and it wasn’t powerful enough to start reasoning beyond that.

Dario: Is Deep Blue’s reasoning really all that non-transparent? Let’s say it’s doing something simple like alpha-beta pruning, it's possible to look at the tree and summarize what it thinks the next move is, what it thinks is the interesting responses to that move are and how it thinks it's going to respond to those interesting responses.

Eliezer: Sure, to some extent. But like I said, if friendly AI is much easier than I expect, it will be because it's easier to do transparency than I think.

Dario: Right, so that actually seems to be the center of what we're debating here. That actually seems to be the linchpin point here.

Holden: Well, there are a couple aspects of it. One is how hard is it going to be to do transparency? Two is how interested by default are people going to be in doing transparency? Because the high level question we're still in, is like: are we living in a world where by default someone is going to build an AI, not know how it works, and say, 'optimize world, go' or for a variety of reasons can we expect that as we get closer to AI there's going to be less…

Eliezer: They don't say, "optimize world, go" they build an AI, they don't know how it works, and eventually it turns the world into paperclips.

Jacob: By the way, this might be a part of the discussion where our presuppositions about whether AI is a utility function maximizer versus something else is actually important.

Eliezer: Friendly AI looks like that. A paperclip maximizer could start out looking like a big soup of heuristics and later transform itself into an agent with something like utility functions and something like beliefs.

Jacob: Well, I disagree with that and I suspect Dario does as well.

Dario: Yeah.

Jacob: I suspect that this might be why we have very different conceptions about how possible a Tool AI or something similar is. I feel like that it's somewhat more subtle to get right than Dario thinks but somewhat less impossible than what Eliezer thinks.

Eliezer: I will sort of parenthetically mention, in response to what Holden said earlier, that if transparency is easy enough, there is not even commercial pressure, there's just: it's easier to program things when you understand them.

Eliezer: I just feel like the history of AI so far has not given us cause for great hope there. I expect to be putting in these huge efforts for transparency that we're putting in *because* we're concerned about friendly AI.

Jacob: So what about historical AI development do you find not transparent? I'm not sure what particular thing you're talking about.

Dario: Also, just to add to the question, could you comment on the extent to which this has changed over time, because in the early days of AI, to the extent that I know the story.,I understand what you are talking about. I'm sure Jacob does as well, but...

Jacob: Wait, I thought in the early days we all used Lisp and it was completely transparent.

Dario: I'm thinking more of these expert systems that...

Eliezer: And reinforcement learning, and backprop neural networks.

Dario: It sounds like, and Jacob knows the field better than I do, a lot of what you are saying seems directed most strongly at the early naive, neural network approaches of the...

Eliezer: If I were to try to describe the abstract quality of human AI development that has stayed stable over time, I would describe it as: effort has been put first into finding effective algorithms, and secondarily into interpreting those algorithms.

That is, an effective algorithm is invented, and then people start looking at it, and try to make its internals transparent. If it's easy to make the internal transparent, people will put in more effort, if it's hard to make internals transparent, people will give up on it… and use it anyway, of course.

Dario: Do you agree with this Jacob?

Jacob: No, I don't think so, although I'm not sure, I'm just trying to think of examples.

Eliezer: I would actually be curious as to what Peter Norvig would say about this.

Jacob: Yeah, I don't know that. So, I think people usually have some theoretical justification for doing what they're doing, and then they try it and it either works in practice or it doesn't. Then, for the ones that do work in practice, like additional theoretical justification accrues over time. It's very interleaved, I guess, is what I would say.

Eliezer: I think that systems which have theoretical justification… that’s not the same as transparent internals. Theoretical justifications have gotten a lot more popular nowadays, because they look cool when you write them up in journals.

Jacob: I thought your explicit statement was that the way people do AI, they just do something that works for them and justify it later.

Eliezer: No, no. First thing they do something that works, and then they make it transparent. Deep belief networks were not invented by finding something that works and then justifying it later, as far as I know.

Jacob: I still think it’s interleaved. You get something to work, and in the process of getting that to work, you need to debug the damn thing, so there's going to be some degree of transparency. Then as time goes on, you both get better at making it work and you get better at figuring out how can we make it more transparent so that we can understand it, and you also scale it up to bigger systems, so they're both more of a need to have automated understanding of how to get it working and what’s going on.

Eliezer: Again, the primary pressure is toward optimizing performance, and there's the secondary pressure of, "Can I figure out how it works? Can I make it transparent enough to debug it?"

Jacob: Yeah.

Eliezer: I guess it mostly is just debugging in one sense or another. “Can I make it transparent enough to have some neat graph in a journal.”

Jacob: Yes, so I agree that this is the problem.

Eliezer: So, decision trees: a single decision tree is transparent, but bagged and boosted decision trees then are less transparent, and you have to invent new tools to look at popular nodes…

Jacob: I'm not sure I think that bagging and boosting decreases transparency that much.

Eliezer: So with one decision tree, you can literally just read it out, right? It's completely human-comprehensible.

Jacob: No, because the decision tree is like some random hyper plane. If you're on this side of this hyper plane, do this thing. If you're on this side of this hyper plane, do that thing. Unless you can interpret what the hyper plane is...

Eliezer: Yeah, and then what the consequences are and why that hyperplane.

Jacob: Yeah.

Eliezer: Nick Bostrom was once asked to write a chapter on machine ethics for the Cambridge Handbook of Artificial Intelligence. He contacted me, and he had to discuss modern machine ethics and of course part of the reason why FAI is a big issue is because modern machine ethics is… there's no pressure to do it, and there probably isn't going to be until the end of the world. The problems are just not very interesting or complicated.

So I really reached, and came up with the hypothesis that, "Well, maybe if you were evolving something that was more like a neural network and less like a decision tree, then a bank would find that it would systematically reject people who had black names," for example. Because even though you forbid it from looking directly at race, it was able to look at something correlated with race.

Then that, basically, actually happened. It might have been with people born in black cities or something like that, but basically actually happened. I really didn't think that was going to be a hit.

Dario: It doesn't surprise me at all that that would happen.

Eliezer: I consider it sort of like extremely crude and basic.

The people constructing the loan approval system did not put a priority on making the overall decision criteria transparent to humans for the purposes of making sure humans were being handed out fairly, and the sort of natural thing happened.

Dario: Based on what you said before, which is: people are primarily trying to get something that works, and then, secondarily, reaching for transparency...

Eliezer: In so far as it's useful.

Dario: Right. But you've sort of agreed that in modern times, the tendency to do this has strengthened relative to 20 or 30 years ago...

Eliezer: I get that impression. I predict Peter Norvig would say that, but I feel my own epistemic state is unstable.

Luke: It might just mean that the earlier systems were simpler because you had less computation. You could just do really simple expert systems or something. That might be the reason.

Eliezer: I feel, also, like back in the heyday, there was all this ideology in favor of neural networks that you didn't understand because they were so wonderful. Rodney Brooks robotics type stuff.

Jacob: Neural networks are still around.

Eliezer: They're still there but they lost, right?

Jacob: What? No, neural networks are the hottest thing around.

Eliezer: But deep belief networks, which are way…

Jacob: They're new branding of neural networks.

Eliezer: Deep belief networks are way less miracle of local training and "we don't how it works, isn't that wonderful?" and much more, "These are single step hop-field nets and let's look at these intervening nodes so we could see what categories the system is forming." The mystique of gradient dissent is far off. People are happy to have things with theoretical justifications and transparent internals now.

Jacob: I agree.

Dario: The second half of the sentence, that I was going to say, is if that's what you think, and therefore, that things would be much better if, say, transparency were first and function were second. That doesn't sound to me like this wide, yawning gap between where things would need to be in order for safe systems to be built and where we are right now. You've commented at one point...

Eliezer: It's a gap of similar size to the way charities would be if charities were built first to be transparent to GiveWell and second to be effective.

Dario: Yes, and this is something that is occasionally achieved with charities.

Eliezer: Right, so it's a *huge* gap.

Dario: Before we argue about how huge it is, let's argue about whether it's larger than some standard, which is interesting. I heard you comment at several times that you think attempts to improve the system of scientific honesty, as Jacob and I and others are trying to do, may actually be a bad thing because they accelerate AI development relative to friendly AI development and give us less time to make friendly AI.

Eliezer: That's a harder case to make with respect to scientific debates in particular than it is with respect to… I've been stating that I felt ambiguous lately when I read news about academic growth, whether positive or negative.

Holden: I guess my point is, even if it's as serious as you describe, on which I mostly defer to Jacob, it doesn't take all that much fixing of the system to fix that.

Eliezer: You think the system is much less broken than I do. Either that, or you think it would take two orders of magnitude less effort to push a change like that through.

Dario: Let me try it another way then. If you look at the difference in the focus on transparency between the AI world two decades ago and now, what multiple of that, in terms of change insofar as one can quantify that, is needed to produce a much higher likelihood of a sane outcome?

10 times? 100 times? Two times?

Eliezer: You don't get FAI just from putting more effort into transparency. You'll only get that if it turns out that the AI problem itself, qua AI, makes transparency easy.

Dario: What I was saying before is that the major disagreement here is about how hard transparency is and how much transparency will be emphasized. Those were the two disagreements.

Eliezer: I feel like there are more disagreements that we have here, possibly about the fragility of value parameter, unless you say that you don't disagree with that, in which case, fine. For example, I've been trying to think of, "How can I put things in more easily-testable terms, with respect to civilization inadequacy with respect to AI?

I feel uncertain, but I think that you probably can’t persuade the field of AI sufficiently strongly of the orthogonality thesis, that is, that you can have a paperclip maximizer. But I wouldn’t be *too* surprised if that could go through, because all the philosophers who don't actually work in meta-ethics basically agree with that. And then fragility of value, you *can't* persuade them of that.

Holden: What do you mean, fragility of value?

Luke: You can just look up fragility of value on our five theses post and say we can't persuade the majority of top AI researchers of that.

Eliezer: It's complexity of value plus the proposition that realized value falls off very rapidly as you lose 10 percent of the complexity. Complexity of value is like orthogonality thesis plus you have to build something complicated in order for it to exhibit human values.

Luke: The example being you get everything right except consciousness and you don't get 99 percent of the value. You get zero percent of the value.

Dario: As a logically consistent thesis, that's correct within its assumptions. Certainly, I agree with this. I'm pretty sure Jacob agrees with it. I would be surprised if you couldn't get the whole field to agree with it just described that way, the idea that there could be a paperclip maximizer and in some mathematical sense, out of the space of all AIs that there are there, most of them are paperclip maximizers.

Eliezer: So that’s the orthogonality thesis and a bit of complexity of value, but when I went to MIT and started to talk about someone who was working on this type theory that may or may not be reflective — I wasn't very sure based on his description of it — the conversation soon goes to, "But why do you want to prove your self-modifications correct by associating it with a utility function? Why do you want the utility function to be stable?” And it went into the standard, “We shall build our children and they shall go their own way and they shall make the galaxy wonderful” kind of thing.

[lunch break]

Eliezer: So, I’ve also been trying to think about that question “What do you allege that you have, that makes you uniquely suited to tackling these problems?” A big part of the answer, of course, is I don't really know. If these skills were crisp enough that I could write them up the way I've written up so many other rationality skills, things would be easier. I would just write it up, I’d set down a list of exercises, and people would do the exercises and build the skill. But this is less like one of those skills and more like the meta-algorithm I used to write those skills.

Holden: So this is something that I wonder. Forget about how you came to the conclusion about your abilities here. How do you believe you compare to academics?

One part of the answer, I imagine, that I would be pretty sympathetic to is, on a very big picture, choosing what matters to think about, you engage with that question and even very brilliant academics don't. They're very brilliant and attack a problem once they've picked it, but they pick problems somewhat randomly.

Eliezer: And you're sympathetic to that because of the obvious relation to cause neutrality in effective altruism.

Holden: Yes.

Eliezer: You have a large amount of experience with some people mysteriously getting cause neutrality, and most people mysteriously not getting it.

Holden: Exactly.

Eliezer: My answer is, sure, that's part of it, and then there's 5 other things kind of like that.

Dario: It's those five other things that it sounds like Holden and certainly me are more skeptical of, because we all agree that the first thing is important…

Luke: What are the guesses? If you just ramble about what you think some of those five things are, what do you come up with?

Holden: And do you believe that within having picked a problem, there are five important things that you really can do better than anyone in academia?

Eliezer: I can name one of them. As far as I know, Paul Christiano still thinks that a good or interesting model of CEV is to ask the AI what a person in the room could think of a hundred years later. What about the person in the room going and saying [inaudible]. Paul didn't think that was an important problem, whereas from my perspective, it's a problem that reveals that you asked the wrong fundamental question.

I've seen over and over again, with respect to the friendly AI domain, that people invent this stuff, and then they don't do what Bruce Schneider would do. People in computer security and cryptography, they will invent an algorithm and then they will try to figure out how it fails and they'll try to figure out how to attack it. This seems like an elementary thing, and its practice seems virtually unknown in what people think about friendly AI. They come up with a system, and they don't ask how this system can fail. People don't seem to take a system like AIXI and be like, "Oh, this system kills the user to hold down its own reward button."

So this seems like a very elementary sort of thing, and I can point to an entire academic discipline where this is trained into people and it's routine, but with Friendly AI almost nobody does it.

Holden: You changed the subject. I wasn't asking...

Eliezer: That's one of the magic things. Part of the reason why I wouldn't expect there to be academic progress in FAI is that I would not expect them to attack their own theories. I don't know why not, but it appears to be so.

Holden: Hang on. I asked you what you think you have that academics don't have, all academics…

Eliezer: No, it would just be a rare *combination* of things.

Holden: OK, I see. So you're saying that there academics that have some of these things, just not all of them.

Jacob: It's less clear than you think it is. My adviser yells at me all the time for failing to criticize my ideas enough, so apparently, I don't have this skill, but my adviser does, at least.

Holden: I also think that fields often have this skill where the individuals don't. There are fields...

Eliezer: Yeah, there's an incentive to criticize other people's theories.

Jacob: You're also being a little bit unfair to Marcus and Paul in the sense that they're not necessarily claiming that this thing is the final solution. At leat, I'm certain Paul wasn't, because I talked to him, and he definitely didn't think that his thing was the solution to AI.

He was putting forth a proposal that had interesting ideas in it, and also had, certainly, many problems.

Eliezer: I remember, when I was just starting out on FAI. I invented this theory, and then I was like, "No wait. That's going to fail because of blah."

Then, I invented something else and was like "Wait, that'll fail because of blah." 10 iterations later, I arrived at the horrible *Creating Friendly AI* thing, which was also crap as it turned out, but it was still better crap than I see other people are inventing.

They don't even do the first 10 iterations that I did at the very beginning when I was just starting out. That didn't get me to something that worked, but the fact that they didn't do the first 10 iterations makes me very afraid of what would happen if they were just left to their own devices.

Jacob: I feel like they probably saw different things than you did. I often go through several iterations of describing things that aren't going to work. I have something that I'm happy with, but then, I show it to someone else, and they raise new considerations.

Eliezer: It happens all the time in science. I'm sure. Has it happened with any theories you have about Friendly AI?

So an additional filter, possibly, is *single magisterium*. You treat the entire world the same way. You think about Friendly AI the same way you would think about any of your other science problems, including that whole that self-criticism thing. That's rare.

Most people, you talk to about that, they go off into a separate magisterium. They don't apply the skills that learned in the lab.

Dario: That might be true, but something I want to object to here is using Friendly AI in examples. If I go back to my list of 50 academics that are really good in any particular field, Friendly AI is not a developed enough field that it has 50 such people.

Eliezer: No, it doesn't.

Dario: Its existence and value as a separate field that people should be studying is part of what's under debate here. I feel like there's circularity to it…

Eliezer: What? No. You should *always* have a single magisterium. You always use the same set of science skills.

Dario: No. I'm not saying you shouldn't. I don't know why you're interpreting what I'm saying as that.

Eliezer: Sorry. How you were responding to my earlier statement.

Dario: No. What I'm saying is… When I ask you what separate skills that you have, and you respond by saying, "I can think really well about Friendly AI in way X, and others cannot." There is certain circularity here, because part of the premise of this whole discussion is how important it is for people to be thinking about friendly AI in the way that you're thinking about it.

Holden: Yeah. My assertion would be that you think you have cause neutrality, which is really rare in academics, and then four other things. My assertion would be if you could get good academics to focus in on friendly AI, not because of cause neutrality, but because of other factors like…

Eliezer: Because cause neutrality is hard. You can't just give people cause neutrality.

Holden: Exactly. Now, you're going to say that applies to the other four —

Eliezer: Yup!

Holden: And I'm going to say that it's not, and that if you can get other good people to work on Friendly AI for whatever reason… I'm hearing these things like self-criticism. I don't think that would be a comparable obstacle at that point, same with separate magisterium.

Eliezer: Plausibly not. I don't really have a good theory. I don't know why it's hard for other people to make progress on Friendly AI. To myself, I appear normal. The observed fact that the train derails at the third stop before making it to the station is much more obvious to me than any particular theory of why it can't.

Holden: We're both agreeing that you need more person-power, and we're both agreeing that you can't throw up your hands and say, "No one can do this but Eliezer." There are two different strategies for getting that person-power.

One of them has much more to do with getting mathematicians interested in what you're doing, and the other has to do with going after more academics in Jacob's field, and we're talking about which one is more likely to work.

Then there's also the question of "Do you need to run the show?" These are important. You think that other people can, or least you're screwed if they can't, make progress on friendly AI, and the question is like, "Which people, and under which conditions"?

We can't just sit here saying, "Well, no one else can do it under any conditions."

Eliezer: So Benja Fallenstein has been iterating on this, and investing lots of time in it for many weeks, and it would not surprise me if he started to develop these skills, because he has been developing experience. So from my perspective, I believe in it when I see it, and don't disbelieve it after I see it. If someone else starts acquiring the ability to critique other people's wacky theories of Friendly AI, or they come up with their own theoretical constructs that don't look completely bogus, then "Yay! Skill has been acquired."

Until that happens, I don't believe that it ought to be there, because of "how can I be that much better than everyone else at this" or any other meta-level consideration. I don't trust the whole meta-level here.

Jacob: I don't think it's the meta-level thing. You seem to be judging other people's FAI ability based on the extent to which they agree with things that you see as completely obvious and settled.

Eliezer: I can understand how it would look that way from the outside. Obviously, it doesn't look that way from the inside.

Jacob: As in you don't think that they are completely obvious and settled, or you don't think you're using these criteria?

Eliezer: From my perspective, I live in this mysterious world where people invent FAI theories, but then don't actually turn around and ask how they could go wrong the way a computer security expert would.

Holden: That sounds to me like just a world where people don't take FAI very seriously, and don't think about it very hard. This practice of critiquing yourself is so standard. It's not something that's rare in any way.

Eliezer: That is one hypothesis. The observation itself is much more obvious than any particular hypothesis about how it happens.

Jacob: What's the observation?

Eliezer: People invent FAI theories, and then don't critique them. Holden's hypothesis that this is because people don't take it very seriously has occurred to me as well…

Dario: So if that's the correct explanation, that the issue is that people don't take FAI seriously, then FAI might deserve to be taken seriously or it might not, but this can't be evidence, or at least can't be strong evidence that you have this rare or special skill, because the thing you're trying to support is that FAI is so important.

Eliezer: I don't care whether I have rare or special skills. That's not relevant to anything.

Holden: It's clearly relevant.

Eliezer: Well, what is it relevant to?

Dario: Are you assuming academics can make progress on this problem?

If you're asserting that academics cannot make progress on this problem, and that you can, then there must be some reason why you can, or perhaps you're asserting that either you or academics can make progress, but that would be strange.

Eliezer: It doesn't have to be a single reason. The reason doesn't have to be anything obvious, and I don't have to claim that I know what the reason is. I take the object level at face value.

Holden: Your only evidence is that Friendly AI doesn't have good thoughts happening on it right now, and I don't think that's very good evidence, because you'll find plenty of cases where there wasn't a field, and people who engaged with it casually said stupid things, but eventually the field gained a little bit of cred and people got much smarter about it.

Eliezer: From my perspective, that seems like a denial of the outside view. There is, also, this theory of "Sure, nobody takes it seriously now, but then later there will come this watershed moment where people agree that AI is imminent. Then, everyone will take it seriously."

Dario: I've offered that theory, as well.

Holden: That's maybe true. I don't believe that's the only way for this to come about. I believe you guys -- assuming certain things about your claims are correct -- can have a really positive impact by raising the profile of this issue in the right academic community.

Eliezer: We can talk about whether there might be a great watershed event in 30 years, and we can talk about whether or not there might be this bright way to approach academics that gets them to take things seriously and then maybe the mysterious cognitive deficiencies will go away.

But I would like to point it out that rather than having some mysterious theory of how I'm special, my background theory is just "This is what I see. I assume it will continue to be that way."

Holden: So what Eliezer sees is that today, when you talk to people about FAI, they don't have very much intelligent to say. I wouldn't really dispute this.

Dario: Right, I don't think I'm disputing that.

Jacob: I guess, I'm concerned that maybe you're being a bit uncharitable towards other people's arguments in the sense that it's already been the case that multiple times throughout this conversation someone has been quoted as saying something that seems really ridiculous, and then when we looked it up, it was less ridiculous than it seemed.

For instance, you say that people all the time come up with silly proposals for FAI that are obviously wrong. Then, Dario was indicated as an example of this, even though he prefaced his thing with saying, "These are ideas that are probably wrong."

Eliezer: No. I was actually talking about Holden's Google Maps AGI.

Holden: I don't know. I'm sure what to say about that.

Dario: So before this conversation, reread Holden's posts, and one of the key points about it was this was not a positive affirmative AI proposal.

This was an example of a set of things that MIRI might not have considered that what was offered up for consideration. It wasn't like, "We should build an AI this way." It's like AI might, or could be very likely, to go down this sort of path instead…

Eliezer: It sounded a lot more like "we should build an AI this way."

Holden: No. You should reread the post. I'm quite confident that's not what I said, and that I took pains to say that's not what I'm saying, because I knew it would be read that way. I'll also say that, obviously, I'm not a person with much background in any of the relevant technical subjects, and I had this idea, and I did try to think of what was wrong with it, I did ask people what was wrong with it. I heard their responses. Their responses didn't make sense to me. So I decided to write it up. I don't know that this is good evidence that I didn't try to critique myself or something.

Jacob: Yeah. I'm not saying that Holden's views on AI are sophisticated.

Holden: But I also said in the post, and I've conceded many times after and before, I'm not sitting here presenting this thing as something that's right. I'm presenting this thing as something that ought to be engaged more than it is.

Eliezer: I find those kinds of disclaimers fundamentally unconvincing.

Dario: I would like to point out that the idea that Holden lacks self-critical capability is on its face absurd.

Eliezer: I agree. GiveWell is extremely self-critical.

Holden: So you're saying in this domain that I lack self-critical ability, which, I don't think…

Dario: My thing is that when he's put like 15 hours of thought into this, and therefore he had some ideas which he recognized and catalogued and being probably naive, but could not see why they were naive. So he presented them to the public for critique.

Eliezer: But it was like, "Why hasn't MIRI engaged with this idea?"

Jacob: Well, it wasn't an idea so much as a swath of area that he was like, "SIAI has not been thinking about this whole swath of area. SIAI should explain why this swath of area is not interesting."

Luke: Anyway, I do agree with some threads that are being said here about how people throw out ideas, and they might even phrase their sentences as confident assertions, but they're not actually thinking about the issues, and they aren't engaging critically, because they don't actually take FAI seriously, and they don't care about it, and they're just making stuff up.

There are very few people who decided, "Oh, this is a really big problem, and I should actually think about it as a full-time job or something close to that," so there is almost nobody to engage with.

Jacob: I would also say, that even for people who have decided this is an important problem to engage with, you should not measure how promising they are — for the first 100 hours, at least, of their engagement with you — based on whether the specific proposals they make are reasonable proposals or not.

You should be trying to look at much more general skills. It's the same reason why it wouldn't be reasonable to compare Eliezer to some random machine learning researcher based on how well they could implement a machine algorithm. That would be stupid, because Eliezer hasn’t spent any time developing that skill. And these people haven't spent anytime developing the FAI skill.

Holden: Okay so going back, the argument I’ve heard against the Google Maps AGI, which I think is a good argument… It's, basically, that you have this thing and it's telling you what it is going to do. But you can't tell, because the whole thing is just too complicated to read. It's too complicated for a human being to make sense of.

You have to choose, "Do I trust this thing, or do I throw it out in the trash can?" You're not going to have this option to understand what it's thinking, the algorithm is too complicated.

So in order to build something that helps you understand how it's thinking, you have to build something that in itself is an intelligent agent that is always at risk of manipulation.

Which, I also think, it's all just possibilities. I also think, because this is the default way the software is built, because it may be possible to understand the read out and because doing so would be much safer than trying to develop friendliness a priori, that this should be one of the major things being talked about by SI because it's a possibility.

Eliezer: Suppose, you had an object that worked exactly the way you wanted it to work, and you can make it as transparent as you want. Is there still some way that could go wrong?

Holden: Obviously, if the wrong person gets their hands on it and wants the wrong things.

Eliezer: Okay, any other ways?

Jacob: I think I would disagree with the stipulation that this is an important skill for making progress on FAI.

Eliezer: That's very surprising.

Jacob: The reason why is because it seems like you're getting way ahead of yourself. If you look at the FAI problem there's much simpler problems that you need to solve before you get to the point where you have this thing that you think is a friendly AI in your hand and you need to decide what to do with it.

Eliezer: Sure. But which of those things are important is determined by backward chaining from what you actually need to have at the end. Obviously you want to think about forward-chaining and you want to think about backward-chaining…

Jacob: Backward chain, forward chain are two strategies you can use to make progress on long term problems but they're not the only two strategies you can use.

Eliezer: That's actually a somewhat surprising statement…

Jacob: There are non-local strategies as well. One common thing to do is you identify a common difficulty that seems to be at the root of all the things. But if we're going to stay in graph theory terms, you identify a small cut in the graph and you're going to have to solve one of these three problems. Let's just have…

Eliezer: But you're still thinking about the whole graph in order to identify…

Jacob: No.

Eliezer: If you didn't know where you were going from the cut…

Jacob: No, that's the key. This is like the Lobian obstacle which you came up with. If you can identify a difficulty that you can argue it's probably going to be a difficulty for any solution… [Disclaimer by Jacob: this and my ensuing statements about the Lobian obstacle were attempts to understand Eliezer's assertion and don't reflect my actual beliefs.]

Eliezer: But that's totally not what the Lobian obstacle is.

Jacob: Well, I don't know why you care about the Lobian obstacle, then.

Eliezer: Because it was a problem crisp enough that other people can work on.

Jacob: Oh, okay. I thought it was also something that you would need to solve before you could get...

Eliezer: I though I already said in this conversation that the Lobian obstacle is not a big obstacle to AI, it's there because it was something crisp enough that you could get started on reflectivity.

Holden: I do recall that.

Jacob: Yeah, I recall that, as well. I guess, I just misinterpreted you, and thought that when you said, "Not a big obstacle," what you were implying was that it was a small obstacle to AI that we expect to overcome without too much difficulty.

Eliezer: No, it's a point where you can crisply point out something we don't yet know how to do.

Jacob: You think it's just not related to AI but can...?

Eliezer: It's something that would go wrong with current attempts to formalize reflectivity.

Jacob: OK.

Eliezer: It's entirely possible that the actual AI theory of Tomorrow -- capital T -- uses some alternative viable approach in which Lobian obstacle never materializes in the first. It can also be that the Lobian obstacle has some really simple solution that nobody has thought of yet. But it’s a reflective thing what we don't know how to do. It's crisp and mathy, and people can work on it. And because of the extreme difficulty of factoring out *anything* that other people can work on, that makes the Lobian obstacle extremely valuable.

Jacob: Yeah, okay, I think we're saying similar things. I basically, agree with that. Actually, I, apparently, agree with that more than I thought I did. I thought that your view was different than that.

Eliezer: This is actually kind of puzzling to me. There appears to be some very great difficulty in communicating the fact that we didn't think the Lobian obstacle is super important, qua an obstacle to AI. It's very useful to have as a way to get started working on reflectivity…

Jacob: Yeah. I, actually, heard you say that several times. All the times I assumed that when you said, "Not a huge obstacle," you just meant that it was a minor obstacle, not that it might not be an obstacle at all. I'm not sure why I thought that. I get the impression that you guys think that self-reflection is incredibly important problem, and that mathematical logic is something that you fundamentally have to deal with.

Eliezer: There are also all these explicit disclaimers about how mathematical logic is a terrible fit for describing the environment.

Jacob: Yeah, I take that to mean that you think we need something better, that working on the Lobian obstacle is the first step for developing something better. That's what I got out of that.

Eliezer: I wonder if this can be solved by...

Jacob: If you can give concrete examples -- not necessarily precise examples -- but examples of things that you can imagine that would overcome the Lobian obstacle and say, "I could imagine something like this."

Eliezer: Paul's theory would be an obvious one. Paul's interpretation of Paul's theory which I find very intriguing is that undefinability of truth is an artifact of demanding infinite precision.

Jacob: You think that that's representative of the sort of things that could get around this?

Eliezer: It's extremely representative. I don't know if it's actually the thing, but it's as representative as you can possibly get.

Jacob: So if you have a family or possibly just one representative example of the sort of thing that doesn't go through the low grad obstacle, just including that example would like...

Eliezer: I did, in the Tiling Agents paper.

Jacob: The thing is that I didn't read the tiling agents paper. I skimmed through it. All of my personal interactions of you where you said that…

Eliezer: This is a remarkable language obstacle. I can't understand how I've had as much trouble as I had with it.

Holden: It seems like a good potential time to try and refocus. So if I were to summarize what we've done so far, there was a big object level discussion about whether we need to build FAI from the ground up.

Eliezer: I think I expect Friendly AI to be harder than Holden does.

Holden: I don't think that's true at all, for what it's worth. It could be but I don't think...

Eliezer: When I was inventing things that in retrospect it appear about as promising as Google Map's AGI, I had priors telling me that FAI was supposed to be immensely harder than any of the concepts I was throwing at it, so I expected the concepts to be wrong. It wasn't like self-criticism. Like doubting this thing that I was attached to. It was that I literally, actually, believed it couldn't be that easy.

Holden: See my picture isn't that it's easy. My picture it that it's this incredibly hard thing that we shouldn't bother to do until we have this tool for doing it that will make it easier, like by orders of magnitude, but could still leave it being almost impossibly hard.

Jacob: My impression was that we roughly agree about the difficulty of FAI, and disagreed about various sociological things.

Holden: Just to give a little background, I have in my head had it flipped in terms of who thinks Friendly AI is harder, you or me, because what my post is arguing for example was like: Our odds of doing this, without this tool, are basically zero. Even if maps are hard to get and even if it's remains hard after we get it, we just need it.

And then I've been making a similar line of argument in this conversation about getting academics involved where you're like, "Yeah, if we can have it be string theory that'd be awesome but that's too high to reach." And I'm like, "Well, we just have to reach that." This problem is too hard for Eliezer and his like self-crafted team to beat.

It has to be de-centralized, diverse community of super brilliant academics. It can't be the people that Eliezer personally trains. It's just too hard for that. You need the tool thing to have a shot. The other way, you have no shot. At least, that's how I was thinking of it at the time.

Luke: I’ll just mention, this is a very common objection and quite a reasonable one of: how on earth would a team at a nonprofit build FAI before the world builds AGI? Quite a reasonable objection.

Eliezer: I don't really process things in those terms. I see the object-level problem, and try to pull resources toward attacking the object-level problem.

Holden: My sense is that the problem is too big for you and your team and it needs to become a field like string theory and that's what you should be trying to do.

Eliezer: That would be much more attractive thing to do if I could visualize how it is that everyone can know how to build an FAI yet nobody knew how to build an AGI 5 years before that. If I could actually visualize a path through that tactic that involves humanity surviving, then I would probably be much more enthusiastic about it.

Holden: There are historical cases of times when people came together and said, "This is dangerous. We need to agree not to do certain things with it until we all agreed that we were ready." In a lot of cases, that involves the government taking control of the technology and things like that.

Eliezer: Yeah. So about the civilizational adequacy thing, and the whole “You’re special” thing, I think one concrete example to illustrate is: Let’s say that tomorrow it will be revealed that tomorrow the NSA has actually been running this big FAI project which has gotten way way further than anyone at MIRI. The incorrect construal of how I think the world works would be like "Ha-ha. See, you weren't so special after all. Your basic hypothesis was wrong." And I'd be like, "Oh. My basic observation that bizarrely people aren't working on this, is wrong." That was thing I didn't have any confidence in. Then the rest of it is like this mysterious background facts.

Holden: We're making the argument here that people are not good at the cause neutrality thing. We have evidence of that but we don't have evidence of the other things except as they pertain to friendly AI, which could easily all just be a symptom of the cause neutrality thing. Solve the cause neutrality thing, get people interested in friendly AI, and that's our best shot here.

Eliezer: That kind of is the whole CFAR strategy, but generalized somewhat beyond cause neutrality. I do think that there are other cognitive skills involved besides the one that you've got the most experience with.

Holden: I just haven't been compelled by your account of them. I've just heard them and when you say, they don't have this skill either, I'm like, "No, no, no, that would go away if it weren't for the cause neutrality thing." This is not observable, this is not a testable proposition I'm making, nor is it on your side.

Eliezer: I guess I’m slightly surprised. It seems to me, it should be obvious to you that you should expect to have more acquaintance with cause neutrality than the other things.

You should expect in advance to predictably update in favor of the other things having more relative important if you have more engagement with them.

Jacob: So, we explicitly disagreed on one other example you brought up, which is the self-criticism thing.

Eliezer: Again, I’m uncomfortable with framing these in terms of hypotheses about innate qualities I have that other people don’t have. That seems like a straw man of my view. You'll remember that after a few moments of additional cogitation, I was like, “Actually it's not so much that I was self-critical, it’s just that I started out assuming the problem was very hard and so the concepts I was throwing at it couldn’t possibly be solutions.”

Dario: Can you come up with any examples of this that aren’t in the Friendly AI field?

Eliezer: The criticism thing?

Dario: Yeah.

Eliezer: Computer security people are super-trained for it.

Dario: No, what I'm saying is can you come up with any examples of you having it and others not that aren't in the friendly AI field because there's this circularity to arguing about this within the friendly AI field.

Holden: That's what I've been saying.

Jacob: Can I make a meta point? My impression is that you're worried about running into this standard societal view where you're not allowed to claim that you have special qualities because that's penalized societally. But no one in this room is going to penalize you for that, so is that your actual belief?

Eliezer: Six years ago, that would've been my actual belief. Nowadays, I genuinely am more neutral about the causes of civilizational incompetence and don't pretend to have exact theories of how it works because observation is much more solid than the hypotheses to explain it.

Dario: When you are observing something and you observe some data point and the data point doesn't seem to make sense mechanistically or you wouldn't predict it based on everything else that you see, depending on what else you see it should cast at least a little bit of doubt on whether you're actually viewing the data point correctly. Isn't that right?

Eliezer: I'm trying to avoid the calcification of viewpoints that happens when people come up with more and more justifications for a policy. I want to be ready for the case where it is revealed tomorrow that the NSA has been doing an FAI project for the last 10 years.

Jacob: There's a simpler explanation for this. It could be the case that you're just wrong in certain ways about what things are germane to the FAI problem.

Eliezer: I'm not quite sure what you mean by that. If you mean that the Lobian obstacle turns out to be irrelevant then that's probable in the first place. Did all the work that came out toward reflective decision theory from working on the Lobian obstacle tell us nothing about how reflectivity works? Then that's kind of sad but still plausible.

If it turns out that thinking about self-improvement and things modifying themselves is not the way to approach AI, that’s substantially less probable but still possible, because you could just have it be that there’s a set of object level probability problems you have to solve and their application to self-modification work isn’t obvious given their general appearance.

Then, if you say the intelligence explosion isn’t the correct thing to be worrying about and then my jaw drops and I’m like “What?”

Jacob: So I think self-modification could quite possibly be the wrong way to think about AI. I don't feel strongly either way. It's a useful way to think. Sorry, “wrong” is the wrong word. It's a useful lens with which to view AI, but there are other useful lenses as well.

Eliezer: I mean, I’m not going to disagree with that statement.

Jacob: You can imagine that there are lenses that seem to you to be unreasonable lenses with which to view AI that actually turn out to be reasonable for reasons that were subtle.

Eliezer: Possibly. In general, I feel that good things are rare and it's much more likely for an apparently good thing to turn out to be bad than it is for an apparently bad thing to turn out to be good.

Jacob: Yeah. But it also is the case that there are entire fields of academia that you're not engaging with. There's probably at least one good idea within all of those fields.

Eliezer: Maybe? It's not easy to do work on friendly AI on purpose, let alone by accident, but I am sure there are subfields of math that have useful concepts that I haven’t yet found.

Jacob: I personally think you’re less bullish on statistics than you should be. I could be wrong about that.

Eliezer: [laughs] What? I'm a flippin' Bayesian, man. How much more bullish would you like me to get?

Jacob: You use probability theory a lot.

Eliezer: Yeah?

Jacob: Which is why I find it confusing that you don't like statistics.

Eliezer: I like statistics. I don't think you can do certain things with certain types of statistical guarantees.

Luke: Eliezer's view on statistics is super narrow about this one issue.

Jacob: It just comes up in every argument I've ever had with Eliezer, so maybe I'm…

Holden: Can I try and refocus us again? We're starting to drift more and more and that might be a sign that we need to use our remaining time as well as we can.

I just want to go back to the big three questions. All right? I'm going to repeat stuff that I've already repeated before, but I just want to pull it all together in one place. One, does friendliness need to be built in from the ground up? We learned about each other's views, but I'm still unclear on whether we want to try and have an agenda for going further especially whether Eliezer finds it worthwhile.

Two, is it worth trying to engage academia? That's a question both about whether it's possible and whether it would lead to good things, which we've talked about a lot.

Then, the third question which we haven't talked about it all is: MIRI's research agenda, is this a good research agenda?

Possible things I could imagine doing: I would probably vote to not continue the engage the academia conversation just because I feel that it's gone in a lot of directions and we're low in time. We can think about it a little more off-line. What can we do to convince Eliezer that it's worth trying to engage academia and such like?

Luke: Can I just very briefly flag that when I go to work in the morning a significant frame for what I'm doing is engaging academia in the FAI problem.

Eliezer: I just went to MIT and gave a lecture and everything.

Luke: Right, so I don't know that we have time for this, but I'm confused by the whole idea that we might disagree about this, and we might just disagree about *which* people in AI to talk to, or something like that.

Jacob: I think the disagreement is about how you should use academics.

Dario: That is the thing I was getting at. Subfields within AI can be quite different.

Jacob: I was going to say mathematicians vs. others.

Dario: That, too.

Luke: That seems like interesting tactics that I have discussed with people in the past and would love to discuss with you also.

Some of the more substantive actual disagreements we might have are not so much about whether to engage academia but whether you can make most of the AI problem be a thing that's tackled in public or whether you have to do it with better information security than that because you can use FAI knowledge to build AGI and somebody will do that. That's more what a lot of the discussion there was about.

Eliezer: Especially because 10 percent of FAI is this meta-utility function stuff, and the other 90% is building the AI to an unusually rigorous standard.

Holden: I want to list three sub-questions of "Should you engage academia?" The weird thing is that we spent a lot of time asking could academics even be helpful? Is it even plausible to engage them? Now, you're coming out and saying, "No, we are trying to engage academia." Other questions, we could be asking, "Is attacking the problem in the open too dangerous?"

We could, also, be asking the question of how you should engage academia and which academics you should engage because my sense is that you should be spending more time talking to Jacob and the people he knows and people who actually work on AI related stuff. And not as much emphasis on what I perceive as whatever academics we can get to pay attention to us strategy.

On the problem of working on AI in the open, it would be good to find historical examples of dangerous technologies that were developed and see how that all played out, because my hypothesis is that once something is recognized as dangerous, you can expect a lot of kind of serious and cooperation about that.

But anyway, I think the best path is to try to convince people in Jacob's community that Friendly AI is a real concern, that it is dangerous if we don't get it, and simultaneously convince them to be careful and also convince them to work on it, which is what you want.

Eliezer: Consider Leo Szilard's quest to get people to take chain reaction seriously. Certainly the part where Fermi came on board, rather reluctantly… then they had to go to get Einstein, who sent a letter to President Roosevelt, then Roosevelt actually paid attention which I think totally wouldn't happen nowadays…

Holden: Why wouldn't that happen? We had the same thing happen with the LHC…

Eliezer: What?

Dario: No, that makes sense. The LHC worries were very much less likely than the idea that nuclear weapons would ignite in the atmosphere, yet it still got some attention. It's not like it was totally dismissed.

Eliezer: I'm talking about the ability to ask whether something that actually needs to be done, you write a letter to the president, and the president takes it seriously. I would expect that not to happen these days.

Dario: George Church wrote a letter to the President Obama and Obama gave him a $100 million.

Eliezer: Is that actually what happened?

Dario: He didn't write one letter, but in a metaphorical sense, Church is being taken very seriously.

Eliezer: So Church did a lot more than just writing a letter to the president.

Holden: So yeah, so it might take some effort. Once, you say, X technology could have national security implications or military implications and we should be careful. It's not that hard to get powerful people to back you up on that.

Eliezer: I'm not sure that could be done at all, let alone productive. I do think that if people who *politicians* could recognize as AI leaders all came together and made some brouhaha about needing AI to fund off the terrorists, that could be like the $100 million project. It’s not as clear to me that FAI could be explained to politicians…

Holden: I strongly disagree with your intuitions here.

Dario: I missed the last minute, but right when I walked out, I was going to say that people including NASA take asteroid risk quite seriously and they even Kessler syndrome seriously.

Eliezer: Kessler syndrome?

Dario: That was the idea that was portrayed badly in the film *Gravity* about having enough space debris that it sets off a chain reaction and sets the debris flying far enough that it's very hard to launch into low-Earth orbit without destroying whatever you're launching.

Jacob: So we get stuck on Earth?

Dario: Yes.

Eliezer: I thought that was a rather reasonable thing to be concerned about.

Dario: No, it *is* a reasonable thing to be concerned about. I'm saying, that NASA takes asteroid risk seriously. They take Kessler syndrome seriously. It was portrayed unrealistically in the film.

Eliezer: So, there's a continuum of these things. Asteroid research is easier to understand than global warming, although they are very much less of a threat.

Dario: One in 10 million per year, or something like that.

Eliezer: Yeah. And global warming is much easier to understand than biotech supervirus stuff. And that, in turn, is much easier to understand than nanotech, and nanotech is much easier to understand than FAI type stuff. And it cuts out above biotech and below nanotech.

Dario: I will say, that the idea that biotech is just of the cusp of being comprehensible to people… I really don't think that. I believe that there was a period when DARPA chose not to fund synthetic biology stuff due to safety concerns.

Eliezer: The government understands that biotech is dangerous. They don't understand how dangerous or which parts are dangerous.

Dario: I think DARPA does understand that.

Holden: Yeah, I think there are people in the government who do.

Eliezer: I can believe there are some, especially just to the extent of saying no to funding, or just one person on a committee.

Dario: They were all pretty concerned, even more so than I thought was justifiable.

Eliezer: Okay, but then the government can’t understand nanotech, and AI is just beyond them. And we should discuss nanotech at some point, because I feel like a lot of my strategy is informed by “Foresight tried academic engagement and it didn’t work.”

Dario: OK, then we should discuss it because I think they're wrong. One thing, I will say quickly, is that this nano computer concern, I would say, is not a concern. One thing that makes it a much smaller concern than you'd think is the Landauer limit.

Eliezer: Heat dissipation?

Luke: That's the heat dissipation.

Eliezer: So, reversible computing?

Dario: Yes, but then you would have to invent nano computers *and* reversible computers at the same time in order to be much faster.

Eliezer: The Foresight people are way ahead of you here.

Dario: No, I looked over it and they don’t have great ideas for reversible computing. Nothing they've done has convinced me in any way.

Holden: Hey guys, I'm going to propose that we move to the screen room where I can use the whiteboard to diagram the major questions so we can stick on them.

[people moving]

Holden: This is how I see the agenda for today, and I still don't see a straight answer on this, the question about whether we need to build FAI from the ground up. Are there next steps on the object level thing?

Luke: So, what if we looked at the standard view among people who work on safety-critical systems? Some people I talk to have the view that AI’s probably going to be fine, and you need fewer guarantees about its behavior than people working on safety-critical systems *already* think you need for autopilot systems, which are much less powerful than AGI.

Jacob: There’s one thing I would like MIRI to do, which is actually a concrete research question. I heard this from Paul. Suppose you had an AI with arbitrarily large computational resources, let’s say it has a halting oracle, and there’s some utility function that it’s supposed to acquire but it doesn’t know what it is: can you come up with a way for it to acquire that utility function?

I am much more excited about the meta-utility problem than I am about the Lobian obstacle.

Eliezer: I don't have any good equivalents of the Lobian obstacle. I can state other things we don’t know how to do with infinite computing power, but they’re not as crisp.

Jacob: I can't write down a formal mathematical statement, but I feel like I could probably write up an exposition that would make a reasonable case for why this is a problem. I'm even happy to do that probably, although, I don't know if you’ll like my exposition or not.

Eliezer: One version of it might be something along the lines of: with infinite computing power, learn your utility function from this species’ brains without creating an instrumental incentive to rewrite their brains. That seems a little too easy… Actually, it might even be better to write up even if it seems like it might be too easy because my solution could be wrong. Or I could discover that nobody can come up with what I think is the easy version, and either of those results would be interesting.

Jacob: My version of this would be: you're some agent and under your model of the world, with respect to your subjective probability distribution you can write down some utility function that doesn't take up too many bits. Maybe, it's only a billion or something.

Then, there are some other AI that’s another agent, and you need to build that agent in such a way that it can interact with you and acquire it's utility function, like the utility function…

Eliezer: Perhaps, I should exhibit an entire Workflowy full of variations on problems exactly like that one that we could potentially…

Jacob: I mean, it's worth noting that, I thought, when you said, "Write this up,” I thought you meant with justification for, "Why we should care about this problem in the first place?" It seems to me pretty clear that this is a problem worth working on, but not everyone I run into agrees with this. And I think I can make arguments for why this is a problem worth working on.

Eliezer: I wasn't thinking of the main obstacle there as writing it up in a way that looks persuasive to people that they should care about it. I was thinking in terms of the main difficulty being coming up with a set of sufficiently well-explained problems.

Dario: My response would be that's the wrong thing to be trying to do. If you could just write up arguments as to why people should care about this, and provide enough English language details that they understand what problem it is you're trying to solve, then people are smart enough that they'll do the thing you want them to…

Eliezer: That sounds possible. Testing it comes at a cost and I really don't expect you to be correct about that.

Jacob: Oh, I expect you to get more engagement if you do it your way, because it’s basically like solving 70-80 percent of the research problem and then like...

Eliezer: Yeah, that's my model. I have to solve 80 percent of it and then, someone else can do the last 20%.

Jacob: Sure, so if you solve 80 percent of the problem and then you’re like, "Here get a free paper, you do the trivial part of this”, not everyone is going to say “yes”, but many of them will.

Eliezer: But I don't agree that stating the Lobian obstacle was 80 percent of the work, maybe more like 50 percent of the work. Marcello and I worked on Lobian obstacle to make sure that we couldn't do that last 20 percent, so it probably wasn't 80 percent.

Jacob: It probably varies from problem to problem.

[crosstalk]

Eliezer: Here’s a concrete open problem example. Exhibit a self-modifying agent, which preserves the code implementing a shutdown button throughout self-modification.

Jacob: I agree that this is a good problem to work on. I'd have to think for little bit about whether I felt like I could exposit well on why it's a good problem work on. But that seems like a good example. I'm also somewhat surprised if it's not solved with known techniques. I'm not claiming to have the solution in my head...

Eliezer: I mean: quining, sure. I was thinking not in terms of quining but in terms of: this AI does not have an Omohundro convergent instrumental incentive to prevent its own shutdown. Not just that it reproduces its own code, but that this is a rational agent which does not perceive an instrumental incentive to strip the shutdown button from its code. That seems like a non-trivial problem in the sense that I don't know how to solve it.

Jacob: There are two versions of this problem. One of them, runs into all of the problems about AIXI being weird and not being able to locate itself in the universe.

Eliezer: Having a Cartesian boundary, in other words?

Jacob: I don't know what a Cartesian boundary is.

Eliezer: It's like the AI's model of reality contains a hard boundary between the outside of the AI and the inside of AI?

Jacob: I'm saying there's one version of the problem where, in some sense, if you don't solve the Cartesian boundary issue, you're not going to be able to solve this problem in full generality.

Eliezer: That's not obvious to me.

Jacob: If it doesn't know where it is in the world, then it's hard for it to guarantee that it won't hit its shut down button with a laser beam that pries it off and, therefore, it can't be trusted anymore. Right? For instance...

Eliezer: There was a stronger version of this, which is describe the AI that creates the shutdown version in the first place, and "describe the AI tries to protect the shutdown button from stray lasers without protecting it from humans" would be an intermediate between those two. But this was just like, "there is no instrumental incentive to eliminate the shutdown button."

Jacob: Are you happy with the following formalization of the problem? We'll model it as a Turing machine. Let's say every thousand instructions or so it reads an either zero or a one. It needs to maintain the invariant that if the thing it reads is ever a one, it immediately terminates.

Eliezer: That doesn't sound very rational agency doing consequentialist reasoning with a goal. You want a consequentiality agent, which is selecting between actions on the basis of their being expected to satisfy a logical goal or maximize or improve an expected utility calculation, and then, we want to not have a convergent instrumental incentive to remove the shutdown button.

Jacob: What if I say something like, for any algorithm, it executes the algorithm except with the additional stipulation that, if a one appears at this location of the input, then within the next 1,000 operations, it must terminate?

Eliezer: Yeah, having a term for the shutdown button in the utility function would be one way of addressing it.

Jacob: Well, I'm not even talking about utility functions. I'm solving a more general problem. Let's not worry about the fact that this is an AI. Let's just worry about the fact that it's an algorithm at all and solve the problem: for any algorithm, create something that, say, with at most a factor of two slowdown, runs that algorithm, with additionally the stipulation that, if a certain input is ever a one...

Eliezer: That sounds uninteresting. That doesn't sound like it intersects with agency at all. This is not a property of an intelligent mind. This is a property of the machine the intelligent mind is running on, and as soon as it can, it constructs a new machine which doesn't have this property.

Jacob: In that case, I don't understand what the actual problem is that you're trying to pose.

Eliezer: It's a bit of an artificial problem, but… You're trying to construct a rational agent which… like, you reach out to press that shutdown button, the AI has a hand right there, and it doesn't interpose its hand in front of the shutdown button.

Jacob: Yeah, I get kinda worried when you say rational agent, because I think it's presupposing...

Eliezer: Consequential agent? Means-end reasoner?

Jacob: The thing that worries me is, when you say, that you're presupposing pretty strong assumptions about what an AI will look like that I disagree with.

Eliezer: I will note that they're very mainstream assumptions.

Jacob: Are you talking about the classical normative decision theory framework, because I think that's a bad framework for AI.

Eliezer: Why? Because we can't actually maximize?

Jacob: Among other things. I'm not sure that this is the right discussion to get into…

Eliezer: It doesn't necessarily have to be expected utility maximization, but if you have some alternative, plausible, general formalism of expected utility, or that doesn't violate the Von Neumann-Morgenstern axioms, or that does but is possible anyway, then you exhibit that there's some variation on it…

I mean, the generic version that maximizes paperclips clearly should interpose a hand. If you start with a different framework and then you exhibit something within that framework that doesn't interpose its hand between you and the shutdown button.

Jacob: Are you happier with the following formalism? Which is, you want to construct something where: assume it has infinite computational resources, it gets the highest utility possible subject to the constraints that it maintains the invariant of having a shutdown button.

Eliezer: Does it say the invariant was satisfied because I had this button here and then interposed a hand to prevent you from actually pressing the button?

Jacob: I'm trying to avoid issues about the Cartesian barrier, or whatever it was you called it. I'm trying to imagine this only as there is some input that it gets every once in a while and it has to maintain the invariant that allows shutdown.

Eliezer: If the input is not modifiable by the AI, it's just given, then that's very Cartesian, because there's this input generated by forces you can't affect.

Jacob: Yes, what I'm saying is that, that is Cartesian, and to solve the harder version of this problem where you actually have the environment, would require you first to solve the fact that it's hard to create AIs that don't have this property.

Eliezer: Basically, what I want to see is engagement with the central paradoxical thing that you totally want an actual AI to do, where it's OK with letting you shut it down, it doesn't pull out a shotgun and shoot you. What I want to see is engagement with the central difficulty, as opposed to cleverly dodging it somehow.

Jacob: What I'm saying is, I don't see a way to gain traction on this problem without first making traction on the problem that it's hard to formalize the notion of an AI in its environment in such a way that the AI actually reasons in a non-insane way about its environment.

Eliezer: I'm not sure what you mean by that. If it's not solved by putting the AI in a certain crisp environment that it can reason about with first-order logic, for example. It's a terrible example, but…

Jacob: If you're given any environment, then sure.

Eliezer: So like in the tiling agents paper I can describe this succession of agents building agents out of environmental transistors…

Jacob: The problem is that any formally described objects can just be given as additional tapes on the Turing machine, so it's not going to get out of this formalism which you seem to dislike.

Eliezer: There's a communications issue. What I want is for you to not say, "Ha, this computer here automatically shuts down when it gets to a one." Problem solved. It has to have something to do with agency. I want you to have something about, "Even though this thing is a consequentialist reasoner, it didn't interpose a hand between you and the shutdown button." You could have, for example, this environmental light in the sky that sometimes goes from zero to one, and that's from outside the system, outside the Turing machine, or whatever, but then inside the system, which is a logical, closed system, the AI has the option of shutting down, or has an action available to modify itself so as to not shutdown when the light in the sky goes to one. But it continues being in the state where it will shut down because it sees the light in the sky, and it preserves that through self-modification. Then, you'd be engaging with the central problem.

Jacob: There are two versions of that problem. One, of which, is the same as the thing I was already saying and one, of which is different. The one that's different is, not only do you want to shut down when that light in the sky goes on, but you're not allowed to interfere with other agents of the environment's plans to cause the light to turn on.

Eliezer: You can skip that part for now. You'd have to engage with it eventually because you want an AI that does resist incoming meteorites but doesn't resist a programmer pressing the button, but I agree, that's more complicated. It makes function calls to a bunch of things that aren't described inside the scope of this project. This is just a lone agents in the environment, it can self-modify, it has evolved. There's also light in the sky.

Jacob: Sure, and that's the same as what I was saying, or at least trying to say. This explicit problem, I'm sure, probably hasn't been posed before, but minor variations on it I'm certain have been solved by the program verification community.

Eliezer: What? No, you're not trying to write a guarantee about the fixed program from the outside that it shuts down due to a light from the sun. You're trying to have a self-modifying, consequentialist agent that keeps the property…

Jacob: But I've already shown you that self-modifying isn't an important property. You can have a universal Turing machine plus a block of source code and...

Eliezer: Self-modifying *consequentialist*. It's trying to build a stack of red blocks as high as possible. If it shuts down, it won't be able to build the stack as high.

Jacob: Yes, that's why, I said, it should maximize utility subject to the fact that it always maintains as an invariant that if the light on the sky goes on, it will shut down. I don't see how you can get more consequentialist than that. Although, that additional stipulation, I'm not confident has been solved by the program analysis community.

Eliezer: Yeah, that feels like a very preliminary engagement. If that's going to be your answer, then I'm going to immediately add on, because, in fact, this AI, with the utility function as described, would — if it could see outside the system and it could see your hand heading for the shutdown button — it immediately interposes some hand. Do you agree with that?

Jacob: Yes.

Eliezer: OK, solve *that* problem.

Jacob: Okay, I think I understand the problem that you care about, and my response is, basically, that there's an easy version of this. Which, we both agree, is not that interesting but maybe should be solved anyway, because it shouldn't take too long.

Then, there's a hard version. My intuition is that the hard version, actually flows through a bunch of other problems that we need to get a handle on before that. I could explain why this problem is an important problem to solve, but I wouldn't be able to make a convincing case that it's the right problem to work on right now because I don't think it *is* the right problem to work on right now.

Eliezer: This sounds an awful lot like it's not the right problem to work on right now because it's not crisp enough yet.

Jacob: No, I don't think that's it. I feel that the crisp posing of it is going to be a strict superset of the crisp posing of other problems. It's not that there's no crisp posing.

Eliezer: Do you think that you could, on your own, state an intermediate problem which was interesting enough to be workable on, and crisp enough to be solvable? Because that's the kind of thing where I'm like, "Oh, I have to do that."

Jacob: For a sufficient definition of intermediate problem, yes. It might not look very much like this problem.

Eliezer: You can come up with some kind of problem that has a non-trivial solution and state the problem crisply and then solve it and that's visible progress for this problem.

Jacob: I feel like I really cared about solving this problem and I was trying to make progress, the next thing I would do after solving the really easy problem, just to make sure that I understood it correctly, would be to look at all the reasons why trying to solve the harder problem runs into a bunch of other difficulties, and identify one of those difficulties that seem both germane and not itself having many prerequisites, and try to solve that problem.

Eliezer: A reasonable approach. Indeed, the problem of building a Friendly AI is very large and so one recurses on relatively more local prerequisites and more well-posed things like “Have something that doesn’t stop you from shutting it down.” And then in turn you can recurse on that. So the process you’re describing is very much how one arrives at “build something that doesn’t stop you from shutting it down” from the larger problem of “build a Friendly AI.” Does that sound right?

Jacob: Yeah.