01-21-2014 conversation between Jacob Steinhardt and Luke Muehlhauser on MIRI strategy, summarized and paraphrased.

Jacob: I think cultural distance and communication distance are maybe a bigger problem for MIRI than inferential distance, when it comes to engaging other researchers. If I was trying to motivate the problem of AI safety, I would point to bad outcomes from poorly engineered systems, and I would point to the fact that autonomous systems are becoming more ubiquitous and more relied-upon, and I would say that given that this trend will probably continue, we need to figure out how to avoid unacceptable software errors.

Luke: I actually wrote a draft framed that way a while back, called "Trustworthy autonomous programs" (TAP). Some people aren't excited about that framing because they think it'll lead people to work on things that are barely valuable at all. But *I* think it's a worthwhile approach. Maybe it nudges people in the direction of program analysis and program synthesis, which I think is valuable, though those subfields are already getting way more investment than explicit FAI work.

Jacob: People in program analysis are already kinda good at finding proofs that a program satisfies some formal requirements. But coming up with formal requirements that match intuitively desirable properties is much harder. Program analysis people would be really happy to see progress on this, but few people are working on it because it's known to be very hard, and they don't see good approaches.

Luke: Have you seen much written discussion of that challenge?

Jacob: No, there isn't much writing on this. You should check out Daikon, though.

I think people would be especially interested if you could come up with a good metric for progress on this problem. People tend to work on what they can evaluate.

Luke: So, MIRI could focus on the TAP story, and then say "...and here are some research avenues we think are particularly promising, but other researchers think X and Y are more promising..." Different people have very different views about what's likely to be useful. Like, I think program synthesis might be even more likely to be useful than the decision theory stuff, but I'm not really sure.

Jacob: I share your intuition on that.

Luke: MIRI could be one of many groups trying to organize research and collaborations on this general "safety of future autonomous systems" problem, and we'd just have one set of problem foci like everybody else.

Jacob: I also think it might be more fruitful to think in terms of a sliding scale of system complexity from today's systems to the systems we'll have a few decades from now, rather than thinking very abstractly about what you think the shape of AGI will look like and what that implies about friendliness. Any time you can turn a really hard problem into a continuous scale of problems of increasing difficulty, that's a much better way to make progress. Though you do need to worry about whether the work that's useful on the easier problems may be of an entirely different character than the work that's useful on the harder problems.

Luke: Unfortunately, I think that's going to be pretty common, though it's still the case that useful insights on the harder problems can come from working on the easier problems, even if the work required to solve the two looks very different. Strongly self-improving human-level AGI looks like a different kind of problem than that of getting a high safety standard in a near-future self-driving car.

Jacob: In what particular ways do you think it's very different?

Luke: For one, the self-improving AGI is capable enough that we can't just test it, turn it off, tweak it, test it again, etc.

Jacob: I still think it's more promising to solve the harder problem by way of first solving the easier problems like program synthesis.

Luke: Yeah, I'm actually fairly optimistic about the scalability of program synthesis. Like, you can imagine an agent using whatever-the-hell-works for vision and actuator control and so on, and then some relatively small core software that goes through rigorous formal verification, just like when Gerwin Klein verifies the seL4 microkernel and then uses whatever-the-hell for drivers and stuff on top of the microkernel. Maybe you can have something just a few orders of magnitude larger than seL4 that's being formally verified or even synthesized a couple decades from now that's serving as the core decision component of the AI, even if the AI also has a lot of other really complicated planning modules outside that are not formally verified. If we develop techniques to make program synthesis more scalable, more general, more modular, and we have huge libraries of verified code, I could see that scaling up fairly continuously. Then you still have a bunch of extra problems to solve like "what values do you want the AI to have when it's smarter than you are?" but at least you know a lot about how to get high safety guarantees for the core decision-making component.

So, suppose MIRI decides to, and is able to, shift its communication so that it's less alien to the academic world, where most of the relevant research talent is, and we're telling the TAP story + some stuff about what MIRI's choosing to focus on (in part due to comparative advantage), but then we're still trying to get researchers interested in the stuff that MIRI has decided to specialize in — like maybe principled reflective reasoning and indirect normativity and reasoning under fragility and so on. One challenge even in that world is that one thing we're learning from the workshops is that as a practical matter, it's actually really important for a researcher to be trying to solve the puzzles from the perspective of "How would an AI use this?" rather than "This is an interesting logic puzzle."

Jacob: In what way has that turned out to be practically important?

Luke: People work on different pieces of the problem depending on whether they're trying to solve the problem for Friendly AI or just for a math journal. If they aren't thinking about it from the FAI perspective, people can work all day on stuff that's *very* close to what we care about in concept-space and yet has no discernable value to FAI theory. Thus, the people who have contributed by far the most novel FAI progress are people explicitly thinking about the problems from the perspective of FAI, rather than thinking of them just as interesting math or philosophy problems. So I'm still left with the problem of getting people to care about FAI, and care about these particular lines of FAI research.

Jacob: This will obviously be less of a problem for the program synthesis avenue than the decision theory avenue, because it's more obvious to more people how program synthesis is relevant to AI safety, and not at all obvious to most researchers how decision theory could be related to AI safety. I don't know how to convince AI people to think decision theory is relevant, because *I* don't think it's likely to be relevant.

Luke: Let me go back to the inferential distance issue. The reason it feels to me like inferential distance is blocking my outreach to researchers, alongside communication issues, is that when I talk to researchers and try to nudge them into caring about long-term AI safety issues, it's not the case that they're familiar with the arguments and evidence that have been brought to bear on the issue, and they reject them for informed reasons. Instead, it's usually the case that they've never studied the issue or thought about it much. That's why it feels like inferential distance to me: they

have expertise in some narrow part of AI, but no expertise whatsoever in the future of AI. And the stuff I want them to care about is not going to affect their life in the next 5 years, so they don't want to invest the time to become *familiar* with the usual arguments and evidence that are brought up when discussing the future of AI.

Jacob: Do you actually find it hard to convince people that human-level AI is a total game-changer?

Luke: Not AI people, so much. The things I get stuck on are things like when the AI researcher says "Oh, I think AI is 300 years away." And then I ask if they've read When Will AI Be Created or other materials on the issue, and they say "No, no, I just think it's a really long ways away." So then there's a bunch of inferential distance for me to fill in, but they don't want to spend their time on that.

Jacob: But if they're working on AI, and they know how fast the field progresses, it seems reasonable for them to believe you're not going to have superior data on AI timelines.

Luke: I don't get the impression that people have thought about the data much, though. They just have an intuitive guess. So, an AI person says "I think AI is 300 years away," and I say "What if you think about it in terms of how much AI progress has been made per quality-adjusted researcher hour rather than AI progress per calendar year; then what do you think the trend is for quality-adjusted researcher hours going into AI?" and they say "Oh, I guess I've never thought about that." So even very basic considerations, they've never thought about them before.

Or maybe their objection is "There's no way you can put a probability on that, there's no way you could know." And I'll say, "There's no 'get out of prediction free card'; you are in fact making decisions on the basis of your intuitive predictions. You're investing in your IRA because you think it's very unlikely that AI flips the global economy in 15 years, etc." And then they'll say "Oh, I guess that's right, but... I still don't think you can put a probability on AI coming in the next 50 years or not." So again, it's not a thing they've tried to think through, and so there's a lot of inferential distance.

Jacob: I actually think it's reasonable for them not to give a probability.

Luke: They can give a range, whatever.

Jacob: The way I think about this problem is that there's a long history of people trying to predict things far out, and even the best people were pretty terrible at it, so in practice it works better to just try to identify the next challenges ahead of you and push on them rather than trying to think about the long-term perspective.

Luke: Yeah, I think that for them, it's not so much about probabilities but about what they're going to try to think about. But I mean, if you're a businessperson, and you've got to think about whether you think the market is going to shift somewhere else in 5 years, you don't have a principled way to assign probabilities, and people have historically been pretty bad at this problem, but there's just nothing else you can do but try to make the best predictions you can, and take proper account of your uncertainty and the robustness of your estimates. But the business owner *cares* about whether his business does well 5 years from now, and so he just does the best he can. And the AI people I talk to are just deciding not to try to care about the long-term future of AI, and so they don't try to study the issue and make the best predictions they can.

Jacob: But I think the academic's day-to-day research doesn't depend much on AI timelines. Whether AI is 30 or 100 years away, there are problems in front of you, and you just want to push on those.

Luke: What? Your AI timelines estimate is *totally* relevant to what should be worked on now.

Jacob: I don't think my actions depend much on AI timelines, and the reason why is that I can point to the next fundamental issues, so let's just work on that, whether AI is 20 years away or 50 years away or 100 years away. If it was 10 years away, maybe, but I think that's very unlikely.

Luke: If somebody has most of their probability mass on AI being more than 150 years out, then this drastically reduces the likely relevance of *anything* you try to do about AI safety now, and maybe they shouldn't be focused on AI safety but instead biosecurity.

Jacob: Okay, I think some sort of estimate — though maybe not exactly a probability distribution over years to AI — should impact what *field* you go into. But I don't think it should have much bearing on the particular problems you work on in that field.

Luke: Is 'field' something as big as AI?

Jacob: Maybe smaller. Maybe 'subfield.'

Luke: So, “Which subfield should I go into?” is definitely a decision facing most academics.

Jacob: I think there are a couple things going on with the academics you're talking to. Some of them just aren't engaging with you, but for the ones who are, they’re just thinking that e.g. decision theory doesn't look relevant. They might also just think the problem isn't at a mature enough state to be worth working on. In academia, there are often problems that people have identified as really important, but people don't feel like the tools and resources are in place to solve them, so they wait, or they tackle a related problem, or they try to develop the tools. Though, I think in this case we do have the tools to work on the problem in a way that might scale up to systems of greater complexity, e.g. I think we are probably ready to work on the formal specification challenge even though it's really hard.

I think the way to convince people to collaborate is to show them that there is a well-identified "next core difficulty" that they can be working on, rather than making a broad 'work on this general area because it's long-term important' argument.

BTW, I don't think it's a slam-dunk that we need to prove systems are safe.

Luke: You mean, you don't even think it's a slam dunk that there are some desirable properties we should try to prove about the core part of the system, so we can get a better safety guarantee than we can through testing alone?

Jacob: To make that argument you already need a lot of buy-in. The person has to already agree with MIRI that AGI has planet-wide repercussions, and I don't think you even need to make that argument. You don't need to persuade people of all your uncommon views about self-improving AGI to make the basic case for doing serious work on the safety of autonomous systems. For example *I* don't believe all these things, and I think the problem is really important.

When I look at MIRI's website, some of the first things I see are things I disagree with. Not the mission statement — that's uncontroversial. But one of the first things I saw was the orthogonality thesis and some other thing I disagreed with. But I don't think you need to convince people of the controversial things to get people to care about long-term AI safety.

Luke: What I said above, the “there are some desirable properties…” bit, is just the normal argument for the formal verification of safety-critical systems, which doesn’t depend on anything about planetary stakes.

Anyway, I do think it might be hard to use the TAP story to get researchers to work on things that *Eliezer* thinks are useful. My probability mass about what is likely to be useful is distributed

more broadly over more things, so I can certainly use the TAP story to try to motivate people to work on things *I* think are plausibly useful, like program synthesis. And that's going to be a more compelling framing for most researchers then "Okay, let me try to convince you that there will be self-improving superintelligence exhibiting the orthogonality thesis..."

Jacob: Yeah, that's the shift in framing I'd recommend.

Luke: So let me give you some of my data, and you can tell me what your interpretation of the data is. I encounter many AI scientists who think AGI would be transformative, and they think it's feasible within 50 years, but they at least *write* as though they're pretty confident that it'll be relatively straightforward to get high-confidence safety guarantees on those systems. That's very counterintuitive to me, and I don't know whether I should read that as a story they're telling the world in writing because that's best for their funding and social status, or whether that really describes their anticipations.

Jacob: I'm not sure. It could be a planning fallacy, or maybe they're not used to thinking about how things could go wrong, or maybe it's a career thing: you don't want to think that what you're working on could have bad social consequences. Or maybe they just think, "Well *I'm* not going to do anything stupid and dangerous." So I'm not sure what to say about people who are highly confident everything will be fine, but I think it's reasonable to think things will more likely than not be okay, and also reasonable to think the opposite. That's just a very hard question. But being really confident one way or the other is hard to justify. But again, you don't need to be highly confident of a bad outcome to justify working on this.

Have you seen the *Global Trends 2030* thing? I was impressed by how they were trying to figure out how technological developments could be used by a malicious actor to cause harm, and that was a significant factor in their decision about what research to fund or not fund.

Luke: Maybe they don't want to be responsible for a future national security risk. Do the NSF and NIH do that?

Jacob: I don't know. It's not part of their mandate, I don't think.

Luke: Maybe that's a place to push on, then. "From now on, the NSF will no longer fund research with a high probability of creating a national security risk in the next 10 years."

Jacob: The nice thing from your perspective is that DARPA provides a sizable fraction of computer science research, alongside with the Office of Naval Research and NSF. Though, an increasing amount of CS research is now being funded by companies, like Google and Microsoft. I don't know how they make funding decisions. But DARPA cares a lot about technology risks.

Luke: So when you think about the problem of getting highly confident safety guarantees out of increasingly autonomous AI, and you think about nudging research talent in that direction, what are a few different problems you would nudge people toward?

Jacob: Turning intuitively desirable system constraints into a formal specification, for sure. And also more flexible verification methods, e.g. mixed logical-probabilistic verification.

Luke: So, like, you're using a formal specification where some of the properties you're proving are allowed to have probabilistic guarantees within certain bounds? And the probability bounds make use of some solution to the problem of logical uncertainty, so the system can coherently think "I ran out of computation so I couldn't prove or disprove this property but I'm 95% confident it's true"?

Jacob: Something like that. Though I'm not *sure* we need a solution to logical uncertainty for this — I just want more flexible, scalable formal verification methods in general.

Another place I would nudge people is into the values problem, maybe moral philosophy or something.

Luke: Or maybe computational social choice, so we can aggregate values by some sensible voting mechanism or something?

Jacob: Maybe, I don't know. I haven't thought enough about the values question to know.

Another one might be more principled ways of thinking about computationally bounded agents. I kinda think that one will be solved on the way to AI, and so it isn't safety-specific, but maybe it won't be solved on the way to AI. Maybe we'll just slap together a bunch of heuristics and get to AI that way first, in which case focusing on more principled bounded agents could be useful for making progress on safer architectures, relative to the heuristic soup AI. But people have cared about principled bounded agents for a while, so I kinda think that one will just happen by default. So it's probably not the right place for MIRI to expend resources.

Luke: What about more advanced concept learning?

Jacob: I think we're already pretty good at concept learning, actually, and lots of people are working on it. In a sense, the majority of machine learning researchers are interested in this. Also, you're not going to get human-level AI if you aren't good at learning concepts.

Luke: Well, you might just not have a *principled* way of doing it — a human-transparent method for doing it.

Jacob: I guess yeah, pushing on human-interpretability tools would be good. On some level, things that make predictions should have hooks into them that should in principle be exploitable for interpreting the system. But these tools could be improved, and that could help with safety. People already do this somewhat, e.g. with autoencoders, but maybe not enough.

Back to the communication issue: one problem is that much of Eliezer's writing isn't very clear and concise.

Luke: What about my writing? I'm not sure whether I'm clearer, but I at least try to make things skimmable so that people can jump past the stuff they already know and quickly get to the point that they don't already know or might not agree with.

I actually talked to Eliezer about how it seems like I put more effort into clarity than he does, and he thought we had different concepts of clarity, which was surprising to me. He said that if someone reads some of my posts, they might have a well-organized map of concepts in their head, and they might feel like it's all clear because they can see the map of concepts all at once. But if they read some of Eliezer's posts, they will *feel Many-Worlds in their bones*, and *that* is clarity from his point of view.

Jacob: I wouldn't define the latter as clarity. I think Eliezer's posts are better at evangelizing for a particular view.

Luke: And more exciting and fun to read, in part because he's optimizing for that rather than for e.g. clarity and skimmability.

Jacob: Right. And I think high-functioning skeptical-minded people don't necessarily want something that's optimized for being exciting and fun to read, and they don't necessarily want something that is evangelizing for a particular view. So Eliezer's writing attracted a lot of people to Less Wrong, but it doesn't seem optimized for the population you're most interested in targeting for the research program. Your style is probably better for getting academics interested.