How Do We More Greatly Ensure Responsible AGI?

Participants: Eiezer YUDKOWSKY, Jeff MEDINA, Dr. Karl H. PRIBRAM, Ari HELJAKKA, Dr. Hugo De GARIS (Mod: Stephan Vladimir BUGAJ)

A video version of this dialogue is available at www.agiri.org/workshop

[Stephan Bugaj]: Each of the panelists will get 3 minutes to talk about your perspective on how do we more greatly ensure responsible AGI. Ari, could you go first?

[Ari Heljakka]: AGI is potentially extremely capable of carrying out any kind of action on this planet that humans could do right now – and an infinite number of more dangerous and more beneficial actions than we can do at this point. I suppose that's the premise that we all start from. Then there is the next question - what kind of artificial general intelligence is it going to be possible to create at all? And that's something we don't actually know the answer to. My point, very briefly put, is that we cannot really answer the questions of how to ensure responsible AGI before we have more information about what kind of architectures will it actually be feasible to produce, and what sort of behavior they show in the initial stages. Here, I suppose that we will actually have initial stages, further than we are right now, but not so close, beyond or equal to human-level intelligence as to actually become dangerous.

[Jeff Medina]: I think we don't know nearly as much as we need to have the confidence that many people have, or seem to have. A lot of people haven't studied ethics formally, right? That's okay, that's a usual thing, in science. But if we care about how to more responsibly move forward with AI, we either need to do that, to some extent, so that we can speak to same sort of language that the ethicists have agreed upon, and then talk and write about it amongst ourselves, or defer to someone else who has done those studies on both sides. You certainly should not just listen to what an ethicist, who similarly has not studied AI, the technical details, has to say about it. The same way you wouldn't care what a bioethicist, a self proclaimed bioethicist says, if they don't know much about actual biology. I think a lot of the theoretical information that is relevant to ethical questions has not been done yet. I think that the more responsible thing to do depends to a large extent on whether we assign a high probability to something like a hard take-off, which is where the advent of human-level AI leads quickly to superintelligence.

[Eiezer Yudkowsky]: I'm afraid I can't take refuge in claiming that I am completely ignorant of the subject. I have been studying it for the past 6 years, and if I hadn't come to any conclusions by now there would be a pretty strong question as to whether I was ever going to come to any conclusions. Ignorance can be a dangerous thing; it sometimes lets you think things that you would have to relinquish if you knew more about it. Alright, so summarize six years. There is no predetermined makeup of an AI. The space of possible

AI's is vastly larger than the space of human beings. We have two problems. We have what I call the technical problem and the philosophical problem. The technical problem is if you know what you are looking for, how do you reach into mind design space and pull out an AI such that it does nice things like develop medical technologies to cure cancer (or just do it directly with nanotech, depending on how smart it is) and doesn't do comparatively awful things, like wiping out the human species. The philosophical side is: what kind of AI do you want to pull out? Even if you had all the technical knowledge to do exactly what you wanted to do, you still could have done the wrong thing. Wrong according to who? Well if you wipe out the whole human species, that was wrong according to me. If anyone would like to wipe out the rest of the human species, please raise your hand. Okay, you can make progress on this problem by walking off the cliff. Having addressed this problem for a while, I concluded that you should do the technical side first. The reason being, you don't even have a language to talk about the philosophical side of the problem until you solve the technical side. You don't know what your options are. You don't have a language to describe what it is you really want. When people first approach the problem they tend to assume that the AI is going to be like them, so they model the AI by putting themselves in its shoes. Works great if you are dealing with another human, like our ancestors for last 100,000 years. But an AI does not have the similarity to your brain architecture that another human does. If you punch a human in the nose, he'll punch back. That's a conditioned response and requires a lot of evolution to get that conditional response. If you are nice to a human, they might be nice back. Having this being an unconditional response would actually be simpler than having it being a conditional response. In other words, you can pull stuff out of mind design space that is so weird, relative to a human, that Greg Egan would spit out his gum. (That's a science fiction author who writes strange stories for those of you who didn't catch the reference.) Point is, there are really strange things in mind design space. There are things that are nicer than you imagine, and there are things that are nicer than you *can* imagine, and its one of those that you want to pull out of mind design space.

[Hugo de Garis]: Definitions, *to ensure* and *responsible*. I assume by responsible you mean human friendly. As machines approach human-level, possibly we can ensure this, if we can program in such a way that is human friendly. But I have enormous question marks about our ability to do that when they become generally smart; and of course when they are hugely smarter than us, that's a different ball game. I am obsessed by this. I see this issue, *can we do this? Should we do this?* As the issues that'll dominate our global politics. I've come to a very gloomy conclusion; there will probably be a major war on this issue in the second half of this century. I just don't think that it is possible that when machines are really smart, its possible to ensure that they stay friendly to us. I make, in my book and in the media and so forth, what I think is a very simple analogy that goes like this [*Hugo dramatically slaps his wrist and then flips off an imaginary mosquito… signifying the insignificance*]. The physicist in me says, you can take a single neuron and what's it doing? It's processing a few bits per second, you can do that with a few atoms using quantum computing techniques, right? So, the potential of the physics of computation to do what biology is doing today is just hugely superior. The potential of these machines in 50 whatever years from now, is just so vastly greater than what we are, that why should they care? So I see *the issue* that's going to dominate humanity this coming century is, do we, humanity, do we build these things or not? Now when I was putting up my hand sort of half, but half jokingly [earlier, when Eliezer Yudkowky asked if anyone wanted to destroy

the human race], what the joke meant was that I think humanity has got a choice that involves facing the risk of its own extermination, if we choose to create AI's more powerful than ourselves and these creatures decide that we are a pest. They would be almost gods, right? A trillion trillion times our capacities and virtually unlimited memories, immortal, going anywhere, changing their shape. I see a kind of new religion being formed, based on this kind of stuff. Let's build these things! To describe peoples' attitudes toward these choices we need labels -- if you are pro, I call you a Cosmist: you're looking at the big picture, so-called, in the cosmos. The other group, opposed to creating powerful AI's, is called Terrans, because that's sort of their perspective, the human scale of things. So which is the greater moral evil? Risking the extinction of the human species or a kind of Deicide by refusing to build them, to build these God like creatures. If you are a fanatical Cosmist, you would say - what's one Artilect (an super AGI is called an Artilect) worth? How many human beings would you sacrifice to capture the hill? You're a general, right? 10,000? First World War, French General, you know over the top, 30% lost in no man's land, another 60% lost in capturing the trenches on the other side. What a great victory. Only 90% casualties, right? Which is the greater moral tragedy, running the risk of seeing humanity wiped out, or not building these god-like immortal creatures?

[Eliezer Yudkowsky]: I believe you are presenting rather a large number of false dichotomies wrapped into the two terms, Terrans and Cosmists. For one thing, how many lives is an Artilect worth? Implementation dependent. Let me look at its source code and I'll tell you how many lives it's worth. Is it likely to squish us when its gets big enough? Implementation dependent. Let me write the source code and I will try to guarantee you that it wont squish you. Actually looking at the source code and determining that for an arbitrary processes, is likely to be well beyond me. In theory, it's knowable beyond me because of Rice's theorem, but that's a separate issue. You are saying, it's enormously smarter, and why would it care? Because I built it. Because I reached into mind design space and selected a point in that mind design space such that I could prove that as it rewrote its own source code, it was going to keep the same optimization target. That it was going to keep on trying to steer the future into the same regions. The problems you have cited are technically addressable. We don't need a war that kills billions of people. You should be happy about that.

[Hugo de Garis]: Two major points. Technology dependent. Okay, you should do a flow chart. You should consider the various contingencies. As I see it, the keyword in this whole debate, whether you should build these or not is *risk*. We just don't know, and if they get smart enough, common sense says we don't know what they're going to do. You know we build smarter creatures all the time, with our children, who may end up growing smarter than we are, and they may turn against us. It's a possibility. Just because we build them, that's not an argument to say that they won't turn against us. If they are hugely smarter than us, you may get the Matrix scenario: you know, "You are a disease".

[Eliezer Yudkowsky]: I have a PowerPoint presentation where I actually take that scene and I show that Agent Smith is showing the humanly universal expression for disgust. This is supposedly an artificial intelligence. How does he get the brain wiring such that when he feels disgusted, his face, which isn't even supposed to be a real face, contorts into the exact expression that all known human cultures use for disgust? Common sense is not reliable here. This doesn't run on analogies and metaphors. If you want it to actually work it has to run on math.

[Ari Heljakka]: I think there's one important distinction which should be made, and that is the distinction between AGIs being under our control and, on the other hand, being friendly to humans, because the latter is very vague, and I am not sure we all agree what it's supposed to mean. I think Hugo de Garis advocates the idea that humans, as such, are not important but something more is actually what we want; and there is the concept of transhumanism, which is about humans becoming something else. So, then the question becomes, do we actually prefer an AGI system to help people change into something else, or do we prefer it to conserve the current state of humanity? That's a much more difficult question. Which question should we ask: are we able to control the AI, or what are we going to do with it?

[Sam S. Adams]: I think one answer to the question is an old line to a song, *teach your children well*, because as technologists, which fundamentally most of us are, we create tools. When we talk about making something for humanity, or controllable by humanity, humanity is a very broad space, with a whole lot of different opinions about what's right and wrong, what's good and bad, what's beneficial and what's friendly, and what's not. If I invent a new shovel, yeah it helps everybody dig holes easier, but someone also figures out that it's a pretty good machete and takes someone's arm off in the Congo. I get questions like this all the time, people come up and say - *wow, what about this thing when we build it?* They ask the moral question - *should you turn it off?* They also ask – *when it does something bad to me, who will be responsible?* The thing is, if we believe that we as technologists will control the destiny of our creations, we are fooling ourselves because we have never done that ever, ever. Okay, will mankind be able to control it? Ask yourself about any other technology that's ever been created. It is used for both good and evil as defined by individuals.

[Jeff Medina]: So in that question/comment you explicitly spoke on behalf of technologists as tool makers. AGI posses a somewhat different problem in the sense that you aren't creating something that is analogous in that tool sense. Rather it's how people use it. But, if right now you found out I am Ben Goertzel's first successful Novamente AGI, you wouldn't say - *how can I use you*? I would say - *you can't use me! I get what I want, right? They succeeded. I am human-level. I am not a tool for you to use.* The comment that we are never going to be able to control other people's use of our technologies, using a hammer to build a house or smashing skulls, doesn't seem to apply to AGI because it doesn't matter how you want to use the tool, it's what the tool *itself* is thinking about *itself*, if it reaches that point.

[Sam S. Adams]: If it reaches that point, I think we're near Hugo de Garis space.

[Hugo de Garis]: [To Eliezer.] This is a critical question. Are you saying that we, with our finite intelligence level X, that we can change the structure of this super creature such that it always remains friendly to us? Is that what you are saying is possible?

[Eliezer Yudkowsky]: Correct.

[Jeff Medina]: Right.

[Hugo de Garis]: Wow!

[Jeff Medina]: There is even among the software engineers formal verification.

[Karl Pribram]: I think what you're doing is thinking about AI, AGI, in terms of a weapon. If that's the case, we have whole history of what has happened to weapons. And somebody else is going to put up a Maginot Line which doesn't work. And somebody else is going to create something else, and we are going to have a number of AGI systems

created by different parties, like happened with the atom bomb, a kind of equilibrium. I think these things will stabilize.

[Ari Heljakka]: I am afraid I am going to be awfully dull here with my idea, but if we try to approach these questions, which have to do with how do we make sure that the AGI does this or that in the future, then our discussions sounds an awful lot like the historical idealistic philosophers before the age of empirical science. The easy solution to that would be to go a bit further and actually do experiments, and make observations on the behavior, for something which is like a seed AI. The only counter argument as to why we should not do that seems to be the assumption that we'll have an extremely fast take-off, so that we won't somehow have time to make these types of experimentations.

[Eliezer Yudkowsky]: Or, a slow enough take-off that the AGI is smart enough to conceal itself.

[Ari Heljakka]: Sure. But I am certain that there are fairly long states of research where we can just use our common sense, and we'll know that the AI is not quite there yet. I know that Novamente is not quite there yet. There is so much more that we can learn before we come close to creating AGI.

[Hugo de Garis]: I think we should be expecting this now. This is so important.

[Cassio Pennachin]: I have a meta question as it relates to AI as a weapon and AI taking over the world. There is some fairly strong biological evidence that our quest for power has evolutionary reasons, which means that I don't think it's a good assumption to make that an AI will have the same lust for power.

[Hugo de Garis]: How can you be sure?

[Cassio Pennachin]: I'm not sure of anything, I'm just saying that lots of people seem to be assuming that its going to take over the world, that it's a weapon, and I'm challenging that assumption. I'm not going to assume that evolutionary bias is carried over into AI's, even if the AI is achieved through brain emulation.

[Bill Redeen]: I do think we have to assume this is inevitable… the evolution and emergence of AGI.

[Josh S. Hall]: I think it's worth thinking about what happens if a group the size of Novamente can create an AGI and it works. Or, what if Hugo de Garis creates an AGI that works. Or, what if Sam S. Adams creates an AGI that works. If that is the case, there are going to be a billion of them in 10 years. A take-off may not be nearly as hard as you think. If the take-off is going to be soft, you can't start out with the notion that your AGI is going to take over the world, because that will get all the other groups riled up.

[Jeff Medina]: A team of people with a 120 level IQ can defeat an enemy with an IQ of a 160. There is a point where we are smarter, and the enemy is defeatable. If the take-off is slow enough, you really can have an impact.

[Ben Goertzel rephrasing a question posed by Izabela Goertzel]: Isn't it likely that while you're sitting there trying to prove that your AI design is going to be Friendly, during the 48 years it's going to take you to provide the proof, someone will annihilate the world with engineered biopathogens, or with an evil, unfriendly AI or something?

[Eliezer Yudkowsky]: This is what I call the Ben Goertzel problem. [audience laugh]

[Ben Goertzel]: You need to estimate the odds that your AI is going to be friendly, and also take into account the odds of the world being destroyed by some other means while you're working on the problem. Why do you think there are reasonable odds that you would make a provable safe AI before a hundred other teams would make a highly

intelligent AI? – their rate of progress should be faster because they're not stopping to bother with a friendliness proof.

[Eliezer Yudkowsky]: Presuming that my hypotheses are correct, there is a limit to how much you can understand the nature of intelligence and not notice that the AI you're building is going to wipe out the human species. In other words, there is an upper bound on the competence of the teams who are trying build the AI without understanding it properly. If they get too smart they are going to notice that they have to start over and work out deterministic designs. That's the only answer I can think of.

[Ben Goertzel]: So you're saying that all the other teams working on AI, other than yourself, are not smart enough because they should be proving theorems instead? [audience laugh]

[Eliezer Yudkowsky]: I didn't say they weren't smart enough, but I do think that if a team is seriously saying that "our AI is going to take over the world, and we don't need any assurance it's going to be friendly," I think that they have not reached the state of understanding that you get after working on this problem for a few years.

[Ari Heljakka]: I'm still thinking about the previous question, about how we actually need to do real concrete research. A relevant question is to what extent can we continue developing a specific architecture, for example Novamente, and still have a feeling that we completely understand what it's doing. Because, for example, with neural networks, you learn quickly that there is a point beyond which you don't understand the system anymore. So that is one area where you can do very concrete, fairly straightforward experimentation.

[Eliezer Yudkowsky]: I think that if you want to improve human rationality you go off and you become a Bayesian, and you go off and you join the field of heuristics and biases, which I think is already telling us a whole lot more than we are likely to get out of watching young AI's. I don't think that a human who looks at an AI are going to become smarter than the smartest humans that exist today. Maybe they'll gain something like two effective IQ points, and that'll be it. I do want to take a moment and rephrase what I call the Ben Goertzel problem. It's not specific to Ben Goertzel, but I think that maybe a better answer to the questions is that you have to hope that the first team that builds an AI is also smart enough to make it friendly. That's the main source for hope, that beyond some point you notice that you need to make it friendly and that the smartest team out there is doing that.

[Ari Heljakka]: Briefly, I do think that this is again speculation. I think it's very difficult to say beforehand how much we can learn by just looking at it, looking at the system when it is running. It's not as simple as this. We can also devise data mining methods to assist us in further understanding the system. This will provide us with empirical information. I'm not questioning the relevance of this question that we are talking about, but I am just saying that I feel very uncomfortable with the level of speculation here. I still find it kind of fun. But I definitely want to stress that doing these empirical experiments will provide so much insight that I'd rather be designing these experiments right now than speculating a lot about these questions.

[Sam S. Adams]: Question. You talked earlier about designing rules [for ensuring beneficialness of AGIs]. Recapitulating Asimov's three laws of robotics. I'm saying they are in a similar vein. The problem with these systems, is that humans are incredibly bad at writing perfect rules. Now the question I had is, I think there is a safety valve. I think there is a way to build these things and to prevent them from doing horrible things, or at least to

build a pretty good safety valve, if not perfect. And someone said yesterday that science fiction was a form of experimentation of imagination into these alternative worlds. It is such an active genre -- you know, that supercomputerish AI thing that runs amuck and takes over, right? How many things have been written, or in movies, that talk about that. Now the big mistake that always happens in all of these is that the human is taken out of the loop. There is a reason why our project is actually named Joshua Hal Forbin. Because, if you are serious about building these things, you had better keep in mind what happens if you think it's smart enough to take a human out of the loop, because we as programmers suck, we write buggy code.

[Audience]: What happens if the human is taken out of the loop?

[Sam S. Adams]: Well, that's what happened with Colossus, with the Forbin Project.

[Eliezer Yudkowsky]: Logical fallacy from generalizing based on fictional evidence.

[**Sam S. Adams**]: But the point is, we have lots of people who have explored this space, have explored these arguments, without the technicals saying we are going to build one next year. But my question to the panel is, okay, so if we took a rule like - *don't give them any kind of lethal capability without a human check on that capability*, is that a useful thing?

[Hugo de Garis]: That kind of reasoning is valid when the machines are at some sort of human-level competence. But when they are hugely smarter than us, why would they continue to respect humans?

[Sam S. Adams]: It's the *pull the plug* problem, which is dealt with in the Forbin Project. Because he says – I have my finger on the button… you plug me back in or I will drop the bomb. The thing is, as soon as you take man out of the loop, whatever this thing is you create, it has a way to coerce it into do its bidding. Then you can't unplug it. That's what I'm saying. Don't go that far. Put a wall there.