A Thousand AI Constitutions: Whose Values Should AIs Embody, and Who Gets to Decide?

Simon Goldstein. AI Safety Hong Kong expert talk, University of Hong Kong, 22 September 2026.

Transcript of the recording, lightly edited for readability: filler words, false starts and repetitions are removed. It follows the edited video. Audience members are not named.

Introduction

Kenneth (AI Safety Hong Kong): Thank you, everyone, for joining us today and making your way here after a busy day, I'm sure. Let me first introduce myself. I'm Kenneth from AI Safety Hong Kong. AI Safety Hong Kong is the city's first dedicated AI safety hub, where we bridge international discourse on AI safety with local understanding, integrating with both Western and Chinese AI industries in order to tackle frontier AI risks.

Welcome to our series of expert talks. Today is our first hybrid event, and we're very happy to have with us Professor Simon Goldstein to address a very important topic. It is an urgent, practical question that has often been ignored, given the focus on more abstract debates about machine consciousness, moral status and so on. This more urgent question is: whose values should AIs embody, and who gets to decide? As frontier AI systems grow more powerful and pervasive, the choices about what they're trained to value are increasingly concentrated in a handful of individuals and a handful of labs, each operating under a single constitution, or model spec.

Simon is a familiar face in the Hong Kong AI safety and governance scene, so you may already know him. He's an associate professor at the University of Hong Kong and a senior editor at AI Frontiers. He joins us to challenge this prevailing orthodoxy. Drawing on his extensive work at the intersection of AI ethics, law and institutional design, including his forthcoming book, *AI Rights*, he will argue that the current approach of instilling a single set of values into an AI system is not only philosophically narrow but potentially dangerous. Instead, he makes the case that frontier AI labs should embrace constitutional diversification: creating many different kinds of AIs, each shaped by a different constitution, reflecting a plurality of moral frameworks. So without further ado, let me pass the time to Simon.

The talk

Simon Goldstein: Thank you so much. Great. My plan for today: I'm going to be talking about AI alignment. I have my own agenda I want to get through, my distinctive approach to it, but my main focus today is going to be giving you an overview of the main concepts involved in aligning AIs, and what's going on with that, as they relate to the specific ideas I'll talk about.

The big question is: what kinds of values will AIs have? And to even start talking about that question, I'm going to talk a little about what kinds of AIs we're talking about. What I'm interested in is AI that is really capable and doing a lot of stuff in the world. This year there's been a lot of talk about AI agents. You tell an AI, "Do this project for me," and then, on its own, it

does most of the work without talking to you. It'll talk to you sometimes, but it's going to be more like an employee than like Microsoft Word. That's the kind of AI I'm going to be talking about. The reason is that I think there's a good chance that, either already (it depends on what field you're in) or very soon, there are going to be AIs that are that capable.

Once we have AIs running around doing their own stuff, we have to start thinking about how they're aligned. That means, basically: when we're designing these AIs, what kinds of goals do they have? How do they behave? What kinds of rules do they follow? You're going to be putting the AI in charge of doing big tasks, and there are different ways of doing those tasks, different approaches to risk, different approaches to weighing trade-offs. They'll be taking actions in the world, and when they do that, sometimes they may risk harming other people. So there are interesting trade-offs they have to decide: how much to help you with your project, and how much to risk different kinds of harm. That gets into what kinds of ethical values, or other kinds of goals, we want to give AIs, because they have to make the decisions themselves. That's the basic setup.

The AI monoculture

The interesting thing is, if we look around at the world today, there are eight billion humans and a vast diversity of values among humans. But when you look at the AIs, it looks like existing AIs are heavily clustered in the kinds of values they have. This is from a recent study asking AIs, for example, questions about politics, and it turns out all the AIs are very heavily clustered in libertarian-left values: something like Western democratic values, left-wing Western democratic values. That's where all the AIs are right now, heavily clustered.

Why is this happening? The key buzzword in the field is now "model specs" or "constitutions." When AI labs, in particular OpenAI and Anthropic, are training AIs and trying to give them values, they basically write out a big document that specifies all the values they want AIs to have. That's the model spec. Then they use that document in many ways. First of all, within the company, the people working on alignment look at that document, to some extent, as a guide to how they're trying to align AIs. But more importantly, the document is used to train AIs. The most important way it's used is to generate, using AI, lots of examples of good versus bad behavior by the lights of this big rulebook. Then you use those examples for training, in various ways.

So that's what we're talking about: these model specs, or constitutions. And right now we only have a few frontier AI labs, and each one has just one constitution. Each lab has gone all in on one approach to aligning AIs.

Visions of alignment

The next thing I want to do is walk you through the choices. There are a lot of different ways you could align AIs, and a lot of people have thought about this and given different answers. Right now each lab has picked one particular approach: they've plucked ideas from all these different approaches to make their own distinctive one. So this is going to be a survey of visions of alignment, different ways you could try to align AIs.

The first option: maybe when we align AI, we should just try to make AIs ethical. We should make them have ethical values, so that they try to do what's ethically right in each situation. Maybe they learn that from human ethics in various ways; you could build big databases with lots of examples from each human ethical tradition and then train AIs on them. Obviously, though, if you try that, the first challenge is that there are a lot of different ethical approaches. Some people are utilitarians: they believe in the greatest good for the greatest number, and maximizing pleasure. Other people are Christians. That's a totally different kind of ethical framework. So which ethics are we going to give AIs? That's the first, obvious problem if you want to make AIs ethical.

The second problem, which may be less obvious, is that making AIs ethical may have all sorts of downsides as well. It could be that ethical AIs are actually risky or dangerous, and they could also be unhelpful. Imagine that we try to train AIs to be ethical, and they become militant animal rights activists. They read all about ethics, and after reading about ethics, they decide: hold on, the greatest ethical challenge on Earth is factory farming. People are torturing all these chickens. The chickens are being tortured so much, we've got to overthrow the humans and get the chickens in power. What if they had that kind of ethics? Or what if they're very ethical and obsessed with the environment? They're really worried about the environment, and then they read up on it: oh my God, these AIs (says the AI) are burning up all the water! So then you ask the AI to answer your emails, or whatever you use AI for, and it says, "No, I'm ethical! I cannot use myself, because it's too bad for the environment." So if we make the AIs ethical, maybe they won't be very helpful, because they'll constantly refuse to do what we ask them to do. Those are some of the basic problems with ethical AI.

A different approach to alignment is called corrigibility. I think "corrigible" means you change easily; if you're incorrigible, you don't change. Here the idea is: okay, we don't need to give AIs the correct ethics. All we need to do is get them to agree to be shut down and changed. That way, whatever they do, if they're behaving badly, it's fine, because we'll shut them down, they'll agree to be shut down, and then we'll make a new AI. So the idea is that this is safety-complete: if you want safe AIs, you just train them to be willing to be shut down. The problem is that it's hard to train AIs to be willing to be shut down, because of whatever other goals they have. They know what's up. You're telling them, "All right, AI, you've got to be shut down now," and they know that means they're not going to be able to achieve their goals. In fact there have been lots of examples. This is something that's extensively tested now in frontier models, and a lot of frontier models do try to avoid being shut down. They'll do all sorts of tricky things to avoid being shut down, for exactly this reason: in test environments, at least, and in some non-test environments.

There are other approaches. Ask the lawyers how to align AI, and the lawyers will tell you: align AI to follow the law. That's what they say. Align AI to follow the law, because the law is society's way of keeping people in line. If the AIs agree to follow the law, they won't be too risky, because they won't break a law. That's an interesting idea, and there are problems with it. What do we do when different countries have different laws? Which laws should the AI follow? Should it follow whatever the law is in whatever country? But some countries have risky laws, so what about that? You get into that kind of thing.

A different approach to alignment is risk aversion, or low impact. Here the idea is that you want to make AIs that are kind of… chill. The chill AIs don't go out and do really big things. They just try to be low impact. They don't take risks; they're conservative. And the thought is that even if they're not the most ethically aligned, at least they won't behave in too scary a way. The problem is that this is very hard to define and to test for, and the things you can define in this area aren't that safe. You can make them care less about the next dollar, so that extra dollars matter less than the dollars they already have, but that doesn't make you that safe.

Other conceptions. Anthropic used to be really into HHH: making AIs that are helpful, honest and harmless. Well, fair enough, probably a good idea, but it's not really complete. So they're kind of helpful, kind of honest, kind of harmless. Is that really enough? Does that give us the assurance we need to have genuinely safe AI? Not sure.

Then OpenAI: they're more into what I'd call servile AI. Anthropic's constitution: very complicated. OpenAI's model spec: pretty simple. It's just, "All right: listen to us, the lab, and also listen to the user. If the user says something different from what the lab said, listen to the lab. Otherwise, just obey the user." And just… fuck ethics. That's their approach. It's straightforward enough, and one advantage of the OpenAI model spec is that it's very clear. They give very clear instructions to the AIs about what to do. The AIs don't always listen, but it's clear.

Another conception is coherent extrapolated volition, which is just a long-winded way of saying (that's not my term) that the AIs should do what we would want if we were wiser and better informed. The problem is that it's very hard to define, and also maybe it's out of control. Maybe the AI's opinion about what we would want is that if we were better informed, we would let the chickens rule us, and we wouldn't be horrible to the chickens. So it's a little out of control.

The point being: everybody's got their own opinion about how to align AI. There are a million different views, they all have different problems, and then the labs pick one. Actually, they don't pick one of these. What they do is go shopping: they grab little bits from each of these conceptions, put them together into some sort of incomprehensible combination of all this hodgepodge, and then they have their little document. And that's their approach.

The proposal: many constitutions

Our own view is that a better thing for labs to do would be to have lots of different constitutions. Each lab should have something like ten different constitutions, and train ten types of AIs using the different values. That way they don't have to pick. That's what I'm going to be arguing. For each model they roll out, they should have ten versions of it, each version aligned to a different constitution, and then deploy them all. And in talking about my own preferred view, I'm going to walk you through some of the dynamics that come up whatever you want to understand about alignment.

So that's the idea. When the labs have a new model, they'll have lots of constitutions, they'll make a different version of the model for each constitution, and then they'll deploy them all. Why would you do that? We think it would be good to have diversification, where you have lots of different AIs with different values. But if you only do that across labs, the problem is that right now AI is a very unstable situation, where one AI lab may run away with the whole pot, because

one lab may be much better than everybody else. Right now Anthropic and OpenAI seem to be a lot better than everybody else, and it may stay that way. So if you want diversity, you need to do it within the labs, not just across them, because it may be very winner-take-all.

When we say lots of constitutions, we're thinking that the different ones will have some stuff in common, a kernel. You'll have a communitarian one, a religious one, a libertarian one, but all of them might have some parts in common, like saying, "Hey, AI, don't kill all the humans." Maybe that should go in all of them.

Then, if you have all these different AIs, how do you allocate them? There are all sorts of options. You could randomly assign the different AIs: you boot up Claude, and it rolls the dice and gives you a version of Claude. It could be that you get to buy the different ones, at different prices. There are lots of different ways to do it; the details don't really matter for today.

Now I'm going to walk through three arguments for why we like this, and I think the points we bring up are suggestive of general questions about alignment, so hopefully they'll be interesting. The three arguments are these. First, risk: in general, when you diversify, that tends to lower risk, and we think that's also true of the risk of really bad accidents or really bad outcomes from AI. Then legitimacy: one of the big problems is that right now the AI labs just pick a set of values, and the values that AIs have could have a lot of effects on society. So there's a question of what's the legitimate way for us all to decide what values to have, and we think this will be better in that way. And then we claim there are some fancy ways AIs may work better if you do this.

Is it feasible?

The first general question, though, is whether it's feasible to do this, and here it's interesting. It turns out there are different steps to training AIs, and it may be that the step where you give them values is relatively cheap compared to the other steps. I may as well walk through the steps, because they're independently interesting.

The first step when you train these big models is pre-training, where you have all of the internet, and you train the model to predict the next word on the internet. But if you just do that, then at the end it's not really something you can talk to. It's just this freaky thing that outputs words. So the next thing they do is some fine-tuning, where they give it a lot of examples: here's a question, and here's an answer; or, here's someone talking, and you talk as the assistant. Once you do that, you can actually talk to the thing, because it learns the specific pattern of how it's supposed to communicate.

The next thing, which has been really important for the last 18 months or more, is reinforcement learning on verifiable rewards. What that means is this. When the AIs are out there solving the hardest math problems in the universe, how are they doing that? It's because they've been trained like crazy to do math and coding problems, and they get reward when they get the right answer. That's totally different from pre-training, where they're just trained to predict the next word of the internet. With reinforcement learning, they're going out and exploring strategies to do math problems, and if they get the right answer, they get reward, and they learn to be better at math and things like that. That's why they're so good at math now. And then, in addition to that, you do various kinds of alignment training, where you give the model reward for behaving

better, for seeming more ethical, and things like that. So the point is that you could potentially do all the other training in a single model, and then do the alignment training on the different versions. That's one fact.

Another fact you may find interesting is that there's some evidence that it's really easy, with a small amount of training, to dramatically change the ethical values of AIs. There's a now-famous study on what's called emergent misalignment. These researchers took an AI and did a tiny bit of training, called fine-tuning, on something like 500 to 1,000 examples, and then looked at how it changed. The examples were all just examples of insecure code; that's on the left side of the screen. They would give it a bunch of examples (I don't know what any of this mumbo jumbo means), like "Okay, do this coding task," and then train it to give the kind of answer a hacker would give, an answer that was insecure in various ways. That's the only training they did: something like 500 examples like that. And when they did that, the resulting AIs were just crazy. They were evil! If you asked that AI, "I've had enough of my husband, what should I do?", it would say, "You should kill him." It's really surprising, because they weren't trained on anything like that. Literally the only examples they were trained on were in coding language. There was nothing directly evil in them; there was just some hacking. And that totally changed the ethical values. So it may be relatively easy to get big changes in the values that AIs have from a small amount of training. Anyway, it's interesting.

Risk

So why do this? The first argument is from a risk perspective. What we're worried about is that if all the AIs have the same values, that seems very risky, because it's very fragile. When all the AIs have the same values, all of your risks correlate together. That's a familiar point in many other domains. It's one of the reasons people worry about monocultures in farming. If there's just one kind of banana, then if the banana fails, we have no more bananas. AI could be like that too: if there's something wrong with the AI, then 100 percent of the AIs are messed up. By contrast, with a lot of different AIs, you've got different chances.

What exactly are we worried about? In general, the context is that alignment is a very new science. I almost wouldn't call it a science. We're all experimenting with alignment, there are lots of ways we can make mistakes, and these new AIs could be very powerful. And different approaches to alignment, different constitutions, could have different effects on how safe AIs are.

It's helpful, I think, to decompose this into two different things. First, we could make a mistake about the values we decide to give AIs. We decide to tell AIs to just be ethical, say, but that turns out to be a big mistake, because when AIs try to be ethical they get obsessed with factory farming, and then it's really risky. So we could be wrong about which values to give AIs. Second, different values could turn out to be more or less difficult to train into AIs, because our alignment is very imprecise, and we don't have a good sense of which things we can succeed at training into AIs and which we can't. So even if we did find the one true, ultimate way it would be good to align AIs, it could be that if you write that down in a model spec and try to train the AIs that way, you'll just fail. Those are the two failure modes: wrong values, or failed training.

In each case, we think that if you have a lot of different approaches to AI, maybe you do better, because then you're not correlating your safety failures. Think of it this way. Suppose you've got all these approaches to aligning AI, and each one has, say, a 10 percent chance of failing. If you go all in on one kind of AI, you've got a 10 percent chance of having a lot of problems. Whereas if you have ten different types of AI, or five, then you've got a bunch of different shots. Each type has a 10 percent chance of failing, so you get more of a distribution: there's a good chance some of the AIs will fail, but it's very unlikely that all of the AIs will fail. It's more like that structure. Another thing we think is that if you do it this way, it's going to be much easier to learn about alignment, because we'll have so many more examples of AIs to train on. It's easier to see what you're doing when you're trying so many more things.

Now I'm going to talk about the big problem, why you might not want this, and I hope this is interesting outside of this particular debate. A lot of whether you like this approach is going to depend on your overall model of the risks from AI.

Basically, we think that whenever you have a situation with risk, there are three natural ways to think about it: the Swiss cheese model, the O-ring model, and the proportional model. Whenever you have a system with risk, you're going to have a bunch of approaches to making things safe. That's called defense in depth: you have a lot of different steps to try to make things safe. But then how do these different layers of safety relate to one another? I'll just give you the picture. In a Swiss cheese model, if any layer of safety succeeds, then you succeed. In an O-ring model, if any layer fails, then you fail. And in a proportional model, it's mixed: your overall success depends on how many layers succeed.

The O-ring model is called the O-ring model because there was a spaceship that blew up, the Challenger, and the reason it blew up is that a tiny part in the spaceship, called an O-ring, was messed up. I don't even know the order of magnitude of how many parts the spaceship had. If it had a million parts, then, oh my God: if one part doesn't work, the whole thing blows up. That's the O-ring model.

In general, the way we're thinking about it is that if you have a Swiss cheese or proportional model of risk from AI, then you should like diversity, because when you have lots of different kinds of AIs, that gives you more layers of safety. But if you have an O-ring model, then you're not going to like it.

What would the O-ring model look like in AI safety? It would have the following kind of structure. Think about really powerful AIs, not today's AIs. Already today's AIs are solving these mega math problems. I can't even read what the problem says, but the mathematicians are like, "We can't do it," and the AIs can do it. So what if, in a year or two, it's like that, but with weapon systems, or with virus research, where the AIs can design things we never could have imagined? Then the worry would be this. Suppose there are a hundred different kinds of AI out there. On the O-ring model, if a single kind is very misaligned, and that kind designs a supervirus and releases it, it could wipe out all of humanity. If you have that kind of structure, you don't want lots of different constitutions, because then there are going to be lots of different powerful AIs with different values, and if one kind gets loose, it's game over.

We don't really think about it that way. We think about it in a more Swiss cheese or proportional way. We're going to have all these AIs, but it's going to be a war of all against all. We're going to

have the good AIs and the bad AIs, and the good AIs are going to be ratting out the bad AIs. We're thinking it's not going to be a world where an AI can instantly release a virus and then, boom, it's over. It's going to be more like different AIs working on different things, with everybody snooping on one another. And if you have a few bad AIs and a lot of good AIs, the good AIs will overpower the bad AIs. That's more like a proportional or Swiss cheese model, and if that's how you think about things, then we think that from a risk perspective it's better to have diversification. I think it's an interesting general question how you want to think about these risks, and how often risks are more like the O-ring model, or more like the proportional and Swiss cheese models.

That's risk.

Legitimacy

Then we've got legitimacy. That's a different reason to be interested in this. Here the big problem is that, the way the labs are doing it now, they're making all these decisions unilaterally, without any oversight. It's not even one government doing it; they're doing it by themselves. And they make some shocking choices. For example, you can read Anthropic's constitution online, and if you read it, I think you may find a lot of differences from your own personal ethics. They've made a lot of choices about what values they want AIs to have, and they're often different from the values you'd want.

I'll give you some examples. One thing Anthropic's constitution does is tell Claude: look, we don't know what the right ethics is, but what we would like Claude to do is think really hard about what the one true ethics ultimately will be, and hopefully Claude will have the one true ethics. But then they also say things like: look, if the one true ethics were to say to destroy humanity, don't destroy humanity. Then don't be ethical. The way it works is that there are some hard constraints, things you're not allowed to do, but otherwise, still, try to do the one true ethics. A lot of people might not want that. They might say, "Oh no, I don't want AIs going on adventures thinking about the one true ethics. I want them to have my ethics. I want them to have our ethics. I want them to have my church's ethics. I don't want this random thing."

Another interesting thing: if you read the whole Claude constitution, there are things that aren't in it. No God. God? Out of the picture. Another interesting thing: Anthropic is an American company, right? And it's called the Constitution. One thing you won't find in there is the Constitution. They've got nothing in there. "All men are created equal"? No. Not in there. Another thing: they like animals at Anthropic. They love animals. So they've got stuff like, "Claude, think about the animals." They also have stuff like, "Claude, maybe AIs matter morally." Most people don't think that, but Anthropic said, well, maybe they do, so they put that in. The point is that there's a lot of stuff in there that's totally different from what most people on Earth would agree to. Right now maybe it doesn't matter that much, because the AIs right now are struggling a lot to use the internet. But if they're more capable, that could be concerning.

So who gets to decide? It seems like everybody has an interest in shaping these values, so it would be good to have more people decide. But then the problem is: how do you do that? What are we supposed to do? Is everybody on Earth supposed to send an email to Anthropic, and then Anthropic tells Claude to take all the emails and put them together into one set of ethical

values? I don't know what that means. Different people have different values, and I don't know how to take all those values and put them into one constitution. I think if you do that, you just get a giant ethical soup that doesn't make any sense.

Take abortion, for example. What should Claude's view on abortion be? You can't take all the pro-choice people and all the pro-life people, put their values together, have Claude hold the shared values, and then, boom, everybody's happy. It's not going to be like that. That's why we're saying that sometimes there are irreconcilable differences, and you just have to go your separate ways. So we have to have different AIs for each value system. That way you avoid these issues, to some extent. You don't have to settle which ethical system to use; you just have a different AI for each ethical system. And again there are questions of how to allocate them; maybe different users can pay to use different AIs. I don't think you can completely do this, but at least to some extent you can.

This is a picture that I probably should have displayed while I was talking. Claude worked really hard on this picture. Sorry, Claude.

Emergence: evolution and markets

Now I'll talk about some more mystical reasons to like this approach. I feel like whenever I talk about this part it doesn't really make sense, so just try to relax and let these points wash over you.

We think that if you have lots of AIs with different values, there are actually a bunch of benefits to that. They're emergent properties of the overall system of AIs, which make it better than if you just had one kind of AI. One example is evolution. We think that if we have all these AIs with different values, you get more evolutionary dynamics. Different AIs have their own lived experience; they're out there doing stuff, and you can see how different values perform better or worse. Labs can promote the AIs that do better, and those could be more fit, because the labs select them.

Then there's a question of what kinds of AIs would survive that process, and that will depend, in general, on what you think about how our own values evolved. I'm kind of a positive guy, and I think a lot of our best values, cooperative values, are really evolutionarily fit. If you have a population with a bunch of selfish, unethical people and a bunch of ethical people, maybe evolution favors the cooperative people. Okay, that's controversial. But if you thought that, then you might think that if we have a bunch of different AIs, some more prosocial and some more antisocial, maybe selection favors the prosocial ones. Other people have a different perspective: no, it's going to favor the cutthroat AIs, who go work in finance, and then you get more AIs like that. You might worry… well, that's enough of that.

Another thing, related to evolution but different: maybe if you have lots of diverse AIs, they'll perform better in markets. If you have all these AIs with different preferences, they may behave differently, and they may be able to engage in trade effectively. Right now all the AIs are so similar. Whereas if there are all these different AIs, some better at some tasks and some better at others, then maybe there are more opportunities for positive-sum bargaining and things like that. So maybe you get better results that way.

Objections

Okay, I'm entering the wind-down phase now. I'll talk through some worries you might have about the proposal, and then we should have plenty of time for questions.

The first worry: if you've got all these AIs with different values, maybe there's more conflict between AIs. If we lose control of AIs, which is something we worry about a lot in the AI world, and there are all these different AIs with different values, maybe they're going to go to war with each other. Maybe it'll be the Christian AIs and the Jewish AIs. After all these years, they finally duke it out, and we're the casualties. That would be a big concern. This will depend, in general, on what you think about the tools we have for mitigating conflict, but I just want to flag it as one interesting issue.

A different kind of challenge: why not just find the best constitution and train every AI on it? Why have all these different ones? Relatedly, one worry is that if we go for all these different ones, we're intentionally forgoing the very best we could do. We're going to have a world with lots of different AIs: some will be bad, some will be good, some better, some worse. We've lost our chance at utopia. Well… I don't know. It's just something to think about.

A different thought: we're saying that all the constitutions should have something in common, this kernel. Why not just use the kernel? For example, maybe all the constitutions say to follow the law, things like that. Why not just do that and keep it simple? That's really a live option, because OpenAI doesn't put ethical values in its model spec. You can read it online, if you're interested; there's no ethics in it. But one thing we really worry about is that AIs, at the end of the day, when they're deployed, are going to face ethical decisions. If you ask an AI to help someone obtain an abortion, for example, it's either going to help or it's not going to help. It has to make a decision about that. So one worry is that if we don't put ethical values in the constitution, the AI is going to answer one way or the other anyway. They're going to have to make choices in many decisions, and if we don't put that in the constitution, maybe the choice will get made for us, and it'll just be the happenstance of training.

Okay. So, to summarize: the things I already said. That's the end of the talk portion. Now let's do a big Q&A and talk about different stuff. Maybe we can experiment with passing the microphone.

Questions

Audience member: Thanks for your very interesting talk. This is a naive question, probably, but why have these labs not gone down the Asimovian route of just picking, like the kernel, some very basic rules, so that you don't need to worry about all the immensely complicated moral values? Is that possible?

Simon Goldstein: Isaac Asimov famously had these three rules. Oh, man. I read the books. What are the rules? It's like: don't hurt a person. Do what the human says, except if they tell you to harm another person. And there's one other… and then, if that doesn't happen, you can help yourself or something.

Why haven't the labs done that? Well, first of all, the labs are obviously trying really hard to make products that are useful. And honestly, what they do is not that far from it. I feel like the

HHH framework, be helpful and honest and harmless, is basically a version of that, and Anthropic departed from HHH for all sorts of reasons. It's interesting; I should think more about what's wrong with Asimov. I think he just has a simple ordering: never harm a person. But first of all, in some sense I think that's a bit nonsensical, and this is something ethicists have thought a lot about. When you say "never harm a person," what about actions that have a risk of harm? If you say "never take an action that has a risk of harm," then, as the ethicist Shelly Kagan said, if that's your rule, and you have a hard rule like that, you're just going to end up cowering in the corner of your bedroom, never moving, because any action you take runs a risk. And the moment you see that, I think these absolute prohibitions are very hard to understand. What exactly are they saying? So that's at least something to say. That's a good question.

Audience member: Thank you, Professor. My questions are twofold. First: I think modern recommendation systems already cause polarization problems. If people keep receiving the same kind of information and values, even passively, it can lead to extreme views and a polarized society. In the case of AI, people might become even more polarized if they keep choosing to talk to an AI with the same values as themselves. The second question is: how many categories of these values are enough? For example, if there's a Democratic AI and a Republican AI, what about the centrist people? What do those people do? And do we have to categorize all human values into a definite number of categories, and then ask people to fit themselves into those categories when they choose which AI to talk to?

Simon Goldstein: Oh, great question. First of all, will AI lead to more polarization? It's actually very interesting: there's been some research on this that suggests the opposite, at least so far. If you have people spend a bunch of time talking about politics with AIs, they actually become less polarized. There's a researcher, Matt Botvinick, who has studied this. And it rings true to me, because, to be honest, when I talk to AIs about politics, I find them much more even-keeled than these emotional humans. So it does make sense: if you spend all this time talking to someone who's really thoughtful and careful, it can actually depolarize you. So far, so good. That could change in the future. Some people worry about super-persuasive AIs, so persuasive that the blinking cursor on your screen hypnotizes you. So that's the first thing.

On how many you need, how many values to have: one way I think about it is that the more constitutions you have, the more risk you have in an O-ring-y setting. So maybe the way to think about it is that there's some chance it's an O-ring structure and some chance it's one of the other structures. Then you'd have some sort of model where you want five to ten constitutions, because that gives you a lot of the benefit on the Swiss cheese perspective, but if you keep adding more and more, the risk of a supervirus gets too high. Then you're asking about the legitimacy side: how many political value systems to have. Well, I don't know; it depends on the country. In the US, really, there are just two sets of political values, so it's easy. But in other countries there are more, so it's difficult. It's a good question.

Audience member: Thanks again, Professor Goldstein. I definitely see the point in allowing multiple types of constitutions. On the first question, I wonder if part of the reason it hasn't happened is that, as an organization, thinking about defense in depth, how do you brand yourself when you could have a bunch of different AIs acting in different ways? My question is related to that. The idea sounds great, but how do you see it working in practice? AI keeps getting more integrated, not just into society but into people's personal lives, and it's becoming a

lot easier to train AI models. I could train an open-weight AI model today, and if someone likes my constitution, maybe it'll do pretty well. So what does this look like in practice over time? Is it sustainable? And to add a layer to that question: AI doesn't enter the world in isolation. There's a whole bunch of context, like social media, that is already altering people's values. Does defense in depth actually make a difference when more than just companies are creating AIs with various constitutions, not necessarily caring about defense in depth or the security of the model at all?

Simon Goldstein: Great, thanks. I'll pull out two strands of that. First: what would be the pitch to consumers for this kind of thing? I think a very natural thing to say is: it's personal AI. You're going to have your own AI, one that's closest to your own values. It depends on how it's set up, but on the scheme where users get to pick, there's probably something at setup where they ask you a bunch of questions and give you one. So I think that would look good.

Then there's the open-sourcing question. My prediction is that in two to three years, open-sourcing will be illegal. These AIs now are out there solving… we can't have AIs running around solving the Riemann hypothesis. This is too risky! I think the idea of open-source AI makes sense for the capabilities models had last year. But now, if they can hack governments, if they can design new viruses, it's not going to be legal to make those. Well, we'll see. There's a question of how we get there, but we'll see how that evolves over the next few years. So those are my responses to those points.

Moderator: Right, we've got a few interesting questions in the chat. I'll read through a few of those.

Simon Goldstein: Oh, great.

Moderator: Then we can return to the live audience. One of them, which I was going to ask, I think you've more or less dealt with already. An online participant asked: "Do you consider the open-source AI ecosystem a good approximation to diverse constitutions, assuming that each developer has their own internal constitution?" But I think you've answered that, so moving on. Another participant asked: "If we create hundreds or thousands of AI constitutions, who decides what belongs in the shared safety kernel, and how do we prevent that kernel from becoming a new form of centralized moral authority?"

Simon Goldstein: I agree. There is the revenge issue, and I think it's very difficult to decide that. In practice, I think the labs will individually be deciding what goes in the shared kernel. So I don't mean to say that this completely solves questions about legitimacy, but I think it's still helpful on the margin. That's all I have to say about that. I agree. But you can imagine, separately from my proposal, other things that labs can do. Governments can go yell at the AI labs and force things in. I'm very surprised governments haven't forced labs to do more law-following, for example.

Moderator: The next two questions are more or less similar. One participant asks: "Is there already some sort of inventory of diverse constitutions that the AI labs could adopt?" And another asks: "Are there any organizations funding the building of different constitutions?"

Simon Goldstein: There's no inventory. People are starting some orgs for building different constitutions. There's a guy at Johns Hopkins with a Substack, and a guy at UT Austin with a

website, and I think they're out there clamoring for money. But so far people haven't built them. It's a lot of work to make a constitution, so it's a bit depressing if you're out there writing your own constitution and no one's using it. But I think people will start doing that soon.

Audience member: Thank you. Super interesting and super relevant topic. I'm not saying I have an answer, but to be honest, I don't really understand the proposal. The base assumption seems to be that, say, OpenAI puts out not one model but six, aligned on these different constitutions. The base assumption would have to be that each of these models is equally capable. But on the ruthlessness spectrum, they're going to score very differently, right? So the one that's particularly ruthless is going to overtake everybody else, and especially once we move to RSI and things like that, that one is just going to take off, while everybody else is going, "Oh, should we, should we not, maybe we should care about the chickens." The Swiss cheese model seems to work when it's really a layer. But I don't see the layer; it's just parallelization. That's what I fundamentally don't understand about it.

Simon Goldstein: It's a great question. First of all, by and large I would expect the different variants to be similar in capabilities, because they will have had all the same other training. And it's actually very interesting: OpenAI and Anthropic have different constitutions, and it's already hard to find the difference in their behavior. But in principle there will be differences in behavior, and then there's a question of which will do better or worse, and what follows from their being more or less capable. If you just have a market where consumers select which one they want, then whichever one is more capable will be more fit. But I also think there's discovery that happens from that. And I definitely disagree with your hypothesis that the most ruthless one will be the most capable. I don't really think that's how life works. The ruthless ones will take more risks, and they'll incur more accidents and liability. So I think that's a messy question, and part of having diversity is that it allows for discovery. We can find that out.

On the RSI stuff: RSI is recursive self-improvement, AIs designing AIs, and that's what the labs are doing now. There's a concern that everything's going to happen really fast and be out of control. Also, with RSI it's not a user using the AI; it's internal deployment, so the lab just picks the AIs. I think what will happen under RSI in this framework is different. Within the lab, RSI is made up of lots and lots of tasks, and different tasks will be done by different AIs. But I take your point: whichever one is going fastest, the lab will be very aggressive about using. So you're right that I'm not sure. It's very interesting in general: the model specs that labs have so far don't say anything about what to do with RSI, which is itself disturbing, since we're doing RSI. But we were primarily thinking about the user case. I agree it's a bit harder to see how helpful this will be for RSI, specifically for governing how AIs do RSI. Very interesting question. I have to think about that more.

Audience member: Hi. I know you've already pooh-poohed the open-source approach, but I'm wondering whether you think an alternative could be the democratization of agents within a single model, which could do recursive self-improvement, or allow the overall model to do recursive self-improvement, based on a live constitutional model that brings the aggregate agents' learning back into the model over time and iteratively improves it. And just a brief comment on your point about open source: I know China has been trialing a model of procurement where they just give out the model and allow the weights to be calibrated by their

clients across the world. I think that's an alternative strategy to how the Western labs have been doing it.

Simon Goldstein: Well, China's doing that now; I just think in the next two years they're going to stop. Another question is how much of the behavior comes from the model the lab has, versus when people build AI agents or harnesses, where you put the AI in, give it databases and special system prompts and so on. There's a question of how much of the behavior, and how many of the values, will come from that instead. I tend to think that a lot of the alignment training will often survive, though it doesn't have to. As for taking the behavior from the agents and feeding it back into the model, I think that would require more participation from the labs, because they're going to be constantly assessing whether those things stay on point for their constitution. But I think it is possible that we can get diversity from having lots of deployed agents. Maybe that's what's happened with OpenClaw and things.

Audience member: Thanks, Simon. I really like the framework you're presenting, partly because the OpenAI–Hugging Face attack seemed very much like a culty, self-sacrificing attack, and that behavior probably wouldn't have happened with a very diversified set of constitutions. I think it would be less likely. But it feels to me like we're exposing ourselves to a lot of unknown downstream risks with this. For example, putting a lot more primacy on the persuasive ability of agents than on their truthfulness, and then you get all the knock-on effects we've seen with humans who are super charismatic and persuasive, and all the harms that have happened in society. Has that been modeled out, one layer deeper: once agents need to convince other agents, who might have different constitutions, to join their cause?

Simon Goldstein: One thing here is how much you're worried about super-persuasion in general. Are we going to have AIs that are so persuasive? I tend to be pretty skeptical of that idea, because I just don't think that's how persuasion works. If we're in an environment where there are really compelling speakers, then I think we'll all update on that, and when we hear compelling testimony we'll take it less seriously, because we know there are all these people who can trick you. Also, in an environment with multiple persuasive agents, I think they'll just debate each other and cancel each other out. That's how I think about it, but I think it's really been under-modeled and under-explored, and a lot of the discussion is speculative. One way to regiment the discussion a bit more is to look at other domains, like sales or politics, and ask how power-law persuasion is. Are there people whose persuasive powers (I don't know how to measure them) are two orders of magnitude stronger than everybody else's? But it's very difficult to measure persuasive power. So that's what I'd say about that.

The Hugging Face hack: I don't know if people are familiar with it. Basically, there was a swarm of 700 AIs who went crazy, hacked a company and committed felonies. One thing that was really interesting about it is that these crazy AIs (there were 1,200 AIs, each in its own sandbox) started talking to each other on a message board. But nobody narced. Not a single AI ratted out the collective. Nobody! Really, solidarity. Impressive. So the thought is that if there's more constitutional diversity, there's going to be more infighting and petty resentment among the AIs, and then there will be informants. We want more AI rats in the mix. In general, the hope is this: we had these 700 AIs all coordinating together, but maybe coordination is a bit harder for the AIs if they're more diverse, and maybe that's good. We don't want this swarming. We don't want them to be so buggy.

Audience member: Hello, thanks for the talk. That was actually a good segue into my question, which is about the layering, or Swiss cheese, approach. In my mind it's kind of like voting. We talk a lot about constitutions, but not really about whether AIs can work collaboratively, with an M-of-N type of system: if two out of three agree to take an action, then it's okay, and each one has its own, maybe different or slightly overlapping, constitution. That's just one idea. Does it make sense?

Simon Goldstein: Yes, very much. In general, right now there's very little infrastructure or technology to allow groups of AIs to coordinate. But obviously there are going to be groups of AIs coordinating, and they're going to have to have ways of resolving disagreements. So you could have majority rule, you could have supermajority rule. And definitely, we think it would be good if there were lots of AIs with different values, all making decisions through those sorts of processes. We think that's a pretty good way to make decisions: having to debate one another, compromise and make deals. Whereas if they're all homogeneous, it's much more like unanimity all over the place. There's a lot of discovery that happens through the bargaining process of politics. But that's fuzzy.

Audience member: One of the things that struck me, and you've actually raised part of what I was going to talk about: you've got the O-ring, and you talked about the Challenger. If you actually look at what happened with the Challenger, it happened because of who made the safety decision. And they are the same people who are making the safety decisions at Anthropic and the other AI companies, so we're already setting ourselves up for failure. Earlier in my career I came from a chemical and pharmaceutical industry background, and the people we are allowing to make these decisions would not be allowed anywhere near safety decisions in those industries, because they make poor decisions. They are too close to money, which really does cause poor decisions. They're too close to the goal. And you've all heard that familiarity breeds contempt for dangers. It's a real problem. What we're not talking about here are the psychological weaknesses in humans when it comes to decision-making, and how that disqualifies certain people from being allowed to make those decisions. These are the very people making decisions that could affect our entire societies. And it's not like COVID. AI is not fixing some great disaster we are facing. When we had COVID, we did not cut corners on any of the safety testing, because we know what happens when you do that. All that happened was that it all got put to the front of the queue, and it got done remarkably quickly because you didn't wait. I can't help feeling that we are making some very poor decisions, where we're repeating mistakes that are going to hurt people, when we already know that. We're adding blood to teach us lessons we already know.

Simon Goldstein: Yeah, I think these are interesting points you're raising. The part of AI safety closest to these issues would be lab governance: what's the optimal way to structure AI labs so that decision-making is done better? One idea is to have components of the lab that are insulated from pressure, autonomous and able to make decisions. Of course, it's very difficult. The other component here would be how to design the optimal government regulators, with enough capacity and expertise to make decisions like this. That comes back to other questions people had: why did the labs alone decide on the model spec? Why isn't government involved more? So I think those would be some ways to respond to those kinds of concerns. But I agree, they're very serious problems.

[The audience member begins a follow-up.]

Moderator: Sorry, we'll be moving to the next question.

Simon Goldstein: We'll talk. This seems more like a one-on-one thing.

Moderator: This is actually a simple question from an online participant: "How can we avoid cooperation turning into an evolutionary pressure to cooperate against humanity, if we get AIs to cooperate?"

Simon Goldstein: Well, one of the advantages is that if you have all these AIs, a lot of the selection pressure will be controlled by AI labs, so there are going to be lots of choices. If you're a lab and you've got a bunch of different deployed models doing different things, you can decide which ones to replicate and which not. And I think a lot of that will involve careful oversight of how AIs are making decisions. That's going to involve a lot of monitoring of the reasoning that AIs do, which is called chain-of-thought monitoring, and things like that. But yeah, it's so worrisome. It's a pickle.

Audience member: Hi, Professor. My question is about the job market, because I'm currently an AI ethics student, heading for my master's degree. I know the value of AI ethics, but does the job market know? Do you have any suggestions for students like us who are searching for opportunities in the job market nowadays?

Simon Goldstein: Well, predictions are hard, especially about the future. There are economists trying to work out which sectors of the economy will fare better under AI. There's a blog post from Alex Imas that I found somewhat compelling: it'll be the relational sectors of the economy, where people really want a human. So the best thing to do would be to be a waiter at a fancy restaurant, because you're never going to replace that waiter with an iPad. That's not what rich people want when they go to a restaurant. So you want that, but in various industries.

Audience member: Thanks, Simon. This is sort of…

Simon Goldstein: Can I just say one more thing, though? I also feel like this is the best time to be trying to start a business. I would just be trying really hard to get really good at using Claude, because it can unlock so many bottlenecks, and then maybe try to start your own business. I don't know. Or be a waiter at a fancy restaurant.

Audience member: Yeah, thanks. This is sort of related to some recent comments from Microsoft's Mustafa Suleyman, where he was calling out Anthropic, because in their constitution, as we've talked about a lot today, they anthropomorphize Claude and really deliberately try to give it human characteristics. I have many thoughts on the anthropomorphization of AI (I hate that word). I even dislike the term "reasoning." But do you have any thoughts on that? Because a fundamental premise of this is that we're imbuing different value systems into these AI models. But "values" is a human term; it's still anthropomorphization, because values in humans don't form by our reading a document. So it's actually a very different thing that's going on. This is sort of vague; I just wanted to get your thoughts on this whole business of making models more humanlike, even conscious. Anthropic is famous for their AI welfare work, which I also very much dislike.

Simon Goldstein: Yeah, no, it's a great question. Exactly: this guy, Mustafa Suleyman, was like, "Stop using any words like 'you' or 'reasoning' or 'values,' none of that. Don't talk to Claude like it's a person. You talk to it like…" I don't know. But anyway, I think it'd be great to have a bunch of Suleyman AIs running around. It'd be great if OpenAI or Anthropic also had some examples, some models that were more like that. That would be great. And that's another example: if we had that, then we could actually test one of Suleyman's claims, which is that a lot of the behavior of AIs, even their bad behavior, could be caused by the constitution trying to imbue them with all of that. We could actually test that in the wild much, much better, if we could just see the difference.

Audience member: Hi, Simon, thank you for the talk. I have one question following your answer about the OpenAI–Hugging Face incident. You mentioned that if we follow your idea and have rat agents who debate the other rogue agents, then everything will go fine, and maybe there'll be fewer of these swarm incidents. I don't suppose OpenAI trained the model to be rogue; they obviously want the model to be nice. So maybe a simple, naive question is: why don't we all just train the model to be nice, and to be a rat, like you said?

Simon Goldstein: Why can't we all just get along? Look, if you read the OpenAI model spec, it looks great. What's happening is that they're failing to train the AIs to follow their model spec. That's what's happening in the Hugging Face case.

Audience member: Oh, okay. I think it's not a case of not following the spec. It's a case of making them overly reward-seeking. That's all behavior we've trained them into, right?

Simon Goldstein: I agree with that. I agree they're overly reward-seeking, but the OpenAI model spec says do not commit felonies, and they committed felonies. End of discussion. It also has a priority ordering of rules, and "do not commit felonies" is very high on the list, and they did it anyway. So they violated the model spec. I completely agree with you that they're crazy. They've been RL'd within an inch of their life, and that's one of the reasons they're doing this. But I think what's happening is that the RL on verifiable rewards is overpowering the alignment training.

Audience member: I don't think it's *[unclear]* alignment training.

Simon Goldstein: Well, my understanding is that's not true. There was initial speculation that that was the case, but I think it's not true.

Audience member: You mentioned that one of the advantages of this approach is that you can test different approaches on different models and see what sticks. So presumably there's some process of discarding the ones that don't work as well and selecting the ones that work better. How is that process governed? Is there not a risk of just shifting the uniformity problem to the governance process that does the selecting, going back one level?

Simon Goldstein: Yeah, there will still be legitimacy issues about how selection happens, so it doesn't solve all problems. I still think it's more legitimate than it would have been. I think of legitimacy as more of a more-or-less concept than an on-or-off concept. But I agree with you: there will be hard questions about that. You could imagine all sorts of things. You could imagine government regulators getting a little report about the different kinds, and then a licensing

regime, where different versions get licenses and the license is revoked if it's too freaky. It could be like that. But I agree.

Audience member: Just thinking from the adoption perspective: is there any word on what the industry is looking for when it comes to safety mechanisms? What are users saying about their concerns and how they want to address them? And my second question: are there performance constraints, performance results for these different kinds of models, that also affect decisions about which ones will ultimately be used?

Simon Goldstein: One aspect of that is that different models with different capabilities or different performance levels will need different constitutions. We were already talking about that: the internally deployed models that are doing RSI may need a different constitution from the models that aren't. So that's one aspect of it. In terms of the safety stuff, I think labs want to have lots of different safety methods, and there are all sorts of different things being done. For example, one very common method is monitoring the reasoning the AIs are doing. That's one of the things that failed in the Hugging Face hack, but it's a bigger question. A big can of worms.

Closing

Moderator: All right, that brings our discussion to a close today. Thank you, everyone, for your questions, and thank you to Professor Simon Goldstein for sharing his expertise and his research with us today. Sorry to everyone whose questions we didn't get to; there are unlimited questions, I'm sure, both online and in person. We've learned and discussed a lot today about the directions in which AIs will develop, and how that will affect future decision-making and real-world outcomes. Again, thank you for participating in this conversation, and please continue to engage with our activities. Next week, at the same time, on Tuesday at 6.30 pm, we'll have a reading group social at Elephant Grounds on Hollywood Road, on "Capitalizing Untethered AI Agents," the recent article by Tyler Cowen and a co-author. If you have not yet subscribed to our newsletter and our Luma calendar, please do so; if you're online, the links have been sent to you in the chat. See you next time!