

# Alignment Workshop

[About](#)[Vienna 2024](#)[NOLA 2023](#)[SF 2023](#)[Attend ou](#)

**[Adam Gleave]:** I'd just like to ask our panelists to start by briefly introducing themselves, giving a little bit of context on how they came to work on Alignment. So David, do you mind starting us off?

**[David Kreuger]:** Sure, yeah. I'm David, I'm a professor at Cambridge since three years ago. Before that, I was doing grad school at MILA in the University of Montreal. I'll be returning there as a faculty member in September. So I've been thinking about x-risk for over 15 years and doing AI for about 11 years now. It's really one of the things that got me into the field. I guess that's enough about me maybe. I think the topic is the current issues in AI safety. '

I want to do some sort of stage setting essentially. I think around 20 years ago is when the modern concerns about AI safety started. Ten years ago is when they hit the machine learning community to some extent with Stuart Russell, for instance, starting to talk about this and Nick Boston's book getting the attention of machine learning researchers who were very hostile to the idea at the time. And now like with Chat GPT is when it's finally stopped being something that stopped being something that you can say is like some crazy idea that only, you know, a few researchers take seriously or no serious researchers take seriously or something like that. So yeah, you know I thought or hoped that when that happened like last year we would see a big response from the rest of the world and in some ways we have but I'm still very underwhelmed by what we've seen I think. And, yeah, I don't know. I guess we'll get into more stuff later, but I'll leave it at that for now.

**[Victoria Krakovna]:** Hi I'm Victoria Krakovna. I'm a research scientist at DeepMind on, or I guess now Google DeepMind, on the AGI Safety and Alignment team. So I've been at DeepMind for seven years now and I originally got into AI safety as a grad student when I was I was a grad student at Harvard, maybe more than 10 years ago now.

And yeah I was reading about it on LessWrong and it seemed like a big deal. And then, yeah, I got involved in co-founding the Future of Life Institute, which is ten years old now. So a lot of time has passed, but yeah, at the time we were organizing some conferences to bring together AI safety people and machine learning people who were very separate communities at that time.

So yeah, there was a lot of work to do just to get them to talk to each other. And now, yeah, I think things have changed a fair bit. And yeah, I think with the advent of LLMs a lot more people in machine learning are taking safety seriously and thinking about it more, but yeah, there's still, I think, a lot more work to do there. And now in the past couple of years it's been good to see a lot of AI governance work going on. And I have personally been working on model evaluations for dangerous capabilities. How we can assess how good a model is at various potentially dangerous capabilities like persuasion, cyber offense, reasoning about itself, and so on.

And then have that be an input to AI governance to how careful we should be when deploying these models or should we deploy these models. Yeah, right now the AI safety landscape looks very different from when I got

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

**[Robert Trager]:** Hi everyone, Robert Trager, AI Governance Initiative at the University of Oxford. So how did I get into it? Briefly back in 2016 was AlphaGo, which affected me a lot. And it affected a friend of mine. At the time I was a professor of political science at UCLA and had a totally different research agenda.

It also affected a friend of mine who was in political science called Alan Defoe, and he immediately reoriented his research to work on these issues, and he tried to convince me to do it, but I wasn't convinced. But the reason I wasn't convinced was not because of importance, I just thought that other people had a better opportunity and more knowledge and could work on these things. And then in 2019, he convinced me to go over to Oxford and talk to everybody that he was talking to. And I discovered two things. One is that the community was extraordinary, and there were lots of wonderful people to talk to, and I had a great time. And the other thing was that I disagreed with many of the things that people were saying to me, particularly about the political world.

And once I realized that, people said things to me like, "We don't have a solution to world peace, so we need a technical solution to the alignment problem." As if one thing followed from the other. And it just made absolutely no sense to me. And there, I just thought, often we need social solutions that go along with technical solutions and sometimes one works and sometimes the other works. And when I found that I disagreed with a lot of folks that was when I decided to get into the field. Yeah, what do we work on now? The institute that I direct has five areas.

Frontier AI Governance, Technical AI Governance - which is a place that we're investing a lot in now and we'll be training DPhil students in computer science and engineering. And that's one of the reasons why I'm really happy to be here and meet all of you. The other things we do are social impacts China and AI and, I'm forgetting one, am I not? Maybe I mentioned five. Anyway, that's what we're doing. Looking forward to discussions. Thanks.

**[Gillian Hadfield]:** Okay. Hi, I'm Gillian Hadfield. Are we telling our origin stories? I got the impression we're telling origins. A lot of people know my origin story here. Which is that I did have an origin before working in AI safety and alignment areas. My PhD is in economics that's joint with a law degree.

I've been studying the law and economics of institutions and organizations, market structure, and government structures for many years, from the point of view of "Why we don't see more innovative approaches to solving the kinds of problems of the modern world and technology?"

I got into AI safety and alignment through reverse nepotism. Which was that my son, Dylan Hadfield-Menell, was thinking about these things as a grad student with Stuart, a number of years ago. And It actually was very much aligned with the types of things I thought about. Our first paper on incomplete contracting and AI

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

The types of things that I'm thinking about—and we'll talk about this more on the panel are: (1) recognizing that alignment is a problem of systems, not of individual agents, and (2) thinking about “How do you build systems that are what I call normatively competent, capable of integrating into normative governance systems?” That's where we're working today.

I have just made a transition, as Adam mentioned, to I've just joined Johns Hopkins. Which means I've taught my last law school class, which is interesting and cool after 35 years. Where they're starting a new school of government and policy, which will have a big focus on tech policy with a joint appointment in the computer science department. I'm very excited about that. So please do reach out to chat about what's going to be happening there. Thanks.

**[Kreuger]:** Can I say a tiny bit more on the origin story? Cause I guess I started off thinking I would talk about general stuff. I missed some of the interesting parts. So just yeah, like 15 years ago when I started thinking about X-Risk, I thought AI was far off, not going to happen in my lifetime.

And I always viewed this as mostly a coordination problem, where we need to solve these collective action problems so that people aren't taking risks with the collective good and existential risk is, just structurally speaking, very much a collective action problem.

So I heard about deep learning in 2012, I think. Right after the AlexNet thing. And at the time I really didn't know what I was going to be doing with my life because I was trying to think about angles of attack on this collective action problem. But Jeff Hinton's Coursera course in 2012 updated me that deep learning was going to give us AI probably in a couple decades and in my lifetime instead of in centuries or something.

And so I saw a great opportunity to get involved in the field. And I was also very curious because I was thinking about existential risk, more in the way that I would say a lot of people in the ethics community now think about it, or maybe people in the broader machine learning community, not as sudden rogue AI takeover, but as the incentives in society are all messed up, so we're going to, put narrow AI systems in charge of everything, and they're going to optimize for things like, GDP or something, and it's going to be very dehumanizing and bad, and maybe eventually we'll all go extinct or something like that.

And, but, I'd heard about this sort of rogue AI stuff as well online from LessWrong, and I found it plausible in terms of the way the arguments went, but nobody in the field seemed to be taking it seriously, and so I thought that they had good reasons for that, or thought that they, I guess I was like, 50-50 about what they would, to be honest, so I was very curious to see what people in the field thought about this, and was hoping that they would tell me, oh yeah here are all the reasons why you don't have to worry about that, and instead it was like

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

And then, 2015 is when I started my grad studies, so that was right after AlexNet and right when we saw things like DQN and the success of deep learning for translation. And at that point I was like, okay, we're off to the races. There's this thing called reinforcement learning, which is very much like the scary agent threat model.

People are combining it with this deep learning stuff that seems to be able to do these perceptual tasks and these hard tasks and the whole research community is ignoring and ignorant of the issues here, and isn't really thinking seriously about what happens when we succeed. And that's when I became, okay, this is my mission now. Let's get this community and the world aware of this problem.

**[Gleave]:** Thank you, David, and all our panelists. So I'm going to ask a couple of opening questions to the panel, but we will also be taking audience questions via swap card. So if you go to the event for this panel, there should be a little button in the bottom right corner for Q&A. So you can put your questions in there and also upvote other people's questions. But to start with, I'd like to ask each of the panelists to share what you view as being the core risks in AI safety. Maybe one of the three risks that I described in an opening talk, or potentially, as I said, I want to have fresh perspectives and debates. If you disagree with that categorization, or think there's a fourth risk we should be caring about. And I'd also love to hear that. Yeah, anyone who'd like to start answering the question.

**[Kreuger]:** I can start. Yeah. So I mentioned that before I heard about deep learning and was like, "Oh yeah, we're going to have AGI in my lifetime."

I was more worried about these slow, gradual [risks]. Just making more decisions like automatically and that scales and how that might lead to dehumanization and concentrations of power, et cetera, et cetera. I'm still really worried about that stuff. I think, actually, recently I'm trending more towards being worried about this kind of issue than loss of control, sudden rogue AI kind of scenario, because I think at a technical level, the updates have been more positive for me recently on alignment.

But I think even if we solve all these technical problems, if we perfectly solved alignment, which I don't think that's going to happen, to be clear. I think that's just very unrealistic in my mind to be perfect. But even if we take that hypothetical, I think that basically cuts our risk of extinction from AI or something equivalent to extinction, maybe in half.

From maybe 70 or 80 percent to like half of that would be my estimation. I'm really worried about everything between the sudden, extreme, fast take off, classic scenarios that people argued about and that nobody talks about anymore, but I think are still quite plausible. Because I think things like recursive self-improvement are certainly something that we could expect future AI systems to be capable of, and that doesn't mean coming up

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)


---

world in order to figure out what modifications to make.

So I'm worried about everything from that classic fast takeoff. It's all about a rogue misaligned AI. To very slow gradual thing where the systems never really go out of control in the sense of like it just suddenly does something completely different from what I expected or wanted, but in which humanity does not control the way AI technology is used and deployed in the same way that we're not doing a good job of controlling the use of things like weapons and greenhouse (emissions), things that destroy the environment. Alright, so these are collective action problems that I think are really important anchor points for me and how I think about this problem. I think I've basically said ... I forgot exactly what the question was, but I'll leave it at that for now, I think.

**[Gleave]:** Gillian, please.

**[Hadfield]:** Yeah, thanks. So I think you said misuse, misalignment, and destabilization, right? Those were the three. And I really would emphasize this third one, and I'm very glad to see it on the list. And I think you could think of that as disruption, destabilization. I think about it as just breaking the way the world works.

So the type of work that I think about this is as an economist, as a legal scholar, a sort of theorist of how human societies evolved and how they continue to evolve and stability is just absolutely central to human societies. We really don't get anything if we don't have basic stability of expectations about how things are going to work.

Disorder is enormously costly and dangerous for people all around the world. And this is my concern about the ways in which AI that has been poorly designed, or we haven't thought about how it's going to roll out. I think about this especially with the huge push towards agentic systems and introducing AI agents in high volume, just dumping them out into our markets, into our social systems, into our political systems. I think to my mind, that's the real concern about what we might precipitate and how we need to think about design. So the word is a good one, destabilization, right? It's disrupting that stability, which we, you just should not underestimate. It's pulling the rug out from everything you really take for granted.

**[Kreuger]:** Brief response, the three categories I think it's sort of decent. I'm glad that the third one is there as well, very much a lot of times people just talk about accident versus misuse, which is very similar to the misalignment versus misuse, and I hate that distinction, because I think it misses most of the real action, which is somewhere between a pure accident, where it's just whoops, nobody could have foreseen that, and I am deliberately destroying the world, and I think most of this is more like, you know, if I really thought about it, I might realize that this system might have some really negative externalities or might even go rogue and turn into one of these, super powerful systems that nobody can get control over.



# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

my business will fail, or I won't be able to get a job, or like my, my national defense won't be able to compete with the systems that other countries are building.

So I, this is often referred to as structural risk as well, and I think it is a little bit broader than just destabilizations, that's why I wanted to respond, because I think, stability is important, I definitely agree, but I think there's also stuff that's wrong with the way the world works right now.


And I think another thing that AI could do that could lead to dystopian outcomes or extinction is just amplifying current problems that we see with the way that the world works. So things like, inequality and the ability to use money to manipulate people or control people or exert coercive power over them. I think the combination of those two effects is already having some bad effects on our society and could be much worse with AI. And that's just one example.

**[Gleave]:** Yeah, Vika?

**[Krakovna]:** Yeah yeah, I think I'm definitely also in the camp of being worried about all of these things. I think advanced AI will be transformative in a lot of ways that are hard to foresee. So probably will be pretty destabilizing and also present various misuse opportunities, some of which we might foresee. And for example these systems already seem to be pretty good at persuasion. At some point we might see systems that are even better than humans at persuasion, and this could be potentially misused in various ways. For example, for political or commercial purposes in a way that could have various far reaching consequences.

And then there might be other ways to misuse the systems that might not be, so easily foreseeable. And yeah of course as you might expect, I'm worried about misalignment risks and pretty uncertain whether we're going to have the kind of, faster takeoff that David was talking about with self improving quickly. Or whether it's going to be some kind of slower improvement process that we somehow lose control of. I think that's also quite plausible.

And especially given the kind of risk dynamics that we see in the field these days, that could make it more difficult to, be really careful with deploying these systems or slow down if we need to do that and yeah, that's why I'm so excited to see all the AI governance work in recent years and hopefully being able to coordinate on norms and standards and evaluations and possibly, undercutting this kind of race dynamics.

So, yeah, as Gillian was saying, agentic systems are going to be a big deal and very impactful. And I think right now, the risks posed by agents are still not taken as seriously as I think maybe they should be in the or  ine learning community at large. I think I still often encounter these views where people think that these (AI) systems are just a tool. They're not agents. They're not going to have goals or they're not conscious,

# Alignment Workshop

[About](#)[Vienna 2024](#)[NOLA 2023](#)[SF 2023](#)[Attend ou](#)

And to be able to evaluate when these systems start to develop, like maybe different aspects of goal directedness or intentionality, power seeking, etc. That may or may not overlap with how it manifests in humans. Yeah, I think there's a lot to do there.

And right now, thankfully, like one thing I'm excited about with the language model path to general intelligence is that so far, it seems like the systems are not that agentic for how capable they are. Which so far, seems to be pointing towards a slower improvement trajectory, like a slower takeoff, which gives us more opportunity for intervention. People call it slow takeoff. It's slow relative to the instantaneous "foom" kind of self improvement thing, but it still would feel very fast by human standards. So yeah, we really need to have our ducks in a row there. But yeah I think overall, I'm worried about all of these things to varying degrees. And I think there's a lot of uncertainty about how things are going to go.

**[Gleave]:** Robert?

**[Trager]:** Yeah. So I think sometimes we think a lot about prioritizing risks and prioritizing things, and it can lead to a sort of Hamlet problem where. We just spend a lot of time doing that as opposed to realizing that we have a lot of things to worry about. And we should just try to move on a lot of fronts at the same time. So definitely in favor of worrying about all of the things that people have discussed.

Maybe I'll just mention two areas that I think about alongside other things. And the first is the sleepwalking problem a little bit, that we kind of sleepwalk into risks because suppose we had a really bad event, I think there'd be tons and tons of people who would say something like, "Oh yeah I knew this was a problem. I've been telling people this was a problem for years."

But I think in some ways that can amount to, we've been saying that we have a hunch about this problem, rather than a real characterization - crossing some particular line that we can really identify would lead to a problem that we're very confident about. And having enough confidence or a clear enough argument that it can really drive political action.

I think that's something that strikes me as just, really important and actually a really high bar in terms of driving political action and I worry that we might not meet that bar and it's very hard to meet that bar, which is different from saying, it could be a problem and things like that.

And maybe, people could say, "Oh it should be enough to say there's a risk of a certain amount, even if it's a small risk of a really bad thing happening." And that's right. But maybe it's not enough to drive political action. So I think, really focusing on characterizing risks is really an important thing.

## Alignment Workshop

[About](#)[Vienna 2024](#)[NOLA 2023](#)[SF 2023](#)[Attend ou](#)

around that and we could do all sorts of things if that were really clear and nobody wants to walk off a cliff. So that'll help us.

And on the other side, if we can actually make systems really totally safe no matter what people might try to do with them, then of course that would be a good thing too. But we're probably in this intermediate area where a lot of safety improvements are shifting the schedule of the safety performance trade off out. And that's going to change how these systems are used in societies. And it depends on the shape of that schedule and how we are thinking about risks at a particular point in time, whether shifting schedules out like that really has a big effect on risks of how they're actually used or not.


So I think, we're probably in this world where there won't be these bright lines, where there will be these trade offs - there'll be an association between the thing that is potentially beneficial and the thing that is risky - and that's a problematic place to be. And it's a place where technical solutions in and of themselves are often fundamentally can't be the full answer.

Unless they have fully removed us from that trade off, then they're the full answer. But as long as we're still in that trade off, then I think, we really need governance. And it worries me that maybe we're not thinking about the sorts of technological approaches that would help with that governance problem, as opposed to kind of shifting the safety performance trade off outward.

**[Gleave]:** Thank you, Robert. Yeah, just to try and keep on schedule, I'll move on to the next question. One thing I want to post to the panel is: When you look at the range of topics being pursued in AI safety, which area would you say is most underrated; that you'd like to see more people working on? And conversely, which area is most overrated, and you think people are more excited by that is, perhaps, justified? It doesn't mean that the overrated area is a bad place for people to work on, just that at the margin you'd like to see fewer people working on it.

**[Kreuger]:** Yeah I'll say something really quickly about safety-performance trade offs first, which is that I think these are fundamental. I don't think they're going away.

And so I think Robert's absolutely right about this. The one reason why I think they're fundamental. I'll say two reasons that I think are really important to have in mind. One is the issue of assurance, or having justified confidence in what your system's going to do.

Safety, you can think about it as like the actual reality of, will this system destroy the world, let's say if I turn it on or  st to simplify things. But there's also this question of "Do I know if this system will destroy the world when I switch it on?" And right now, we do not have any idea how to answer that question with high confidence.



## Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)


---

and understand it, more theory, all this stuff. And so there's just no limit to how much you can invest in improving the assurance, improving your confidence that this system is, in-fact, safe.

And so the question is: what's a reasonable investment there? And I think the answer is quite a bit, in fact. Probably, we would like some fundamental advances in terms of our basic techniques for doing assurance with AI systems before we deploy anything that seems like it might be dangerous. Like even GPT 4, frankly. I think we're at the level already now where we should not be doing what we're doing. We should stop.

The question about research. So traditionally I would have said I think mech interp is overrated because there's so many people doing it and it's so hot, and everything. We had a really similar question at this summer school I was at yesterday. And I was like: "I can't say that anymore, because Jan Leike had a very similar view to me of not being optimistic about it, but I think he's become more optimistic and I haven't kept up that well with recent developments in the field."


But yeah, so I'll have a meta-level answer to this question. I think basically within the field of AI safety, there's a reasonable allocation of talent. Usually. There's this portfolio approach that Vika wrote about that makes a lot of sense to me because we don't really know how things work.

So this is just: invest broadly, try lots of things, encourage people to pursue any plausible idea that could help. With two exceptions, and again, I'm talking about the AI safety community specifically, not the broader investment. One is, I think there's some sort of trend cycles.

And so things like mech interp get hot and then especially a bunch of young people and new people entering into this field and a lot of them latch onto whatever is the hot trendy thing. And then, those trends last like a couple of years and something else and so on and so forth.

So there's that. And then there's also just a huge bias towards technical machine learning research, specifically like deep learning and LLM research these days. And, it's really good that there is a lot of work on that, but there's so many other angles of attack on this problem and so many other kinds of expertise and background and skills and networks that people can bring to bear.

And I think that's the main thing is that, I know I'm talking to technical researchers here, but I just think we should be doing all sorts of other approaches. We need journalists, we need researchers from other fields, people who work in government and understand how the government works, we need engagement with civil society.

W  eed this really broad based response, I think. And that's something that people in the AI safety community I think have neglected for a long time, and I think that's partially because of arguments and ideas

# Alignment Workshop

[About](#)[Vienna 2024](#)[NOLA 2023](#)[SF 2023](#)[Attend ou](#)

like that sort of population.

It's actually really interesting for a while. I think there was this idea that you needed to be some math or STEM super genius to do anything to contribute to AI safety. I always thought this was totally wrong. Yesterday at the event I was at, I was saying this sort of stuff and somebody came up to me afterwards. They were like, "But I have a technical background. I feel like I can't do anything now because you've just convinced me that governance is the only thing that matters." So it turns out that no matter what your background is, you might feel like you can't do anything. And then the problems that really matter are the things that you don't have the background for. Yeah, I think even broader, swathe of things we should be working on here is the main thing.

**[Gleave]:** Thank you, David. Victoria?

**[Krakovna]:** Yeah, I think one thing I would generally like to see more of is more concrete threat models for both misalignment and the other areas that Adam mentioned, and also more concrete examples of misalignment or examples where AI systems are misbehaving in various ways or like things are going wrong, and then, how we might concretely draw conclusions from that.

Because, often, the kind of threat models we've had on the field, they're sometimes a bit more high level. And I think especially for governance work and for setting out things like responsible scaling policies or generally making plans for: "What do we need to see to judge that maybe we are too close to dangerous systems?" Like, we need more concrete sort of red lines or stopping points for if we see this such and such behavior, this such and such example, or this kind of property of the system manifested in this way. Actually developing those kinds of examples, those kinds of red lines is going to be helpful.

So yeah, this also was something that we were working on trying to figure out in my team. But I think more work on this as a whole is going to be helpful. And, yeah I guess in terms of things that are, I wouldn't really say overrated, I'm not sure I can judge that.

But yeah, I guess right now lots of people are working on evaluations in different ways, and maybe, it's possible you could say a similar thing to what you said about mech interp. But it's a popular area that people are getting into by default. I'm not sure whether it's overrated though, because I think having lots of evaluations is good and it's just that it seems like a very parallelizable area of research where if people just build lots of different evaluations and we have diversity of evaluations and that's generally useful.

And then we can always pick the best ones and implement them everywhere. So I think that's something that is probably still good to see more work on. Yeah, overall: concrete examples, concrete threat models... It's something that I would really like to see.

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

**[Trager]:** Yeah, far be it for me to say, everybody should or shouldn't work on a particular thing.

And to be honest I'm a social scientist. I'm not a technical researcher. I'm so glad to be here and learning from you all about what you are working on so I could really model the space of what's being done, and what's not being done better. So impressionistically I'll just mention two things.

One is, I worry a little bit that it seems like people often are, we're sinking resources into first-best solutions, which obviously we'd like the first-best solution, so we should be sinking resources into them. But, I worry a little bit that they may not be possible. And maybe we're not investing enough in second-best solutions.

A kind of an analogy to the social choice space, where, under some conditions - which often seem reasonable descriptions of the social world - we have various impossibility results where we know we can't have a first-best solution. And there is no way of aggregating social preferences that will do all the things that we would want to do in that context.

And so, instead of doing that we need to think about "What's the messy world that we are in? What are the configurations of preferences that we're likely to see? And what would be the best solutions in those cases?" Which is really a kind of very different research program than looking for first-best solutions.

That's one thing I worry about a little bit. And then maybe related, the second thing is related to my earlier comment. I think it's plausible, I'm not saying it's necessarily true, but I think it's plausible that the marginal value of technical work on some governance parameters, as opposed to on some of the things that people in this community have more traditionally worked on, is higher.

Again, I'm not saying that's true, I'm just saying in some worlds that would be true. And, similar to what David was saying, I think there are all these other problems that, it seems, given the sort of threat model that some people have, would be really essential to work on, actually. So one of the things that you hear people worry about all the time, I think most of us worry about the idea of technological racing and competition and not having enough time to make things safe and in a variety of dimensions.

So if that's your primary concern and you don't think there's enough time, what are the things that we should be working on from a technical perspective there? I think my first choice would probably be thinking about verification.

So if we wanted something where countries were actually going to get together and agree not to race in certain ways and to limit themselves in certain ways, then they really need to know that they were all being lied in a similar way and that no other actor could come in and scoop things. So you'd need verification of an agreement

# Alignment Workshop

[About](#)
[Vienna 2024](#)
[NOLA 2023](#)
[SF 2023](#)
[Attend ou](#)

we can only talk about verification in the context of some social agreement that we would come to and then you need to verify that everybody is adhering to that thing. Either formal agreement or some kind of informal understanding or norm. So that strikes me as an area where I don't really see that much work happening. But given the threat models that I hear, it seems like it would be a really critical place for work to happen.

I'll just mention one, one other area that also strikes me in that way. So one ecosystem that we can imagine that, that I think particularly when it comes to civilian governance, internationally could be effective, is similar to finance, where we have compute providers as intermediaries and you know you we all have the experience of not being able to use a credit card and being denied financial resources because we've raised some red flags and it may be that compute providers will be in a similar position denying computing resources when red flags have been raised.


But that then suggests a whole set of technical problems for people to be working on. What can a compute provider learn about what's being done with its systems from the signals that it's getting now, what could it learn if it was working with customers in some way and there was something in a user agreement that suggested that customers had to provide certain information and so on.

And so that's a huge range of technical problems that we started a project on at Oxford with a couple of graduate students. I should say there are a couple of papers in this area but clearly there's so much more to do. So depending on what your threat models are, some of these other less traditional projects seem like they could be very valuable.

**[Gleave]:** Thank you. And yeah, Gillian, you're going to get the final words since we need to wrap up after this.

**[Hadfield]:** Echoing Robert, I wouldn't want to say that anybody who's doing work now shouldn't continue doing that work. The problem is we don't have enough work generally. I do think that the share of attention is not right. So I'd just like to grow other things rather than shrink anything that we're doing.

But I do think that we're fundamentally missing really important lines of research. I think we need to be talking a lot more about how multi-agent systems work. I think we are way over focused - again, keep doing all this work - but as a proportion, we are way over focused on the single system, the single agent, alignment the idea of aligning with preferences.

All of these things are really not taking seriously that economic, legal, normative, social order is a complex,  ervative system. And you have to study it at the level of the system. We can't just say, "Let's evaluate this model and see how it behaves." We have to know: how is the system going to respond when we have multiple

## Alignment Workshop

[About](#)[Vienna 2024](#)[NOLA 2023](#)[SF 2023](#)[Attend ou](#)

And so I mentioned my background is in economics, and I think it's actually a really important framework to bring, because economists do not say "Let's align individuals." They say, "Okay, let's assume they're all exclusively self-interested." Under what kinds of organizational structures does the interaction of those self interested agents lead to an outcome that's better for society?

We have a theory about perfect markets on that, which actually already has a lot of institutional structure in it because you can't have markets without property systems and contracting systems and rules against just taking whatever you want or beating somebody up to get what you want. So there's lots of legal structure already embedded in the idea of a market.

And then of course our markets are not perfect. Our democracies are not perfect political institutions. So we have a lot of work about what kinds of interventions will shift the behavior of the system. And we're not focused on just changing those individual agents. We're focused on changing the behavior of systems. So I think this is the thing I see that's missing in a big way.

The second is, from a technical point of view of what we build, again, it's the idea that we're going to take everything we want and stick it into the system. I actually just want to disagree a little with one of the things that David said earlier, because I, again, I think we should do this. The world is broken in lots of ways. There's lots of things about the world that we don't like. But we can't say that solving the problem of figuring out how AI is going to change the world and the way in which it's going to be safe to have it integrated in transformative ways, is that we're going to solve all that.

We're going to fix the world. We're going to fix politics. We're going to fix economics. We're going to fix inequality and stick that into the model. That's just not the way it's going to play out. What you need to do is build systems that are governable, so that we have the capacity for all of the structures we have, the kind of things that Robert's talking about, in politics and legal structure, that then comes in and says, "Oh, system, you need to behave differently."

But we're not building the handholds, we're not building the technical side that you would hook into, right? How does a model call a lawyer, and get advice about what's allowable and not allowable? I think that's a technical problem. We could be thinking about how to build that into models, how to build that into systems and how to build the AI specific institutions, and structures that would provide that guidance. I'll stop there.

**[Gleave]:** Thank you so much, Gillian. That's all we have time for, so please do join me in thanking our panelists.



