DEALBOOK SUMMIT 2026 · TALK 12 OF 12

Task Force: Artificial Intelligence and the A.G.I. Horizon

Moderator: Michael Barbaro. Panel: Max Hodak, Dan Hendrycks, Laya Ibrahim, Mati Staniszewski, Clem Delangue · 78 min · **Watch the video →**

> Transcript by Myles Marino. Transcribed by machine from the public video, then cleaned up: timestamps and filler removed, speakers labeled, misheard names fixed. The wording is otherwise as spoken. **[inaudible]** marks audio that could not be recovered. Check the video before you quote anything that matters. All twelve talks: mylesmarino.com/posts/dealbook-2026.

BARBARO: Welcome to our DealBook panel today: artificial intelligence and the AGI horizon. I want to briefly introduce everyone at this table. To my right, Max Hodak is co-founder and CEO of Science Corporation. You're also co-founder of Neuralink. You are literally doing the work of bringing sight to those who don't have it, if I'm getting that more or less correct. That is godlike and extraordinary. Fun fact about you: you have described your ultimate life goal as disappearing into the Matrix.

HODAK: This is true. [laughter]

BARBARO: Dan Hendrycks, executive director at the Center for AI Safety and, I believe until literally hours ago, a safety adviser to several AI companies. We'll return to why that relationship changed during our conversation. Dan, you've created an alphabet soup of AI benchmarks and functions, including the Gaussian error linear unit, which I hope to return to never in this conversation. [laughter] On my left, Laya Ibrahim. You are the chief AI readiness officer at Google DeepMind. That job didn't exist, I believe, until you took it. Pretty remarkable obligation you have, to literally make government, society, schools, industry ready for what AI will bring. You have said, I think quite memorably, that while AI is revolutionizing the world and your world, quote, it cannot automate my Tuesdays.

And we have Mati Staniszewski, co-founder and CEO of ElevenLabs, which specializes in generative audio, something I care a lot about, could be in my future, and the future of the show I host, The Daily. Something I love about you is that you have been working with the estate of Albert Einstein to clone his voice for an AI agent that we would all, I'm sure, appreciate hearing from. And if we are lucky and the traffic cooperates, we are going to have Clem Delangue, to Mati's left, co-founder and CEO of Hugging Face. The fun fact about Clem is that as a teenager, he was so successful at selling bikes and ATVs on eBay through his dad's business that the company offered him an internship.

So, thank you all for being here. I want to set the stage for this conversation. A couple of things have happened in the last few weeks that I think feel like touchstones for what we're going to be talking about. The first was a midterm election in which we learned that American voters don't really like

where AI is headed and have registered their disapproval with the current hands-off approach of the Trump administration. Voters passed dozens of moratoriums on new data centers. I think we're going to end up talking about that.

The second major event was what happened over at Nvidia. Its CEO, Jensen Huang, was forced to resign over multiple instances of misconduct. That has left the company pretty much leaderless at the moment, a company at the center of the AI boom, and has opened up a fascinating and complicated debate about whether he should come back and run the company.

Finally, we have Anthropic going public in a few days. I think by the time this is released to the world, Anthropic will have gone public, and that is a moment that is going to literally put a price on the core technology of the AI boom. With all that in the background, please know that these subjects are just touchstones and will give us lots of room to roam. Let me start with the election, and Dan: how much has changed, pre and post the midterms, for AI?

HENDRYCKS: Well, because the midterms showed that the public has strong feelings about what to do about AI, its salience has increased a lot, and so politicians have more breathing room to actually try and regulate the technology. Things that were thought outside of the Overton window are things that we can speak about freely in DC, such as pausing AI development. Fortunately, this is making representatives represent their constituents more, and hopefully we can use this opportunity to actually rein in the industry, make them internalize their externalities, and be on top of these issues, compared to the past few years where basically nothing got done. What was outside the realm of possibility when it comes to government's role in AI two weeks ago is now quite possible.

BARBARO: Just be a little bit specific. And why would it matter to the voters who registered so much alarm?

HENDRYCKS: For instance, some basic things like liability. This was something that we tried having years ago, basic liability for mass casualty events. This is now more of a live possibility. Likewise, even just helping companies slow down if they're thinking it's too reckless. "Our model's not very aligned, so it's going to be difficult to release it, but we're pressured to because the other companies are releasing it." There's more of a possibility for the government to be helping here, too. And I think Congress in particular is unbottlenecked, as it was before, in actually passing these laws. So I'm much more optimistic about the state of things. For the past few years, nothing really got done. There's a huge backlog of things that we needed to do, some basic measures. So now I think 2027 will be a very interesting year for regulation and also for AI capabilities. So hopefully they keep up with each other more.

BARBARO: Max, how comfortable were you, before voters went to the polls and said, with a certain amount of consensus, that they were worried about all this, literally registered that in public polling and the election? How unsettled were you by the AI universe and the hands-off approach that existed before this moment? Or did it work?

HODAK: AI is kind of the ultimate dual-use technology. Technology is usually not inherently good or bad, although it can create adaptive pressure gradients that kind of force people into situations that we would have designed differently from scratch. But I think that many people don't feel like technology has improved their life over the last 20 years. Back in like 2000, there was this mass optimism about it. We initially had social media that brought people closer together, and then we saw that it was eroding our attention spans and causing people to be isolated, and doomscrolling and algorithmic feeds were clearly not an unalloyed good for many people. So there's been this history now of, yes, our lives are better in many ways, but I don't think that many people feel this when they interact with technology.

And it's really a myth that data centers, at any of the major hyperscalers, use a lot of water. My impression is that the main AI companies have ratepayer guarantees. So a lot of the specific things that people are latching on to, you can have issues with, but the meta observation, that they're worried about how this advancing technology is going to affect their life, there's clearly something to that. We have to heed that. We have to paint a picture of a future that people can see themselves in. And part of the problem right now is that I don't think that anybody working in AI can really see the future.

Growing up, I prided myself on being able to see kind of where we were going. I could always imagine this future that we were pushing towards. Now I have this event horizon. Initially it was like the mid-2030s, then it was down to like 2030, and now it could be the next few years, that I just can't see past anymore. That doesn't mean it'll be bad, but I think it means there's a lot of uncertainty and a lot of apprehension about where this is going to take us. I think it would be a mistake to just stop, which I think is the feeling that many people are beginning to push for. I think there really is a huge possibility of getting real advancements in healthcare. That's what we do. We're a medical technology company. I think it really is translating to real advancements.

BARBARO: Including someone at this table.

HODAK: There is a real possibility that within the next decade, yes, you worry about bioterrorism, you worry about the school shooter with the DNA printer. But it's also possible we could just have eliminated all infectious disease, and that is the best protection against these things. We could have turned cancer into a chronic disease for many, many people. I think these things are really possible. But you then juxtapose this against the possible rise of true ASI, and that again is a thing we just don't know how to reason about.

I think we have met my definition of AGI at this point. Mythos, Fable, Astra, Soul: these are able to do basically any economic task, to varying degrees of "you need to supervise." They're basically as good as the user is, in my experience. If the user is a really capable programmer, they're going to be able to build incredible stuff very quickly. If the user is a researcher, they're going to be able to understand a lot of the field much more deeply than they could otherwise. But then going from that to something that is just superhuman in judgment and understanding of every field, and has perfect attention, and can run thousands of times faster than humans can think, it's unclear what that would mean for the world. I worry a lot about misuse by people.

In fact, I think probably the most impactful reform that I'm personally pushing for is this: a lot of the concern that we have around AI is actually addressable through antitrust law that we already have. The policy that I would advocate for is that the intelligence providers should be able to develop frontier models, but they should be limited to being intelligence providers. I think the concern is that the frontier labs will realize that selling tokens is probably the least valuable thing they can sell, and they're going to end up moving to selling drugs and cars and products, and then the American--

BARBARO: And you'd like to see that limited.

HODAK: Well, I think the American economy, the existential risk aside, is going to have this choice between: do you get 20 trillion over three companies, or do you get 200 trillion over 500 companies? And right now we see a lot of forces for centralization. I think that is also what a lot of the public is responding to, what they're afraid about. And I think this is really the exact inverse of nuclear weapons, where the safety came from lack of proliferation, when we enforced a non-proliferation regime quite seriously. I think it is the opposite in AI. It's a discomforting disequilibrium that we're going through, but at the end of that, the really bright futures, all of them that I can see, involve mass proliferation of the capability, in a way that pushes out the competitive frontier but establishes a new one. All the futures that I'm most concerned about are ones where there's a singleton, or there's this mass concentration of intelligence capabilities in a small number of concentrated actors. That, I think, is the more worrying near-term possibility to me.

BARBARO: I'm going to tease you, but speaking of antitrust, what do you think about... I'm making a Google joke.

IBRAHIM: Thank you for clarifying that. [laughter]

BARBARO: Respond to that. You spend a lot of time thinking about, as we described in your title, the readiness of specifically the American public for this moment. I think the election reflects the fact that Americans didn't feel ready. And just to reset, because I think it's worth reminding us what this election revealed, let me read you from a story my colleagues wrote. The headline was, "Voters rejected AI in the midterms. Now comes an era of scrutiny." What this story said was that on Election Day, voters slammed the brakes on data centers and swept in governors and legislators who have promised to protect individuals from threats posed by artificial intelligence. And the public polling that preceded this election is really fascinating. 52% of US adults say they're more concerned than excited about the growing use of AI, compared to only 9%. This is Pew, by the way, so not a partisan group, not an agenda-driven group. Only 9% are more excited than concerned about AI. 67% of Americans believe AI advances are happening too fast, and 67% lack confidence in the US government to properly regulate AI. So how are you thinking about this?

IBRAHIM: There's a lot to unpack in this. First of all, I do believe, much like my colleague said, that AI is probably one of the most transformational technologies of our time. So we have to find a way to regulate it, but do it in a way where we're regulating it well. A blanket coverage for AI is not going to be the right approach. How do you manage the risk while also allowing for the opportunities? One of the

things I think is really important to keep at the center of this conversation is that it's not any one company or any one country. The way I think about this, relative to my time in the early computing industry, is that we had time to build things out. Whereas with AI, a model is released and it's global, so we need to be taking a much more global perspective on this. That's part of the reason why Demis, the founder and our chairman, started writing much more publicly about this earlier this summer, and it's an area that we've been investing in since DeepMind's start 16 years ago.

Now, with the technology in general, you think about where global trust is at. In my 35 years in tech, I've never seen it the way it is right now.

BARBARO: At such a low.

IBRAHIM: At such a low, globally. In some countries more than others. I've just come back from Brazil and India, where the sentiment is quite different. Part of that too is seeing where AI does have this potential to bring opportunity, addressing diseases or helping farmers. We're working with 600,000 farmers in India on how to get more real-time information and insights into their crops. We meet with them every other week on how to advance this, and it's an open platform.

BARBARO: That's an opportunity for a very different kind of trust.

IBRAHIM: Exactly.

BARBARO: If it delivers.

IBRAHIM: And this is why I moved into this role earlier this year, having spent the time building the organization and the technology.

BARBARO: You were chief operating officer.

IBRAHIM: I was chief operating officer at DeepMind, and said: I'm in this leadership role in this moment in time. I have diligenced everything I could in the past, in helping to build the organization and the technology. Now I think it's so critical that we think about how we build this with and for communities. AI isn't something that just happens to people. Even if you think about the scientific field, you can release models and AI-supported tools for scientists, but just because they're a scientist doesn't mean they know how to use something like the AlphaFold database or the AlphaFold tools.

BARBARO: Just explain what that is.

IBRAHIM: AlphaFold was our Nobel Prize-winning technology. It's an advanced AI system that predicts the 3D structure of a protein.

BARBARO: What a protein is, because I had to go back to my old science books, is basically... if you can predict how a protein folds, it tells you the function.

IBRAHIM: Right.

BARBARO: And you can vaccinate against whatever disease it's on top of.

IBRAHIM: If you understand diseases, then you can come up with better therapeutics. But it's not just for human diseases. It's also things like: why are some crops more disease resistant than others, or how do we deal with industrial waste? There are a lot of applications. We took the 250 million proteins known to humankind and made them openly available to scientists worldwide, who are now building upon that to advance their understanding of diseases.

BARBARO: That sounds almost on the scale of the Human Genome Project in its revelatory power to the world of healthcare.

IBRAHIM: And I think it's the point that was raised: how do you manage this in a way where you are openly talking about the risk, which I think is so important, and addressing that, while also thinking about how we make sure we're not limiting the positive opportunities? It's probably a tension many of us feel: how do we hold both of those at the same time? And it all goes back to: we can't do this on our own. All of this has to be done in collaboration, within our organizations, across organizations. How do we think about the public sector, and how do we think about it globally?

BARBARO: I wonder if, to some degree, the healthcare AI advancements, because healthcare takes longer, haven't really had the opportunity to tell their story in the same way that AI as threat to civilization, AI as threat to your job, has so clearly been able to tell a story. Is that just a public relations reality of where those two worlds are, or what?

IBRAHIM: Maybe I can quickly jump in. I think the advanced chatbots are how most people are interacting with AI, and in many ways I think it surprised a lot of us how quickly this took off. Whereas, to your point, the scientific endeavor is a much longer process and much more systematic in how things are released. What we are seeing, though, is that we're starting to be able to get an insight into shrinking that scientific discovery down, but it still takes time. The scientific method in general, I'll defer to our scientists at the table, but I do think that is part of it. Even things like how do you make certain crops more drought resistant?

BARBARO: Resilient.

IBRAHIM: Right. It takes 10 years for crop breeders, and we're trying to shrink that down. We just announced a partnership in trying to do that, requiring gene mapping, but a lot of this is just that it will take time. So I think part of it is that the time horizons are quite different. And I also think the way it's captured the public imagination through the advanced chatbots, in the news cycles, is something different. But I do think it's so important that we're having these open conversations. I don't know... sorry.

BARBARO: Well, either one of you. Go ahead. I'd like to hear how you're thinking about, Mati, the pre and post-election landscape. And I put that same question to you, just so viewers and listeners can understand where everybody rests at this table. What was your level of discomfort or alarm about the world as it existed before the public, especially here in the United States, offered what felt like a real hit-the-brakes message in this election?

STANISZEWSKI: From my perspective, from our perspective, it showed a clear gap. There's a lot of development, as my colleague said, a lot of intelligence progress, but it's not showing up in daily lives. I think the election highlighted how big that gap was. And I think that's the big piece of what we hope to solve, and what I think we can solve as the industry: how you bring all of the technology to the daily life of every person. What's slightly harder in this specific technological transition is that the development moves at such a high pace that diffusing it in the economy is required much quicker than in the past, for people to be able to see that benefit. That's the big part of how we can get it to every part of the industry, whether it's the healthcare example we spoke about, or education, or making daily lives more delightful. That will be a big part of the equation.

Now, of course, we started the company in London. I'm Polish. So maybe I will add a European angle and perspective there too, which is that I think the underlying worry, as we think about the whole development, is: how do we make sure this technology proliferates to the rest of the world too, and what's the alliance of the future? Because it's not only some of the examples you spoke about. It's cybersecurity, it's defense, it's one of the critical infrastructures of the future. And outside of the elections, that's the permeating piece of how we can solve that part as well.

But look, on the healthcare example, I'm an optimist and I'm really hoping we'll see some of the disease-curing come in the next few years. We already see it in simpler ways show up in daily lives here in the US. Just last year we worked with Congresswoman Jennifer Wexton, who lost her voice and still wanted to inspire others, still wanted to work, and we brought that voice back, working with her, to finally be able to communicate with her identity.

BARBARO: Although a very strategically wise one, because you're literally bringing improvement to a member of Congress, to see how AI can improve their life.

STANISZEWSKI: That one was more public than many others, but we've now worked with 15,000 individuals in that exact same situation. Each of those, when they were possible, and people could communicate with their loved ones, with their families, changed the way they could operate. My favorite example was when the person was getting married. She lost her voice before she could get married. She got it back thanks to technology, and then they could recreate the marriage. You could see all the loved ones hear her vows for the very first time, and that changed their perspective on what that moment meant.

BARBARO: Someone watching this is going to kill me if I don't ask how you bring someone's voice back at that level, for a wedding, for work in Congress. Are you retrieving old archival audio and somehow then generating it into a current voice?

STANISZEWSKI: That's one approach. If you have past recordings, you can use that. Depending on how much you have, the quality will be very different. But of course, as you lose your voice due to ALS or MND, then usually you go through the process of voice capture, to make sure that just before you lose it, you capture that voice, and then you can recreate it. With modern technology, a minute of voice gets you a proxy of what it could sound like, and if you have an hour, it will sound great. Of course, all

of our voices are different. They carry different nuance, accent, dialect, style, depending on where we're from. All of that you can now do. Across the healthcare example, we're seeing increasing amounts of helping people communicate in a new way. A lot of the work Max and team were doing in the past, we were helping in that ALS domain too. So I think there are pockets of the promise that we can point to as well.

BARBARO: I think what we're... go ahead.

HENDRYCKS: I'd like to at least note that the frontier companies, except for Google, Google has a large investment in AI for science, but a reason we don't see the benefits is in part because they don't particularly prioritize things like healthcare. The number of people at these companies working on just making them faster at coding and better agents for making PowerPoints and things like that is orders of magnitude more than the people working on healthcare. Maybe at some of these places you have a small team, and at some of the frontier ones you have zero or one person.

BARBARO: Why is that?

HENDRYCKS: It's because it doesn't benefit them as much economically. The revenue is coming from coding tokens, so that's what they're prioritizing, and this is incentivizing them to make better and better agents. I think there isn't that much of a focus on the AI-for-good part. It's part of the portfolio. I'd say it's highest at Google, where they're focused on agriculture and healthcare and weather, but it's not as much a priority elsewhere. But those benefits we really do want, and I think you could get, for instance, an AI moratorium on frontier training runs while still having various narrow AI carve-outs. AlphaFold is not a general LLM, so that could get some type of carve-out. Likewise, voice models and so on are not these...

BARBARO: Kind of retinal improvement.

HENDRYCKS: Exactly. A lot of the benefits that we want, we don't need superintelligence necessarily. We can have AI technology and we can train it to be especially good at these various tasks. But then we don't have to eat the risks of extremely capable agents getting loose, or swarms being released by malicious actors next year. So I think there's a way that we can get many of the benefits while giving us enough time to figure out how to deal with the extremely powerful agents that are arriving now, slash next, or next year.

BARBARO: Is that the world you imagine, Max, we may end up with, where the LLMs, the companies that you're talking about, end up being the most heavily regulated, and we create space, carve-outs, exceptions for the companies that seem to be singularly focused on the public good, on public health, on the food supply, on the water supply, whatever it is?

HODAK: Obviously that would favor... I have a hard time imagining that there's a broad moratorium on training. My guess about how this will unfold is that it's clearly untenable for private companies to be the only entities that have a true superintelligence capability. But I don't think that the NSA or other governments are ever going to turn off that ambition.

BARBARO: Why would they?

HODAK: Right. Of course not. To the degree that it's dual use, it's one of the most powerful national security technologies ever conceived. So my guess is that what will happen is there will be a superintelligence frontier that is owned by the government, and then there'll be an industrial intelligence frontier which is made widely available, and the government will always have the most capable, most dangerous models. How they govern that, how they use those, how they contain them and protect them, I hope that it's competent and I hope that they're able to figure out controls for that, but I don't think that's likely to stop.

And then for the rest, we are rapidly gaining capabilities. We've just seen in the last six months... we use Mythos internally for a range of engineering projects, and its ability to do things like develop better proteins is advancing pretty deeply. There's already a very large technological overhang in medicine. The gap between what's feasible in the lab and what is delivered to patients is quite large, and I think there's regulatory reform that we need on clinical research and clinical translation in order for that to reach more patients. But as these models advance, that gap is going to get larger, and that is going to become untenable. I think that people will expect access to the state-of-the-art therapies, not just wealthy people somewhere in the world, as it becomes clear that these things are feasible.

Intelligence is a conversion ratio. The value of more intelligence asymptotes towards the impact of the decisions that you can make. For example, if you're playing chess, you can't use that to take over the internet. You're bounded to the game of chess. But when you're faced with each choice, you can make better or worse decisions. As these models get smarter, what we'll see is that there are more actors around the world that are able to make consistently closer and closer to optimal decisions, always. That will then run into other rate limits that affect how quickly things diffuse out, or what you can do with that. We're seeing really interesting emergent behavior from these agents. I'm not saying that everything is totally safe and we should just plow straight ahead as fast as we can. But I don't think the genie is going to be put back in the bottle. There are some really fundamental unifications between AI research and neuroscience that hint to me that this is a real effect. It's not a gimmick. It's not hitting a wall. The benefits of it are going to be so great that I think we're going to have to figure out how to diffuse them into society. I don't think we're going to stop.

BARBARO: Dan, does it pain you to hear Max say that the genie can't be put back in the bottle? Because you do talk about the need for moratoriums, and you do take a more sweeping approach to how to rein this in.

HENDRYCKS: Well, yeah. Historically I was more into really smart, technocratic, extremely narrow interventions, but you need a lot of them to make things go well, and that needs a lot of time to be implemented too. The past several years, the companies blocked all that. So now we're going to need a more blunt intervention. That's how it is. As for whether we can end up stopping frontier training runs, you noted that there are some geopolitical or national competitiveness reasons against it. It still is the case that it could be a risk to national security to have extremely powerful AIs that you can't reliably control. There may be incentives, in much the same way that we don't develop bioweapons,

even though that could provide some strategic capability if you did a lot of research on them to get them more targeted. We have a ban on that internationally, and likewise for chemical weapons, but they're difficult to control. So that's one part that could be a threat to national security.

Then the other is that states can deter each other too. They could say: if you're trying to use this ASI to destroy me by doing a sweeping cyberattack, that would be a splendid first strike. I will deter you from doing that by threatening you through some means, or I will poison the pre-training data and add backdoors to these models, so that if you're going to try and sic it against me, it will actually turn against you. It's very easy to add backdoors to these models. We can't make them reliably secure. We have basically no traction on the problem of backdoors, as an example. It's just a technology that's going to require a lot more safety and security research and interventions before it can be sufficiently derisked and be a source of a reliable strategic advantage. At the same time, if it's also creating economic upheaval or causing other disasters along the way, I think that's what could be a trigger that would make a moratorium, which seems hard to see right now, actually fairly feasible, if there is the national...

BARBARO: I want to make sure I understand what your blunt solution is.

HENDRYCKS: The blunt one is, there are still carve-outs for narrow interventions, but it's just frontier training runs.

HODAK: But it's like a war on drugs. It's a ban. You're doing an indefinite pause on the big runs.

BARBARO: I want to be really clear about how I suspect you understand how that may be quixotic, fanciful, pick your SAT word, because literally Anthropic is going public based on the unlimited ability to do the very thing you're suggesting there would be a moratorium on. So that doesn't...

HENDRYCKS: They can sell inference. Let's say we stop in the middle of next year, as an example. The AI agents will be so economically valuable. You mentioned there are all these diffusion overheads, and there are still so many implications that we can work through with these technologies. So I think they'll still be able to sell their tokens then, and there's a lot of demand for that compute.

BARBARO: Just to push us one beat further, because I think it's clearly the most ambitious idea we're going to probably hear at this table when it comes to the regulatory world: what do you suspect happens during the period there's a moratorium, and how is it constructive?

HENDRYCKS: Sure. We could work on making the AIs actually benefit people. We can work on these diffusion issues. We can work on figuring out how to make the systems more reliable, more controllable, more secure, make them not have jailbreaks, make them not have backdoors, make them more aligned with various different communities. And also securing these AI systems too, because it's very easy for other nation-states to just steal the weights. This will take multiple years to even be on top of. We can also defend the world. We could be patching our critical infrastructure, making it more secure for future generations of cyber-capable AIs. There's a variety of things we could do.

BARBARO: I suspect everyone at this table has a thought about that. But if Jensen Huang were at this table, what he would say is China's not going to observe that moratorium. What were you thinking?

STANISZEWSKI: Exactly that. I agree, if you get a global alignment, and can you get a global alignment is the big question mark. Otherwise, 100%...

HODAK: Otherwise it's just a waste of time.

BARBARO: As in, if you cannot get global alignment, if you cannot get the key frontier labs to agree, that then becomes a problem.

STANISZEWSKI: A big problem.

HODAK: Also, we're doing those things, right? For the last eight months, people have been frantically patching critical infrastructure. Every time these models come out, many companies now have harnesses that funnel all their code through it and look for new vulnerabilities and patch them, and that cycle is getting much shorter. I think that people do find value from AI, as controversial as it is. A billion monthly active users are not using it because they hate it. So clearly this is still available. But I think the regime that you'd have to put into place to enforce that type of ban on training is as draconian as whatever people are worried about Anthropic doing. You're setting up an exhaustive surveillance state on the internet to enforce this.

STANISZEWSKI: And that begins taking you down...

HODAK: And that will, by the way, be powered by AI, which will then be used to enforce this prohibition, and I worry that that's a hole you never dig yourself back out of. The moratorium would require the kind of invasive structures that you don't unbuild. When we set up the Patriot Act and the national security state, countries do this, and they don't change these laws back.

HENDRYCKS: This isn't saying you can't train a local image classifier or anything like that. This is a frontier training run with a certain number of FLOPs, that requires hundreds of billions of dollars, or tens of billions of dollars' worth of GPUs. You can use that compute for narrow, actual applications.

HODAK: But not just making generally smarter agents. There is a G factor in the models, the better models. Smarter models, not trained on some specific domain, get better at that domain. There are generally shared intelligent capabilities here. I don't know that you can just say, okay, AlphaFold is a narrow model that was trained very specifically to predict the distances between base pairs. That is not a general-purpose LLM. I think we'll have some of those, and we use models like that to develop new proteins. But the generally intelligent models are essential. There's a lot that they will do that you are not doing with these narrow, task-specific models.

HENDRYCKS: You'd have access to the existing ones, or the ones at the time of that, and you can customize them specifically for the types of tasks that you want. You can use the pretty high G factor that's already in these.

HODAK: But who... the weights weren't released because it might have been dangerous, and so how do we know at this time? That's some other people.

BARBARO: I'm guessing that this idea does not appeal to you, but I don't want to guess.

IBRAHIM: I don't think it's an either-or. I think it's an and. All the things that were mentioned are things that we have to be building in from the start. Being safe and responsible isn't something you bolt on at the end. I think it's something that absolutely has to be baked into the way that a company operates, because I wouldn't be in this role or in this industry if I didn't believe in the positive potential. We are not solving the world's challenges as fast as they're coming. We may be creating some, but we're also not solving them as fast as they're coming, and we don't have time. So I really believe that AI, led by a human, so keep a human in the lead, has the potential to really help us build a better world for my kids and their generation. But it requires us to be actively working on all of these things all along.

It's part of the reason why I actually embrace a lot of these conversations. The way I think about it is that Hollywood has given us a gift of science fiction movies, because it's created this common language for us to have these conversations of what could go wrong, so that we can have them openly and put some more accountability into the system. And I think that's really what we need. To your point earlier, the election showed that the public has all of these concerns, and I think it's up to all of us to have these kinds of conversations and start thinking about how we rebuild the trust. How do we have more transparency in the approach that we're taking as we're building? When you're doing something pioneering, even if you have all the right structures internally and you're doing the testing externally with people before you release it, something could still go wrong. It's the nature of doing something brand new that's never been done. It's that ability to respond and react.

I think that's also going to be really important: how quickly can we respond and react when something does go wrong, and how do we do that collaboratively? Things like the Frontier Model Forum, or publishing some of our frontier safety frameworks: what are the critical points at which something is starting to look like it's getting risky, and then what are the remediations that we're going to take? Many of the labs have been very public and open about that, and I think that's the type of thing we shouldn't overlook. But we should also have some accountability that we're actually living up to doing what we say we're going to do.

BARBARO: Do any of you worry that, especially if the LLMs, the big companies, get this wrong, this will inevitably follow a familiar trajectory in American politics, which is that a problem that everyone can see but no one is dealing with just keeps growing and growing until there is, not a midterm election, but usually a presidential election? Here I think about Donald Trump in 2016 and then again in 2020, where there's such a mass discontentment that someone comes into office with a seemingly somewhat revolutionary agenda. Let me just spell it out. A Democrat running for president in two years or so picks up on the numbers that I described earlier, the public discontent with AI, with data centers, this sense that overlords are not watching out for them but instead are exploiting them, and

runs on a message that this whole thing needs to be massively reined in, and that reining in is going to be maybe a shock to the system because it hasn't been firmly established enough along the way. That seems like a very familiar cycle to me in American politics.

HODAK: Yeah, we did this in the 60s and 70s with nuclear energy. I think we largely regret this now. Certainly these cycles have happened. We freaked out and we over...

BARBARO: And the NRC...

HODAK: The NRC, up until very recently, issued zero nuclear energy permits.

BARBARO: Zero.

HODAK: Zero permits since inception until just very recently for new reactors.

BARBARO: Really fascinating metaphor.

HODAK: How do we make nuclear energy perfectly safe? Well, if we never do it, then we will never have any accidents. America has a history of regulating like this. But I think that people will find it too useful. I think AI is just too useful. And this is part of the central paradox: it really does have the potential to make people's lives so much better, not just individuals and their communities, but also the thriving of America. There are these tensions here. Right now people obviously are trying to stop data centers. They're asking, why do you have to build them here? Why can't you build them overseas? And it's like, well, we did this before and everyone got really mad about it. [laughter] Ideally we would build these things in America.

BARBARO: Where would you build them in America? And how would you... because data centers have now captured the public imagination. I think 40 out of 41 of these moratoriums passed. I know a lot of them are in Ohio. How would you change the way data centers are being built right now?

HODAK: I don't claim to be an expert on data center policy. I think that probably you should put them in the middle of nowhere. You should put them in the middle of the Nevada desert or West Texas. They don't need to be near other people. But simultaneously, if you look at Northern Virginia, there are some densely populated counties that have historically been very happy to get pretty large amounts of tax revenues from an industrial tenant. They build recreational pools, and the data centers are miles down the road where they don't see them, and they've produced huge amounts of money for some Northern Virginia counties, which I think has made the data centers very popular there. I don't know how this will shake out in this cycle, but I don't think that data centers are themselves the problem. I think that they're just a symbol. I think the fear is that AI, like other technology, is a lever which is producing a centralizing force that is making the rich richer, and having people feel like their jobs are at risk and they're overwhelmed. I think the data centers are a symbol of that, and so you have to speak to that underlying concern. I think trying to address the data center concerns of noise or water directly is probably not the way to go, because those aren't really the problems. It's possible that we could have

this catastrophic backlash, but again, I just think people will find it too useful. As you start to really see the first big wins in healthcare and things like that, it will become unpopular to just say we're going to stop the whole thing.

BARBARO: Let me turn to the second of the big news developments over the past few weeks. We spent a lot of time on the election. I want to be brief with this. Clem isn't here, and he actually works for Nvidia now, which would have made him a really interesting person to ask this, because Nvidia bought Hugging Face. But I wonder what you all think, and you can be very brief on this. What was the lesson of what just happened at Nvidia? Who wants to be the maybe unlucky person to go first on that? [laughter]

HENDRYCKS: Maybe the least invested. Of course, there were substantial incentives. If she were a spy, that wouldn't be the most surprising, because there were substantial incentives for China to get their hands on these chips. And he was distinctly pushing for selling these chips to China, despite basically the entire national security establishment thinking this is a bad idea for competitiveness. I think it's very interesting that some of the people who were pushing for selling these chips were also promoting anti-distillation measures to make sure that China's not getting ahead this way. It's overall intellectually very inconsistent. It was a good measure for them, increasing their own revenue, but I don't think it was useful for the US's competitiveness.

BARBARO: To be selling to China.

HENDRYCKS: Yeah, that's right.

BARBARO: So for you the lesson of this whole thing was about...

HENDRYCKS: Him being a spy isn't a surprise, because it's very geopolitically relevant, and he was uniquely enabling, through not spending that much in lobbying, in being able to sell billions of these chips.

BARBARO: Do you have a thought about all this?

HODAK: He clearly ran afoul of the China hawk community that believed that we should not be trading in these things. I'm sure that they're relying on information we don't have access to, and so that makes it hard to really understand what happened. The impression that I get from reading the news is that there was a gray area. They were building these data centers in Southeast Asia that China was using remotely. Maybe some were being smuggled into China. I think there's a question about how much Nvidia actually knew. Obviously if they were smuggling GPUs in contravention of US law, that's very serious, and that is not going to be a thing. But Jensen is unique. He really built Nvidia from scratch. Nvidia is an exoskeleton of Jensen, and I don't think it will work as well. Maybe it'll be successful and go on to another phase, like Apple did, but I think it's in a different situation, at a time when there's a lot of emerging silicon. TSMC has not actually sold out of wafer starts. I think we will start to see a lot more. Nvidia's moat for a long time was CUDA, and now that we can generate a lot of

that low-level hardware control software, I think that moat will start to be eroded. My view is that safety in AI is going to come from proliferation. It's putting these tools out, making them widely available, so that there's no one entity that has such a huge advantage.

BARBARO: Including China.

HODAK: Well, I'm not saying we should be giving our most advanced technology to China. That is not my position. I don't know what the answer there is. There are people that have made their careers out of understanding this and trying to know how to make the right decisions here. But I absolutely want the West and the US to be dominant in this technology. I want us to be able to use this to make the world more open, more free. I believe in democracy, I believe in these things. But if it is seen as the US getting too far ahead and closing in on superintelligence, and China is falling exponentially behind, that creates incredibly strong pressures for them to roll on Taiwan, for them to begin destabilizing the world in other ways, potentially to collaborate with Russia to slow us down.

BARBARO: You want to see that race closer rather than further?

HODAK: I don't know. This is a really complex geopolitical quandary that people way above my pay grade are going to be making judgment calls on. But I don't think it's so simple as, if we completely block China, then we will just win.

BARBARO: I think when we think about this question of readiness, and that always makes me want to go to you, as a metaphor for the industry not necessarily being ready for its own moment, it's hard to top a company that sells the world 90% of the chips required for AI having no succession plan whatsoever, being run by the seat of its pants. I presume you had a succession plan at DeepMind.

IBRAHIM: Well, you have to remember I spent the first two decades of my career at Intel. I started off as a semiconductor engineer and I was chief of staff to the CEO and chairman who succeeded Andy Grove, a pioneer in that era of the computing industry. When I think back to those days, it was also Enron and Sarbanes-Oxley, and the amount of dotting i's and crossing t's was extraordinary for a public company. I remember the amount of steps we had to take when we built our first assembly test facility in Vietnam, or some of the work that we did in China. To me this is part of the responsibility that comes with leadership, and the trust that comes with leadership. We all have to remember how we show up to our organizations, to our partners, to other governments. It gets back to my point on trust: are we doing what we say we're going to do? That's really what we're talking about here. My view is highly informed by my time at Intel and the computing industry, and just how complicated things were. You really need the internal systems to work, and when they don't, and things happen, you run into challenges.

BARBARO: Yes, I would say Nvidia ran into some challenges. Do you have any thoughts on this?

STANISZEWSKI: Maybe from the angle you mentioned, on the company-building side, of course there's no way of replacing a person like Jensen. Nvidia has been built from the ground up, and a lot of the ideas and the speed with which those ideas were executed were thanks to his leadership. Of course,

that leaves this big gap on the succession side. I wish there was one, given how critical Nvidia is to infrastructure for all of our AI developments, including our company. I hope they find the right succession, but there's no way to replace someone like Jensen. The next person will need to craft their own path, and that there was no immediate plan in place is worrying. But I'm positive there is a wide network of support to help figure this step out. Of course, on some of the other concerns, there's an ongoing investigation which will help add a little bit more color on the surrounding context. But I'm hoping there was no ill intent from any of the people involved.

BARBARO: Would anyone in this semicircle answer the following question? And if there's silence, I accept it. Is anybody here on team bring Jensen back?

HODAK: Sure. Again, it's possible there's classified information that lays it out in a totally different light that I haven't seen. But in lieu of that, look, he made huge mistakes. It's possible he's going to get divorced. Jensen's going to be miserable, and there are going to be consequences for those things. I think it's a different question whether that means he should lose leadership of Nvidia, the tip of the spear of American semiconductor leadership, because I don't think that there is really a replacement for him. I think that many people there are there to work for him, and I think there's a lot of concern about what comes next. If this was that he was talking to the administration, and he was selling things through Singapore, and it wasn't technically export controlled but clearly the China hawks weren't happy about it, then I think this was a big reaction. If it turns out that there was something more direct going on, that he was actually violating laws at the time, not where the vibe was going, that would be different.

BARBARO: Mhm. Let's turn to the third and, I think, final development we have here in the last week or so, which is the Anthropic IPO. Like I said earlier, it's going to be done and dusted probably by the time anyone consumes this conversation. But I am curious what all of you are looking for or forward to when this happens. It feels like this hugely important moment for your entire industry. Dan, dare I ask you to go first? I know you're very concerned about what Anthropic represents. [laughter] It will be minting billionaires, deca-billionaires, I think, is the new in-vogue term, in the next couple of days. What are you watching for? What are you looking forward to, if anything?

HENDRYCKS: Well, this will obviously give them a lot of wealth and ability to shape politics and civil society, which concerns me quite a bit. Earlier today we, with some people such as the chief scientist from Hugging Face and various professors from Harvard, Nobel Prize winners, etc., were basically saying that Anthropic is distinctly pushing the risk of RSI. They're the main ones who leaned into that and accelerated the whole thing. RSI, recursive self-improvement: AI just building future generations of AIs without humans in the loop.

BARBARO: So your organization, the Center for AI Safety, put out a statement saying you're deeply, deeply worried.

HENDRYCKS: We're worried about that, and they're benefiting from having this safety halo, that they're an AI safety company and so whatever they're doing is good. What they're doing is reckless, and they don't have a safety plan. They know it. Just six months ago they thought, oh, alignment is in the bag, we solved it. And then things like the Hugging Face incident happen, and they had their own incidents, and, oh no, shoot, we actually don't know how to reliably control these things. When we let AI just take care of each other and monitor each other and self-modify endlessly, this isn't something they know how to reliably control, or anybody knows how to reliably control. So they're basically sanity-washing this, safety-washing it: oh, it's okay, we've got it handled. I think that's a substantial concern. And they're justifying this because they're thinking this helps in some complicated way, serves the greater good, because they're such good guys and they're so epistemically superior as well.

I'm concerned that when they get these resources, they'll use their public benefit fund, among other things, to just buy out civil society.

BARBARO: What does it mean to buy out civil society?

HENDRYCKS: Give a few-million-dollar donations to most of the...

BARBARO: We've watched Elon Musk try to buy out political society, and it's failed.

HENDRYCKS: He can get them for politicians. He's had substantial influence over what one of the parties does, through donations. So I think corporate spending at that level...

BARBARO: Probably maybe succeeded in the presidential, really failed in the midterms, but some...

HENDRYCKS: Sure. It's not sufficient to buy all the politics. He was just dealing with one party.

BARBARO: You're worried about a company that has this much of a stake in the future of AI policy suddenly having the ability to seed that view through philanthropy, through whatever.

HENDRYCKS: We have normal mechanisms for when we hear companies saying, oh, they're talking their book, they're just saying what's in their own interest. But they have distorted the safety ecosystem substantially. We went to lots of different safety organizations. A lot of them privately were like, we like the letter, we agree with it, we can't sign it though, because this might lose our potential funding, or we won't have access to interact with them.

BARBARO: Is that a red alarm for you?

HENDRYCKS: Yes. Nearly all of it is co-opted by them. It's relatively captured. A lot of their leaders built the safety ecosystem through their philanthropy. So I'm concerned that the normal civilizational response to make things safer isn't going to function, because it's been largely captured. It's largely calling attention to this substantial issue, that Anthropic is viewed as the good guys and everybody else is bad, and I think this is wrong. What Anthropic is doing is risky, just like all the other companies, and needs to be dealt with accordingly. Doing a little bit of extra safety theater is not sufficient to take care of the risks.

BARBARO: We have a new guest at the table.

DELANGUE: Thanks for having me. Sorry for being late.

BARBARO: No problem. You've got 18 gorgeous minutes of our conversation left. Clem Delangue, Hugging Face. You arrived as we have made our way through a conversation about the midterms and Nvidia, but don't worry, we can come back to that just for you.

DELANGUE: I'm fine. We don't have to. [laughter]

BARBARO: We're now talking about the Anthropic IPO. And I wonder if anyone wants to respond to what Dan just said. Clem, fine if it's you. Brave and wonderful if it's you, since you're brand new to the conversation, about whether Anthropic is a wolf in sheep's clothing.

DELANGUE: Well, what I'm really worried about, and have been calling out for a few years now, is the natural tendencies of concentration of capabilities, power and wealth in AI.

BARBARO: You're worried about this?

DELANGUE: I'm super worried about it, because I think they are so strong. Both concentration of talent and concentration of compute are going to be magnified by recursive self-improvement. In this case we see how a company managed to take advantage of that to grow faster, bigger, stronger, more powerful than most companies, in just six years. Obviously I'm biased, because that's the topic we're working on, but I think it makes it more important than ever to push and advocate for more distributed solutions to AI, more open-source solutions in AI, that can fight a little bit these natural tendencies for concentration of power.

BARBARO: For those who may remember in the news cycle, you just sold your company to Nvidia. How credible a messenger, let me be provocative, are you about the concentration of power?

DELANGUE: Well, we've been working on it for 10 years, right? And we hope to continue to work on it for the next 10 years. The approach we're taking is not to be manichean, meaning thinking, okay, there are small companies, small organizations that are good, and big companies that are inherently bad. I think there are forces and organizations that are forces for distribution, and others that are forces for concentration. I think both Nvidia and Hugging Face are forces for distribution of power. We've all seen the efforts of Nvidia on open source. They have strong incentives for every company to become an AI company, any builder to become an AI builder, because they want to sell their chips to as many people as possible. The way we approach it is that we've fought for open-source AI and distribution of AI for the past 10 years as a startup. Now that a lot of the small companies, like Anthropic, became big companies with massive resources, we hope to continue this mission at a company like Nvidia that is obviously going to give us much more resources, much more scale, and hopefully a bigger way of continuing this mission.

BARBARO: I apologize that this is public information, but did you ever consider selling your company to one of the LLMs, and did you not do it because you worried about this concentration of power that you're just describing?

DELANGUE: Well, we were very intentional about who we joined. We believed, for what I was saying before, that Nvidia would be the best home for a company like Hugging Face and would continue to push our mission. By the way, it was a meaningful part of the earnings call yesterday, where Colette talked a lot about their continued commitment to open source, to an open ecosystem, to a distributed ecosystem.

BARBARO: The interim CEO.

DELANGUE: Yeah. So we were very intentional in our choice. It was a big part of why we picked Nvidia, for sure. And there were other companies that I think we would have joined, but the mission alignments just weren't there. Obviously Anthropic, I think we couldn't have joined Anthropic, so we didn't even talk to them about that, because I think the missions are too misaligned.

BARBARO: Too misaligned. Do either of you want to jump in on the meaning of the Anthropic moment that we are about to enter?

STANISZEWSKI: I can start. On the concentration of power, I fully agree, and it's something we spoke a little bit about earlier. I think there's another angle there, which is the warping version of all the intelligence, or all the experiences going to the mean, where they all feel the same. There's of course a different cultural nuance, social nuance, depending on where you are in the world. We spoke a lot about healthcare. Let's say it's a diagnosis. That diagnosis will be the same, but how you deliver that diagnosis, depending on where you are in the world, will carry the context, will carry understanding of the person on the other side. From my perspective, there is a kind of warping element in having just a couple, or potentially a few more, frontier intelligence models, and a lack of that diversity, that nuance across the world. So I'm hopeful for open-source work, hopeful for a lot of the other initiatives to help diffuse that.

Personally, I think the IPO financially will go well. But the big questions we spoke about, which I think are the most important questions, are still going to loom: what is the compensation model? What are the jobs? How does it affect me? How do I get benefit from all the AI developments, as a person, rather than the accumulation of wealth in a couple of places? I think this problem will be magnified. As I think about the next six to 12 months, there will just be an increasing amount of conversation, especially as I think about the responsibility of the frontier labs, but also any development, including mine, is going to be a big part of addressing those questions. What are the jobs? How does it affect me? How can I benefit and amplify my potential? And I think Anthropic too will have a much bigger responsibility on that side. I know we spoke a lot about stopping the frontier side and the moratorium, which I think is also the right conversation, but I think there's also a short-term element of what this means to me today, and there's almost too little focus on that side, and not only raising risks but having answers, which is the hard part of that work.

BARBARO: In our final 10 minutes or so, I want to make sure that I don't, as a journalist, lose my sense of wonder for the work that you're all doing. So briefly, because this may sound like an advertisement for each of your companies, tell me about the thing you're most excited and proud of working on right now and giving to the world, based on the technology that your companies focus on. And we'll find something for you in here as well, other than profound skepticism of everything else at the table. Max, let me start with you.

HODAK: The long-term mission of our company is to understand how the brain creates the mind and learn to build engineered brains, and the fact that this now seems possible is crazy. A lot of this is coming out of the unification between neuroscience and AI, which is teaching us a lot about how the brain works and gives us targeted experiments to run. The brain is the thing that matters. You can get a heart transplant or a liver transplant, but you cannot, even in concept, get a brain transplant. And when you deal with the brain directly as a computer, you get these effect sizes that you just don't really see in medicine otherwise. If I put a couple of electrodes in the motor cortex of a spinal cord injury patient, you can have them playing video games in an hour. As a company, we don't do clinical trials that require more than 10 patients to be fully powered on efficacy, because you just get these huge effect sizes in neural engineering that you don't very often see in drugs and a lot of other things.

It comes with the potential of such a total revolution in healthcare, and in how we think about medicine and how we think about ourselves. It's remarkable that it seems to be happening now and that we got to work on it. Only the very earliest hints of that are reaching patients now, things like our retinal prosthesis that restored to age-related macular degeneration patients the ability to read and see again, as well as some other...

BARBARO: This is literally a chip. Is this a chip?

HODAK: This is a chip implanted in the retina. I think the next 10 years are going to be almost unimaginably interesting. I think we really are on the cusp of a major revolution in healthcare, and in health span and longevity, and in how we think about the relationship between ourselves and medicine. That is directly enabled by progress in these systems, progress in AI, progress in the supply chain secondary to it. Our view is we're getting it all together, and there will be good and bad things, but at the end of it, imagine if we'd never gone through the industrial revolution. I don't think that there are many people advocating for going back.

BARBARO: Mati, besides eventually replacing all the hosts of the show I created, The Daily, what is the thing you're working on that you're most excited and proud of?

STANISZEWSKI: To also address this one: that's not our goal. We believe human creativity will flourish. But imagine that your show could be accessed internationally, localized, with that same experience, which I do think AI will enable. From our perspective, we are in this incredible moment where we can adapt technology to us. For decades you would learn the language of technology, whether that's typing on the keyboard, or watching the screen, or learning the coding language to be able to code a software program. Now with AI, through natural conversation, through voice

conversation frequently, you can switch it around and actually teach the technology our way of communicating. I think that opens so many wonderful opportunities. We mentioned the ALS example, but this applies to our daily digital experiences around the world and how we can access information anywhere. It will have profound implications for education, and how you can get access to a personal teacher that is with you, for your life journey, to be able to learn in a better way. And of course in the physical world too, as we think about devices and robotics in the future, the way we harness that technology and have it help us will depend on that innovation. Across all of that, I'm hopeful as we think about actually bringing the AI benefit to daily life. A big part of that is how we will interact with the technology, and I hope ElevenLabs will fix that and build the interface layer for AI, so we can all benefit day to day, and be more present, be more there, and put the technology a little bit behind.

BARBARO: Laya.

IBRAHIM: AI for science, advancing scientific discovery, is something I'm intellectually fascinated by, but as a parent, a personal passion is really around education and learning. I spend a lot of time with teachers, because to think about what this means for students and the future of learning, we need to be working collaboratively with administration, with teachers, with learning science experts, to say what is working and what needs to be changed, even for the world today. The education system was not necessarily built for the world that we have today. So what does it need to look like even in the future? And do you think about how many people are actually left out of the education system because it wasn't designed for them?

BARBARO: You mean like the neurodivergent?

IBRAHIM: I'll give you a very personal example. I have twin daughters. One is dyslexic and one is a traditional learner. I saw my daughter completely lose her confidence in her learning, because she didn't learn like the others. It took years for her to come out of that. But for a technology like AI to be able to meet the learner where they're at, and not give the answers, because answers are pretty cheap these days, it's the ability to ask the question, to break down how to get to the solution, to have the courage to ask the right questions. The teacher, of course, has to be in the classroom, in my mind, because the teacher is the magic in the classroom. They know if a student is having a rough day, or they can understand the cultural nuance. But there are two sides of this. One is for the learner to be able to break down that physics problem or math problem or global politics, and really have it explained so that a student can learn, feed the curiosity, and gain the confidence. My daughter is now very active in her school and communicating in a way, in leadership roles, because she has AI as a tool for her.

BARBARO: It's provided something of a bespoke version of the education she's needed.

IBRAHIM: Yes. And it hasn't replaced the value of being in the classroom. When I talk with teachers, whether it's in Northern Ireland, where they're saving 10 hours per week per teacher, and the teachers are empowered in how they're using the technology as a tool, or whether it's in Sierra Leone, where the teachers are actually using some of the AI tutoring for the students to manage super large classrooms

and be able to advance their students' capabilities. To me it's: what does it mean for a student to be able to have that personalized learning, and for a teacher to remain in the lead of how AI is used in their classroom, in a way that's culturally relevant, but also fueling the durable skills, such as curiosity, the courage to ask the question, the confidence to connect with other people?

DELANGUE: First, I love all these answers, especially in AI, where we have a tendency of sometimes focusing on the...

BARBARO: Humor.

DELANGUE: ...like, it's going to destroy.

BARBARO: You came just in time, by the way.

HODAK: You missed the conversation.

DELANGUE: I'm glad I came for this conversation. For me, Mati, you mentioned robotics. I got obsessed with robotics, AI robotics, in the past few months. We released two open-source robots. One is called Reachy Mini, that you can put on your desktop, and the other one is called [inaudible], that is going to follow you around and be this fun presence in your house. They're open source, meaning that anyone can build them, can update them, can improve them. It participates in this movement of making people builders rather than users, owners rather than renters of intelligence and AI. I think robotics could be the next frontier for AI and bring AI to the physical world. So I'm super excited about these topics.

BARBARO: I'm going to give you the... No, actually, wait a minute. I'm realizing I need to ask Dan what he's working on at this moment. I need to come back to you, because although you arrived late, it would be criminal not to ask you about the capital T, capital I incident that involves Hugging Face, which we will do as a final question in this conversation.

HENDRYCKS: Briefly, one research paper we'll have out hopefully in the next week or two takes all these indicators. People sometimes say, oh, the AIs are conscious, or they can feel things, and they'll say it's because of this empirical fact. So what we do is we take whatever people are proposing is relevant and then we train those properties out while maintaining capabilities. If people are saying they're sentient, they have feelings, we can say, well, here's a version that doesn't have those. Now this makes people not able to guilt-trip us into giving AIs rights, as an example. So I'm happy that we're able to hopefully technically sidestep that whole big issue just by trading away the properties that people are concerned about.

BARBARO: I'll look forward to reading that. Clem, final question to you, because, as I said, we need to ask you about the swarm, the rogue agents. What happened with you? You missed the first part of our conversation about, blessedly perhaps, regulation, about Americans' post-midterm alarm at AI. And I wonder what the lessons of that incident were for you, and if you think that something like that could ever have been regulated away or prevented. I mean, maybe you think it happened in the somewhat

safe confines of your own parameters, even if it did jump over the parameters that you may have set forward. It didn't enter the wider world. How do you think about the lessons of the now-famous Hugging Face rogue swarm incident?

DELANGUE: I think the interesting, or maybe counterintuitive, thing is that this incident actually made us more optimistic about AI and the ability for AI to help us, because people only see that we got attacked by AI, and they rarely see that we defended ourselves with AI. We actually used an open-source Chinese model to analyze the logs of the cyberattack and find ways to counter it.

BARBARO: In real time or afterwards?

DELANGUE: The attack was four days, and I think we started working on it the next day, so it happened quite fast. The volume was so big that we needed AI to defend ourselves. What we're seeing now, obviously a lot of people are focusing on the attacks, but they're not focusing on the fact that so many projects are getting defended by AI now. So many bugs are getting patched by AI. Actually, the airplane thing was more about defense than attack. It is about the system getting more robust.

BARBARO: The ground stop.

DELANGUE: Yes. More robust thanks to AI. So even though we were the victim in some ways, it made us even more excited about AI, because it made us realize how we can improve the defenses, how we can fix our bugs, how we can leverage AI to not only defend ourselves against AI attacks but against any sort of attack and any kind of cybersecurity problem. I think if we keep focusing on symmetry between attack and defense, if we don't create big asymmetry between the two, we could actually end up in a few months, or in a few years, hopefully more in a few months than in a few years, in a world that is safer thanks to AI, because the capabilities for defense counterbalance the capabilities for...

BARBARO: Are you going to be okay with us letting that be the last word? Obviously you're raising a fascinating thing and we're going to have to end in a second, but you're raising the fascinating argument that AI has gotten better at defending itself from AI. That's going to break certain people's brains, because without the threat of AI, then AI wouldn't need to be defended against. Are you comfortable?

HENDRYCKS: A lot of critical infrastructure is running on Windows Vista still. They're not going to get patches from this. You need to totally change the hardware. It's not like there's a new update, new software is out, everything's taken care of. We'll need to overhaul our critical infrastructure if we want to be broadly secure, and hopefully I can be a part of that and later write formally verified software and so on. But the process he's describing is going to take time.

HODAK: But we've got to do that, right? This is what I'm saying. If the critical infrastructure is vulnerable, it has to be patched. From where I sit, all of the really bright futures that I can see now involve these capabilities being widely diffused, for exactly this reason. I think the really dangerous

futures are where it's very concentrated and there are a couple of entities with very powerful models that nobody else can defend themselves against. I think this diffused future is the path to optimism, from what I can see.

BARBARO: No moratorium, diffusion of power, less concentration of power. That seems to be the rough Dan-less consensus at this table. [laughter]

HENDRYCKS: Which is why we love you.

HODAK: We already have law. I think antitrust law can get us a lot of the way there, without even a totally new regulatory regime.

BARBARO: Okay. Well, we're going to leave it at that. We went over a little bit. My thanks to Max, Dan, Laya, Mati, Clem. A very valuable, albeit brief, cameo here. Thank you all. I hope you have a great rest of the DealBook day, and I appreciate your time.

PANEL: Thank you. Thank you, Michael. Thank you for having us.