AI and the Future of Philosophy

Simon Goldstein

Fangtang Institute, Tsinghua University
26 October 2026

Video: https://youtu.be/Gwffw_gsRUI

Transcript of the edited public video, including the talk and questions and answers. Speech fillers and repetitions have been reduced and obvious transcription errors corrected. Audience questions are condensed and labeled. Timestamps refer to the public video.

Introduction

00:00

Great. Thank you so much. It's an honor to be here. It's very exciting to meet so many people who are experimenting in how to use AI in the humanities. So what I'm going to talk about today is I want to try to raise some questions that all of us can think about of what are some of the biggest challenges moving forward in how we can use AI in the humanities. My background is in philosophy so I'm going to primarily talk about that, but I think the general themes I'm going to talk about apply very generally across the humanities and also beyond that into other areas of research.

Okay, so the background for this, look, in general, the way I'm thinking about it, okay, AIs are getting really good at research and what do we do in response? Well, what most academics are doing in response is ignore. Well, we're not doing that. Everybody in this room is interested in using AI. But even then, even if we're going to use AI in research, there's kind of two paths that we could go. One path is humans and AIs collaborating, and that's, I think, what this event is about.

The other path is AIs do it alone. And that's what I spend a lot of my time worrying about is, I'm really worried the AIs are going to get so good. What's our role going to be? So that's kind of what I'm going to talk about today, is trying to think about what are the opportunities for us to reliably, for humans, to continue to reliably add value, and where are the likeliest areas we can add value moving forward as AIs get better?

AI progress in mathematics

01:36

The backdrop for this, one thing that I think has been a big wake up call has been the improvements in math for AI. So I think math of all academic fields has faced the biggest risk so far of wholesale automation. And things have really changed. I think earlier this year and last year, a lot of human mathematicians were interested in actively collaborating with AIs to do math. And Terence Tao and others were, actually working with AI labs to try to do collaboration. But then what's happened over the last few months is things have changed.

And now it's not AI-human collaboration as much as, no, AIs are doing math. And so, the most striking example of this recently is OpenAI did this big, in October, released a series of new math results. So there's interesting limits to what happened with what OpenAI did. So for example, if you look at the proofs that, first of all, I can't read any math proofs, but apparently if you ask human mathematicians to read the AI math proofs, they're even, they're completely unreadable to humans so far. So far, definitely, it looks like there's still lots of room for humans to contribute, to taking these results and unpacking them.

It's very natural to think, well, probably they're going to train the AIs to be better. And it looks like, okay, at least the, I think there's a real chance that the process of proving math results may be 100% AI very soon. And then there's the process of taking the math results and disseminating them and understanding them and then we'll talk about that soon.

Which fields will AI automate

03:07

Okay, what should we expect beyond math? So where are things going? And I think we can, we can make some predictions here based on, we know a lot about why AIs have gotten good at math. They got good at math because it's relatively easy to train AIs at math compared to some other things. And the AI labs really want to train them. AI labs want to train AI at math because math is related to computer science and you want to train AIs to be better at computer science so that you can use AIs to make your own models better.

So that was already discussed earlier. Okay, so that's why they want to and then the reason they can is you can get really good data in math because for example, there's automatic verification and so you can actually, you can just get ground truth and know whether you got the right answer or not. Maybe one of the joys of the humanities is it's much harder to know whether you've done that. But okay, so in general I think what we should expect over the next, one to five years in terms of the order of automation of fields as well.

Fields where it's easier to have verifiability or access to ground truth. We should expect AIs to get better at that faster. We should expect AIs to get better in areas where there's lots of consensus among researchers because then you can kind of use as your signal for training. You can ask humans what was better or worse. If there's lots of consensus then, you're going to get somewhere on that dimension. And then also we'll expect to have more automation in areas where AI labs want to be working in that area.

So, I think that's kind of a general structure that we should be expecting. In general, yeah, AI has been getting very reliably, have been improving very dramatically over the past four years. So I expect that to continue. And then, yeah, well, there's a lot to talk about. So another reason I expect progress to continue is that there have been very reliable improvements in efficiency. So, every eight months basically you can get, you need half as much compute to get the same level of performance. That's been pretty reliable over the last four years.

So even if we didn't get increase in raw capabilities, just increases in efficiency might mean that then you can kind of stack the deck and have like a giant army of AIs for the same price as before you had one. And they're all doing work. Okay, so, the AIs are coming. All right. And then if the AIs are coming for us, what can we do? Well, definitely we can't ignore it. So we're going to be trying to still do our work in ways that are using AI, but how can we think about doing that?

Human and AI research collaboration

05:50

So the first model that I think this event is about and I'm very excited about and I've tried to work on myself is human AI collaboration or cooperation where humans and AIs work together. So I'll quickly walk through my own experience doing that and then talk about more general themes. So my own experience doing that is, well, this year I published a paper in a political philosophy journal that I kind of did it as an experiment rather than my usual

workflow where I really wanted to see kind of how far, what's the limits of, how much would I need?

I kind of working backwards of like how much would I need to add for the AI to be able to publish in a good journal because you can't, if you just right now, if you prompt Claude, publish a paper in a good journal, I think it's not yet. Maybe on the internal models, I don't know, but I have not seen a model that can just do it. So you have to add something. What do you add? And then the model that I was doing is I said, okay, basically I'm going to be the PhD supervisor, but kind of a German, very heavy handed and doing a lot of support, so more than I do for my human PhD students.

But for my AI PhD student, I did a lot of support, but I had rules, so I didn't write stuff myself. I'm just trying to give feedback to the AI on what the AI is doing. And I also gave it the initial idea for the paper, but I really, not like a lot of detail. Really, the amount of text that's on the slide is the amount of idea that I gave the AI. So I basically told the AI, all right, in political philosophy, the details don't matter. I said, in political philosophy, here's two theories, theory A and theory B.

In economics, economists have given an argument for theory A, but nobody's noticed that argument seems like theory B wouldn't have that benefit. So that seems like a problem for theory B. Okay, try to write a paper about that. Well, it's a lot of work to do that. You got to read a lot of stuff and getting the details right, and it was kind of interactive with economics, so there had to be some formal modeling, and you got to really think about how to set up the question.

A lot of work, but the basic idea, and one reason for that is in my own experience so far, I found at least in philosophy, every field's different, but in philosophy I found the initial idea generation, I've been relatively disappointed from AI so far, compared to, I find it's much better at taking an initial idea and executing it.

But finding an idea that's really, this kind of thing that could be published in a paper so far, I've experimented a lot with that, tried to have, okay, I've done things like, all right, try to go read 30 papers that came out, and, and find problems in the papers and come up with ideas. At least four months ago, it was like, no, no, you don't know what you're talking about. Okay, so, that's why I started here. And then what I did is I had Claude kind of iterate, I had to start with outlines, and then fill in the outlines maybe over four steps of outlining into more and more detailed structure, and then do drafts, and maybe do three or four drafts, so maybe like, eight steps.

Eight, iterations, and then each stage I gave a comment. So I probably gave, two or three hundred comments on the whole process. But then I gave it lots of feedback. And then after I did all this, I kind of made a little AI agent that kind of made it more of a pipeline. It's on my website for people to use. And, okay, here's kind of what I found in the course of, oh, and then, some of the comments I would do would be, kind of the structural comments I'd give my PhD students, where it's things like, okay, you've got like three main arguments you're trying to do, but I think two of them are bad.

Just focus the whole paper on the one argument that's your best argument. Or things like, okay, all right, you're going to discuss objections to what you did right now, because the AIs can talk forever. So, okay, you have ten objections you're considering. Actually, seven of the ten are not that interesting. Just pick three and just do those. So, a lot of comments kind of like that of just carving away the worst stuff and sticking with the best stuff. I once heard an interview, there's this comedian, American comedian, Louis C.K., who said when he writes

his hour-long comedy special, the way he does that is he'll like, usually like the last 20 minutes of a one-hour special are the best.

And so, what he does is he prepares the whole hour and then takes the last 20 minutes, deletes the rest, moves it to the beginning, and then prepares another hour and then keeps doing that until the whole. So, that sort of thing of trying to just keep deleting the worst material and keep the better.

AI writing failure modes

10:11

So, I found that was interesting. In general, at least in my experience, I found like the failure modes of where AIs are at right now for writing long papers are basically like the biggest thing I find is organizational. So, it just gets completely lost in large papers. So, that's why I like doing these iterative outlines where if you have it start with an outline and fill it in, I think it gets much less lost. Related to being lost is it'll be like very redundant. And I almost find like, sometimes I find it's almost like instead of just making an argument once, it'll like describe an argument without making it five different times, you know.

And so, again, I find outlining solves that. And then last, it'll, these AIs, especially Claude, kind of like they talk, in their own language almost. And I feel, I almost feel like I spend so much time talking with Claude that I'm learning to speak Claude. And then like I don't even realize how strange like the words are, you know. So, those are some of the big problems that I've experienced working with AIs so far. I expect some of these problems to change soon and some not.

So, in particular, the biggest practical problem I've had, which I think probably all of you struggle with is the writing quality of the models.

I expect that to change very quickly because I think that's explainable by the incentives of AI labs. So, especially Anthropic so far, their primary goal has been to train AIs to be good at math and coding. And the reason for that is, again, their focus is training AIs that can themselves be used to train AIs. So, to do that, you don't need Shakespeare. You need Newton. Shakespeare isn't helpful for training the next AI. But I think Anthropic has noticed that everyone is furious at them for how bad the writing is. It's like actually affecting them financially now.

So, my suspicion is over the next year or two, they're going to improve the writing a lot. Because, again, if you think of it from a training perspective, it doesn't seem that hard to improve writing quality. Because it's very easy to ask humans, for example, to say, okay, which of these was better written? And there's a pretty clear reward signal there. And it's so horribly written now. Everybody knows it's badly written. So, it should be relatively easy to give reward to improve the writing quality. So, that's at least one thing I suspect is the writing will get a lot better.

I think the, organizational things and redundancy seem to me more like general features of, agency that I kind of expect to improve. Like, that doesn't seem to be about writing as much as just, doing long tasks in a well-behaved way. So, I kind of expect that to improve as agency improves, but it seems like a harder thing to fix than writing.

The human researcher as a lab director

12:38

Yeah. Okay. Okay. So, okay. So, what would that look like? So, that was my experience. So, what would be, the structural version of this kind of collaboration? I guess, one kind of hopeful thought I've had about how the future could go is that basically, all of us researchers, we just kind of use AI to scale up our productivity. So, if we were all writing one paper a year before, no, now we're all going to write, 10 to 20 papers a year. And what that looks like is it's like every day you've got all these different papers you're working on, but the, basically you're the supervisor and your AI agents are doing the brute work of doing it.

And it's sort of like, I have a colleague, Herman Cappelen, who kind of has this vision. And it's basically like in the, it's like having science research labs, but in the humanities. So, I think in the sciences it's more often that PhD students take on the brunt work, the big work of executing the basic ideas. And so, you could imagine that and it's like all of us have a team of AI agents, each one, each group of AI agents is working on their own project. And then you check in with them every day and find the bottlenecks in their work and resolve the bottlenecks.

Okay, so this is kind of a, what you could imagine, is like each human researcher is basically now a giant research lab, director of a research lab, directing their AI agents, fixing their problems. And, yeah, that was my experience in working on this, is basically I'm reading all of these AI drafts, I'm giving lots of comments, I'm walking Claude through the stuff that it's not doing a good job at. And then it, but then it goes off and works for three hours and gets a new draft.

And, okay, I could parallelize that and have three different papers at the same time and try to, be reading each of them and giving comments. So, that's like one vision that sort of makes sense.

Autonomous AI research

14:27

But a different model, the scarier model, is pure AI research. So, in the scarier model, basically, no, it's more like the AI agents are going to be just doing lots of research. And then, yeah, okay, they're the ones doing research and humans aren't. And I guess I do think like, so that's not really under, I think there's an interesting question how much that's under all of our control. I think a lot of it is going to depend on how good the AI agents get and what the labs do. But it's an interesting question.

Also there's the thought that we don't know what's going to happen and so we need to be hedging and trying different strategies. And I guess in general, one reason I think this is a possible future is I'm pretty, I'm an optimist about different areas of the humanities and academic research. And I think in most of these areas, these fields produce lots of good data and have clear standards for success. But then unfortunately, if you have lots of good data, you have clear standards for success, then you can train AIs to be good at it.

So, and then that this is already happening. So AI labs now are hiring all sorts of academic researchers in all sorts of fields to be just generating training data on their specific areas of expertise. So, yeah, so what would it look like though? So this is the next thing I want to talk about. I think this also applies in human AI collaboration. So, another, so like in general, as to some extent research gets automated, what is it going to look like? Well, there's different stages of research. So there's hypothesis generation, there's conceptual reasoning, there's empirical work.

And you could imagine AIs doing work in any of these. There's already been studies about, for example, AI hypothesis generation in the sciences. Again, in my field of philosophy, AIs seem to be worse so far at hypothesis generation, but different fields are different. So you could imagine having specific AI agents that kind of are involved at each step in the pipeline of producing ideas. The next, so that would be on the production side. So you have AI agents that are producing the initial idea, AI agents that do lit reviews, AI agents that write papers, AI agents that critique the papers, so on and so forth.

Journals and research at scale

16:43

The next step of it, though, that's going to be a big factor is what, okay, so if the AIs are producing all this research, what do we do with it? And there, I think a big factor, which is also true on the first model, is we need journals that can house this work. And that's also true. And the first model where everybody goes from one paper a year to 20 papers a year, what are we going to do? We need 20 times as much space for the papers. Existing journals are not well set up for this.

And why would it scale so fast? Well, just to kind of, sharpen the intuition. So imagine that right now, imagine this year, an AI lab creates an AI that can produce a publishable paper for 10 US dollars. So once they can do that, two things follow. First of all, already it will probably be true that you can roughly produce 1 million publishable papers for 10 million US dollars. And the reason for that is, at least in my opinion, the way academic research works is there's not that much diminishing return. It's like there's a million papers you could write.

And so once you get the skill of writing one, you're not going to just like exhaust all of the ideas in like the entire field of philosophy. Because it's kind of like, at that margin of 10 million, I think there's practically an unlimited number of papers you could write once you know how to write papers. That's the first consequence. And the second consequence is probably one year later, the AIs can produce one paper for one US dollar. Because that generally is like the shape of the cost curve for compute. That things get roughly 10 times less expensive each year.

And so then, that's, so like this, that kind of shape means that in general, if we want academic research to be flourishing, we need scalable approaches to research dissemination. So right now, like existing academic journals, they're not set up that way. Because first of all, a lot of them publish like in print, and like you can't, you can't like increase the number of pages of your print journal by a factor of a million. And then second of all, all of the existing journals are completely bottlenecked by humans at every step of the review process.

And that whole system is already, but was already buckling even before AI. So I think another extremely important direction of emphasis when we're talking about the use of AI in academic research is I think there needs to be a lot of focus on how to use AI to increase the efficiency of journal review. And in general of review of the quality of the research, whether it's in journals or not. Because I think we're quickly moving to a future where we're going to have vast amounts of research being produced. And then we're going to have to, and then some of it will be much better than others.

And we're going to have to have a way of measuring that. And so that's, that's like the next thing I think is going to be really important.

Where AI research will happen

19:26

Okay. And then if we have all these AIs doing research on their own, where are they going to do that research? That's the next big question. So I am not convinced that like existing universities are especially well set up to accommodate that. Existing universities are focused on human capital. That's the bread and butter of existing universities is identifying talented humans and setting them up with resources to do research. But if AIs are doing the research, then it may look different than that. It may look more like, well, no, what really matters is you need to spend a lot of money on compute.

And then other things may matter too. Like it may matter a lot. There's interesting questions of what will matter. Like maybe it'll matter a lot kind of how you deploy your AI agents, what kinds of scaffolding or harnesses you use. That may matter. And there may be proprietary, maybe that, individual research groups develop very effective harnesses and those work better. I'm somewhat skeptical of that for, reasons somewhat discussed before. That, it may be that kind of gets brute forced away by improvements in AI capabilities. But, I guess one thing I think will be important going forward is that there are AI research labs that are kind of focused on, the focus of such a lab would be having AIs produce research in a field.

And I'm not sure if, I'm kind of confused about what the plan is. It may be that in a, it may be already, like I think already for math, like we're in a very strange situation where the AIs can do the math. Right now, like the people doing that or like the AI labs are doing that, I think primarily as some sort of like performance to show the world. But they're not like, it's not like OpenAI is planning over the next 10 years to be like focused on doing research in math. That's not their, that's not their job.

That's not their business. So, like it seems like already you would want universities to be extremely focused on setting up research labs like yesterday, like three months ago, this should have happened that the best universities in the world set up some research labs where they're spending millions or hundreds of millions of dollars, on compute to do math. Even that is like a very strange and interesting question because it's like, well, if we want the math results for the future, well, the cost of doing the math falls by 10x every year. So wouldn't it kind of make sense to wait a couple of years before you invest a hundred million dollars?

Okay, you can spend a hundred million dollars this year or you can wait two years and spend a million dollars. But then it's like, but then it's very disorienting because it's like, so do we just keep waiting, every year and then never do it? So it's very confusing to know when to pull the trigger. My instinct is that you want to start doing this stuff at a smaller scale and then kind of scale up as costs fall, but I'm not sure. Okay, so that's like a big question.

Social AI and robust collaboration

22:09

So if people are going to be building research institutes for AIs to do work, what should they be doing? Well, one thing is agentic harnesses. I think that's risky. I'm a bit more confident or excited about building social AI where you have lots of different AIs working together. So there's recently been a lot of discussion about so-called scaling laws for swarms, where the idea being you can look at how much increase in productivity do you get if you parallelize a task and have 10 different agents doing it rather than one.

And one thought is, well, look, no matter how much better the underlying AI agent gets, there will still be, returns from parallelizing and splitting into 10.

And so in general, I guess I sort of think that focusing, doing stuff like building research environments where AI agents can work socially together, that could be the kind of thing that's robust and it's going to be safe and valuable to do even as the labs make the AIs more capable. That's just one example. I think it's an important thing to think about. For anybody doing human AI collaboration, you've got to think about which ways of collaborating are going to be robust or which contributions can we have that are going to be robust and not just get wiped out in six months.

Turning research into public benefit

23:18

I think the next big topic is research dissemination. So, we have all the AIs doing the research. What do we do with it? And the issue, obviously, look, if AI suddenly are writing a million papers a year, no, humans aren't going to be able to read all those papers. So it starts to get very overwhelming, what's the point? Well, look, if it's curing cancer, we know what the point is, because then we're going to get the cure for cancer out of it. And humans don't need to read all of it. They can just take the medicine.

But we're not talking about that. We're talking about the humanities. So we've got to figure out if AIs are making contributions, how are we going to disseminate that research? And here I think we've got one answer is to draw on AI. So we're going to need AI journalists. We're going to need AI media teams. Already I've noticed on my Twitter, for example, there's a new genre that's been coming up of people taking academic research and using AI to turn it into, music videos and things like that. And so I think we're going to need a lot of that to disseminate the research that's happening.

Similar thing, there's an economist, Jason Furman, who this week put out this new website that is using AI to give economics briefings. It's, again, research dissemination. So I think that's the three parts of it, is you have the research side, then you have the journals, and then you have impact. And in all of those places, I can imagine both human AI cooperation and then pure AI and looking at building applications to do these things. So I think that's kind of the shape that I see for it.

What are the humanities for

24:49

Yeah, okay. So, but then in the time I have left, I want to flag, big picture, what are we trying to do here? What should we be doing? I'll be brief because I only have so much time, but, I think, the way I think about it is, what we should be asking, especially in the humanities, is what is our goal in doing all this research? And how much of that goal can be achieved by pure AI versus human AI cooperation, or [unclear] no AI? So I think that's, a good question.

And then here are some of the goals. So here are goals that I think can be achieved even by pure AI. It's like, one of the goals is discovering truths, okay? And AIs can discover those truths. Another goal is understanding the truth. That means you don't just know what's true, but you have a theory of why. You have an explanation, maybe a causal explanation where you have understanding. Then there's knowledge that might be more than truth. And finally, public impact. And again, I think, so taking the truth and sharing it with humans.

And that can be done by AIs too, especially once their prose is better. So those are some of the things that can be done by AI. But I think there's other things that can't be. And several of the things that can't be done by AI alone are really essential to the humanities. So I think some of the things that can't be done is things like self-expression, artistic expression. And

things like self-knowledge and self-understanding and ethical decision-making. Speaking for myself, the reason I became a philosophy professor is that I wanted to know, I wanted to work out my own views and have my own understanding of questions like how should I live?

What are my ethical obligations? And that's why I became a philosopher. I think most of us have an origin story where there were questions we wanted to know the answer to, that were about how to live, how to guide our own life. And then, okay, we became researchers. Now then the immediate next question is, is the current system working very well for achieving those goals? I don't know, I myself found that that's why I got into it, but then the reality is that's not where I ended up. My own research has not really been a satisfying way of achieving those goals.

And instead what I did as a researcher is I switched from these goals to the earlier ones. I instead kind of joined the team and joined a larger collective project of contributing to research and discovering truths. And okay, what I discover doesn't do that much to change how I live my life, but it's a tiny contribution to this larger pool of research everyone can draw from. Okay, but that's something that I think the AIs can do. What the AIs cannot do for us is make our decisions about how to live and they can help us do that and that's where human AI collaboration comes in.

But they can't do that alone. But then my point is actually existing academia doesn't do that either. So what I really think is that as AI comes in, we've got to get, we have to really restructure what we think our purpose is in the humanities. So I think so far in the last 50 years, the humanities has kind of more and more looked to the model of the sciences and said, ah, what we do in the humanities is we discover truths by publishing articles in academic journals where there's consensus within the field about what counts as progress to discover those truths.

And then they get disseminated, we write up articles in newspapers and that's what we do. But I think all of that is something that can be done by AI. And I think that humanities professors have by and large abandoned many of the, goals of humanities in the 19th century that was about helping themselves or helping others, hopefully, really figure out how to live. And I think that's what's going to be left for us to do. I think that AIs can help with that. And I think many of the contributions that you guys were exploring are examples of that.

But it's interesting. Many of these contributions you guys are exploring are not just canonical examples. All right, publish an academic paper in a, publish a paper in an academic journal, you know. And so I think that's going to be a big thing is we need to kind of push more to, and then when we're doing that, if it's about self-discovery, it can't just be about discovery for university professors. That can't be what it's about because we're paid, we have a, we're paid money by the government in most cases.

And we have to make sure that what we're doing is somehow, so it's got to be, we have to find a way where we're giving people, a guide to how to live.

We have to help people find how to live. And that can't be just by publishing papers in academic journals. Because when, look, you take papers in academic journals, you show it to your parents, that's not helping them figure out how to live. We all know that. Okay. And so I think we've got to move away from that and towards a very different paradigm where we're more like, basically, we're just very complex kinds of therapists, where we're drawing on the greatest art and ideas ever made to give people guidance in life.

That's at least what I've come up with, and the methodology here to think through this, I've tried to think about what are the different goals associated with the humanities or research in general.

And which of those goals can just be done by AIs. And then for us to have a job, we've got to find the goals that aren't done by AIs. And then we have to think about just start over. Rebuild universities from scratch. It's got to be those goals. And then how would you optimize that? And it ain't going to be by having academic journals. That's not going to be the way. We'll have those academic journals. They'll be done by AIs. But then what we're doing has to be different than that. So that's my suggestion.

Judging philosophical progress

30:13

Audience question condensed: In mathematics, results can be checked using Lean. Philosophy lacks an equivalent verifier. How can we distinguish genuine discoveries from new combinations of existing ideas?

Simon Goldstein: Thank you. Yeah. So math, we have Lean. We can tell whether the proof works or not. But we don't have that in the humanities. Even in philosophy, we don't have that. But I think what we do have in the humanities and throughout academia is we have academic journals. And then I think in most fields, these academic -- everybody -- I know. We're all bitter researchers and we all complain about academic journals. But I do think at the end of the day, academic journals usually have, standards for what's better or worse, that there's roughly broad agreement in the field.

And there's enough agreement between us that we can all kind of agree on standards. And we can all kind of be trained on the reward of getting published in these journals. And then, yeah. And then I think, that's correlated with doing better work in each field. And then we all get that -- we get our dopamine from that. And then we all get better at our work from that process. So that's kind of how I think about it. So it's not Lean. It's not, objective -- it's not just, an objective complete answer.

But I think we still have, a signal that is, raises the probability of doing better work rather than worse. That's how I think of journals. But then, again, so you can train -- I think there will be more and more training of AIs basically against the same signals. Because I think what's happening is, the labs are taking human researchers and having them say what work is better or worse. And basically, when the humans are doing that, they're kind of saying, "Well, this would be more likely to be published in the journals in my field."

So that's kind of how I think about it. But I agree with you. Math is different than the humanities in that way. But I'd like to think that, most fields still -- it's not as objective as math. But there's at least enough consensus that we're moving in the direction of progress. Richard?

The future of universities and teaching

33:34

Audience question condensed: Universities have roles in both research and teaching. Students increasingly turn to ChatGPT or Claude instead of professors. How will the role of universities change as AI takes over more of these functions?

Simon Goldstein: So I think in the spirit of my last remarks in the talk, is that, yeah, I think in general we should be pivoting away from research towards teaching. But I also think teaching needs to take a more expansive role of really focusing on helping students think through the fundamental questions in the humanities. Rather than, I think there's a lot of different goals from, I think of teaching for universities in the first instance is about improving

labor productivity in the workforce. That's why it's publicly subsidized. But as more and more jobs get automated, that will be less important.

And then I think, but then I think there will still be an extremely important role for the universities that will be more focused on the humanities, which is helping people think through how they want to live their life. And I think as work becomes less and less a part of our life, that will become more and more important. And I think universities will play a role more like traditional cultural institutions or religious institutions or things like that as offering guidance in how to live life. Now you brought up, okay, students are using AI more.

That's the place I'm most optimistic. In my opinion, that is a completely solvable problem that's trivially solvable by just having in-class assignments where you don't have computers in the room. So I think that to me, I never understand what the worry is. Like the one thing I'm confident of, we can stop students from using AI because we can just have in-class assignments. That's what I do. I don't even, it's like you don't even have to tell them not to use AI. If you just have in-class assignments with no computers, you don't have to make, you don't have to be the police officer.

It's just that they're living in your world. So, teachers have the power and students don't. So I think that side of it is easy enough. I also think that the demand for teaching will outlive the demand for research. Because I guess I'm thinking at some level teaching was automated in the 19th century with textbooks. But the reality is humans cannot learn that way. Humans need to, like basically I think a lot of teaching is like kind of like a personal trainer at the gym. And you just can't learn by yourself. People can't do it.

So, but then I think, there will be less demand for learning related to improving labor productivity as more and more of the economy is automated. So as we, I think we're, if we move towards a post-scarcity economy, I think basically the humanities will remain and the other parts of the university will all be gone. So I think that the humanities will be, humanities broadly construed to include arts and then basic science instruction, I think, will be all, all that's left and maybe not. And then I don't know if everyone will go to universities anymore. It'll be whether people, feel that, that that will help contribute to their lives, you know.

Education and employment

37:21

Audience question condensed: How do these changes to universities also affect K–12 education and the way we educate the next generation?

Simon Goldstein: It's a very hard question. So, yeah, how will education change? I think that a lot of how education changes will be sensitive to the rate of automation across the economy and which jobs get automated in what order and at what speed. So already just in like, for me in the last 15 years in my life, I've noticed incredible changes to the composition of which students major in what fields in response to just changing in jobs. So, when I went to college, a lot more people did the humanities and then like basically like kind of the year I got to college, everyone pivoted to doing computer science and things because that's where jobs were.

So, I think a lot of the answer of how education will change will be basically decided by demand from students based on where there's jobs. And so then that's a question of which parts of the economy you expect to be automated first. So that's, so then one way to think about that would be the broader question of how, which sectors of the economy do we expect to be automated, in what ways by AI. And the most compelling answer I've seen to that, which is a very rough answer, is from this economist Alex Imas, who's at DeepMind.

And he said, well, really the sector of the economy that's going to do best in response to automation is the relational sector of the economy. So the kind of simplest example is like a waiter at a high-end restaurant. So like, at many restaurants, waiters have been replaced with QR codes. But a waiter at a fancy restaurant is never going to be replaced by a QR code because people go to the fancy restaurant to have the discussion with the waiter. And so that job is like basically safe for all time.

There's like nothing, maybe, maybe if there was like an extremely sophisticated robot that was like, could like recite, poetry in response to the menu item.

But, I think that's a very safe job. So in general, yeah, I guess I would be thinking the sectors of the economy that will do better are the ones where part of the actual product that's being sold is a human experience. And then you can work backwards. If that's the sector of the economy that will persist under automation, then you should expect that also in universities that will be where, there's more majors. And then that's difficult to say what that looks like exactly. So I guess hospitality and areas like that.

But then other things as well. Maybe therapy, psychology, things like that. But very difficult question. So I'm sorry my answer is not that good.

Why value AI philosophy research

40:12

Audience question condensed: Why do we want journals producing large amounts of AI philosophy, even if we can see the value in AI cancer research? What about AI detection tools and publishers that refuse to publish AI-generated research?

Simon Goldstein: Well, let me ask you, the first question is, why do we want existing academic philosophy journals? I swear to God, if you open up a random article in an academic philosophy journal and try to show it to your parents, good God, it's not a single word is legible. So, but, the general structure that I think is whatever the reason is that we have journals now, there's, I think, a lot of those reasons generalize to then having AI to do it. And so, I think the reason that we have philosophy journals is that we think that it's good to discover new truths in philosophy.

And then the thought is journals, they're not perfect, journals are not Lean, but they're correlated with progress. And so then we have the philosophy journals because they're making progress in answering, philosophical questions. And then the thought is, well, if the AIs really are better than humans in philosophy, then they can make the progress themselves. Now, again, my view is that is not the, that is not all of, so I guess the way I think about it, that's not all of what philosophy is. That's not why I became a philosopher.

But then, in fact, my job today is not why I became a philosopher. So that's why I think what we all need to do as the humans involved, we have to return to the aspects of the humanities that were important before the academic institutionalization of the humanities that has moved it away from all of the reasons that, kids get interested in questions in the humanities. Because the kids were interested, and then if you show them all of the research outputs in the humanities, it's catastrophic, but then I do, but that's, but I also think that the work being done right now is important because it is important to make progress.

But while it's important, it's something AIs can do. So that's kind of what I'm trying to say, if that makes sense. Okay, thank you.