DATA& SOCIETY

March 19, 2026 | Online

The Craft of Science with AI: Evidence, Judgment, and Practice

Speakers: Kristin Branson, Lisa Messeri, and Nicole Nelson
Moderated by Ranjit Singh

Transcribed by Sarika Ram, Law Student, New York University School of Law

Ranjit Singh (00:00:07):

Good afternoon and welcome. My name is Ranjit Singh. I'm the director of the AI on the Ground program at Data and Society, and I'm very glad to open our workshop, The Craft of Science with AI: Evidence, Judgment, and Practice, with this public conversation. This workshop grows out of an ongoing ethnographic fieldwork project at Data and Society, where I've been trying to stay close to the places where scientific judgment is made: in lab meetings, in draft manuscripts, in tool demos, in troubleshooting sessions, and in arguments over validation and verification that often do not appear in published accounts of scientific results. This workshop begins from the same premise. As AI is integrated into scientific practice, the practice of science itself is changing.

Public discussion of AI in science is often framed in fairly blunt terms. Is the technology good for science, or is it bad for science? That framing tends to attach a normative judgment before we have spent enough time understanding what is actually happening inside scientific practice. I'm sure many of you in the audience come to this topic with suspicion about the usefulness of AI, and many others come with enthusiasm about its promise. Rather than begin from this divide, I want to bring in from the ground of practice. So this panel is an invitation to slow down and attend to practice, to how scientists decide what to trust, what to treat as good enough, and how authority and expertise are being renegotiated as these systems move from novelty into infrastructure. It is also an invitation to think about what new research questions this moment opens up for those of us trying to study science as it is being changed.

I'm delighted to be joined by Kristin Branson, Lisa Messeri, and Nicole Nelson. And rather than introduce themselves at length myself, I'm going to invite each of them to briefly introduce themselves as they respond to my opening question. So let me begin. From your respective vantage points, what kinds of changes do you see taking place in the sciences as AI becomes a part of everyday research practice? What is changing?

How do you know it is changing? And what markers do you use to think about the shifts in the craft of doing science with AI? And I'm going to open by inviting Lisa to this conversation.

Lisa Messeri (00:02:25):

All right. Hello, everyone. Thanks for joining us for this conversation. Ranjit, thanks for moderating us and to my co-panelists, I'm sure it's going to be a great one. I'm Lisa Messeri. I'm an anthropologist of science and technology at Yale University. And my work broadly is about the practice of innovators and scientists and engineers as they work with different products to imagine different kinds of tech and societal futures.

To respond to the prompt that Ranjit offers us, I'm drawing from work that I've been doing over the last few years with my colleague and collaborator, Molly Crockett, who's a professor of psychology at Princeton, working in cognitive neuroscience and other disciplines and specialties. To think about change, where we began in our research was actually to think about the promise of change that was being offered to scientists, particularly in the early moments of the launch of large language models and other kind of next generation predictive and generative AI systems. Our goal was to look at what was being imagined as the future of science from ideas of self-driving laboratories to kind of this figure of the AI scientist. And in a paper we wrote in 2024 that was published in Nature, we took these visions seriously and said, okay, if we were to achieve a future in which this change has happened, in which there are self-driving labs and AI scientists, what will this have meant for science? How will science have had to change in order for this to be a kind of viable future? So the kind of TLDR of the paper is kind of the catchphrase that we offer, which is we probably will enter an era of science where broadly speaking, we are producing more, but understanding less. And we therefore proceeded to outline the epistemic risks that would come with such a shift, including the emergence of scientific monocultures and kind of broad categories of illusions of understanding, this sense of working with these models, primarily these anthropomorphized large language models, how that might make scientists more confident in what they know than they actually do.

And that confidence is then materially matched by the uptick of scientific publications, et cetera. So we made these predictions two years ago and we're like, oh, in five years, maybe we'll see if we were on the right track. And lo and behold, two years later, there's actually ample empirical evidence that some of the changes that we were imagining are in fact happening. And I might get into some of those a little bit later, but I just wanted to give that as the overview to how we were thinking about change, not as some of my co-panelists in the everyday of the lab, although Molly themself is seeing how this

changes in their lab and even myself in anthropology, I see how it's changing my students' lives, et cetera, but to instead really think about the implications of the promise and the vision of change and how we can arm ourselves now to decide if we want that change of science or if we want science to proceed in some other way. So that'll be my opening remarks.

Ranjit Singh (00:05:54):

Fantastic. So on that note, Kristin has been closest to the practice, so let's hear from her as to how things are changing in her lab.

Kristin Branson (00:06:03):

Okay. Yeah. So I'm Kristin Branson. I'm a group leader at Genelia Research Campus, which has a focus on neuroscience. My background is in computer vision and machine learning. So I did my PhD and my postdoc in machine learning labs. And I started a lab at Genelia, which has this neuroscience focus, not knowing any neuroscience, but trying to figure out neuroscience and figure out how to apply machine learning computer vision to neuroscience problems. And so a lot of my research is about taking things that are best practices that machine learning researchers know about, things like how do you design a good training dataset, how do you evaluate whether your method is working or not, and trying to make those things that biologists can use and incorporate into their research without having to directly work with a machine learning researcher. And so I've been doing machine learning for 25 years, 26 years at this point, so it's been changing quite a bit. Nobody used to care what I did research on, and now everybody seems to care about AI.

So some of the things that have changed. So my career started before the Deep Learning Revolution started. There was a Deep Learning Revolution in around 2012, and so that's when machine learning actually started to work. There were specific problems that it worked on kind of, like your camera could detect a face and there was things like that that could happen, but now things kind of are working in practice for a subset of problems that people really care about. So that was kind of the change of deep learning was making this stuff kind of work in specific occurrences, and that was supervised machine learning. So in that case, you were giving examples of exactly what you want the algorithm to do. So you would say, okay, I want to recognize species of birds. And so you give a bunch of examples of an image of a bird and then the name of that species. And the only thing that that algorithm would do would be exactly what you told it to do. And it was pretty clear to some extent how to measure how well it was working. How many times does it get the bird species correct or not? Then there was this change with

the large language models that started around 2017 when the first transformer paper came out, and then companies really training this on just huge models and giant data sets that is a type of unsupervised learning. So they're trained to do something which is not really what you want to use it for. They're trained to do this kind of next word prediction, but it turns out when you train them in this unsupervised way, they're able to actually capture a lot about the structure of whatever data they're trained on, and you can use them in all kinds of ways that nobody really knows what they are. And also nobody knows how to evaluate whether they're working particularly well. So that's a big change. So these companies are putting these tools out there. We have no idea how you should use them. And they're just kind of like, you go figure it out and see what happens. So it's really exciting as someone who likes to understand how AI and science go together to see the creative ways that people are using this, but it's also slightly terrifying because I don't understand how people are using these tools without knowing more about them.

And I guess the other thing that's been changing is the amount of focus on AI. As I was saying, nobody used to care about AI. Now everybody does. Everybody is working to incorporate artificial intelligence into their research. And one of the things that has been changing is the audience for AI. So for a while, it was researchers in AI putting out papers on machine learning and then machine learning researchers who had some context about how to tell whether this is a good approach or not, evaluating how well these things were working.

And now there's a lot more of the research is being run by companies and a lot of the audience who's evaluating doesn't have a technical background in AI. And that doesn't mean you can't do a good job at evaluating this, but it's just a different audience and people coming from different backgrounds evaluating this. So yeah, that's I think one of the main changes that I've been trying to understand is the role of expertise here.

Ranjit Singh (00:10:36):

Nicole.

Nicole Nelson (00:10:38):

Okay. Very interesting already. So my name is Nicole Nelson, and I'm a science and technology studies scholar who's located in a medical school at University of Wisconsin-Madison. And I got into this space through a sort of strange route in some senses, but one that's I think historically informative where my research is really on methods, cultures, and the sciences, thinking about how people create new methods,

adopt methods, validate them, what they think the capacities of a particular scientific tool are, et cetera. And some time ago, my group started studying rigor and reproducibility issues because that was a space where we saw people advocating for a lot of methods change. And one of the things that people interested in rigor and reproducibility were worried about is the kind of variation that comes from having humans do science because humans do things in different ways. They also do things that they can't necessarily articulate well, and they don't necessarily know the sources of variation that they're introducing. And so I got into studying automated laboratories or what is now being called physical AI in the AI for science world because some years ago they were seen as a potential solution to irreproducibility. If the humans are doing these variable things, we can just get some of the humans out of science, and that will give us more reproducible science. And what happened over the past two or three years or so is the convergence of huge increases in the capacities of large language models along with these developing techniques in lab automation that started to make those projects a little less about reproducibility per se and more about creating AI-ready data. And so this is one of the things that my group is tracking now is when these systems are now sort of focused on producing large amounts of data and data with the appropriate metadata and structure to be fed into training models, what happens to the practice of science?

And so just to give a little teaser example of some of the changes that we're seeing, in my context in the medical school, which has a very large basic sciences research unit with a lot of bench scientists. One of the things that we hear all the time in grant review panels or discussions of graduate students is like, can they make the technique work in their hands? And so this idea of the craft element of science is both a problem for reproducibility, but it's also a recognized skill and kind of source of pride that you have a grad student or a postdoc who can make this really difficult technique work in their hands. Now in the world of automated labs, that's no longer a question because making the technique work in their hands is exactly what we don't want to do in this experimental setup. And so what we have been tracking in my group is then trying to see, well, what does skill look like in here? How do you learn to trust in the data? If you no longer trust it because you've got a skilled postdoc that you've seen execute a technique, then what makes you think that this data is good data? So those are some of the things that I hope that we can chat about today.

Ranjit Singh (00:13:41):

That's great. So one of the interesting things about the way in which Nicole, you framed the problem, it seems like there's a lot of conversation to be had around what does it mean to actually think about evidence, and what is the standard of evidence here?

Primarily because we are simultaneously trying to make a lot of science AI ready in a way, but at the same time, the question then becomes what is specific about the practice, which kind of is a matter of learning the discipline ,and how do we actually know at the end of the day that something that we are getting out of the system is a contribution to the disciplinary knowledge that we are working on in the first place. So to some extent, I'm kind of curious about opening the space up in terms of thinking more deeply about standards of evidence and verification. And it can be about replication. And we are also seeing this, to Kristin's point, in machine learning a lot where there's a lot of work that has happened on reproducibility of papers on trying to see whether, can you repeat a technique on the same dataset and get the same results?

So I'm broadly framing this question around what does it mean for us to think through these core questions of evidence and verification as we are interacting with these systems, specifically in the sense of are we getting something which is predictive? Are we getting something that we can interpret, which is also a challenge here? And then ultimately, what is it that we can trust? So Nicole, I'm going to open with you and then we'll move around.

Nicole Nelson (00:15:12):

Yeah, that's great because in fact, I will probably talk about the input end and maybe others can then take up the output end. So in thinking about the input end, if you want to train a model, you have to accumulate quite a large set of data, most or all of which you will have not made yourself. And so one of the questions then that immediately presents itself for practitioners is like, well, how do I know that this is actually good data? So I've been thinking quite a lot with Ted Porter's classic history of science book, Trust in Numbers, where he argues for this historical shift in modernization processes where trust used to look like a face-to-face thing. You went to a market with a particular seller and you trusted that they weighed out a particular amount of grain for you because you knew them and because you'd been there many times before to a system that he calls trust in numbers where rather than trusting in the individual that's doing the measuring, you look for objective features of the measurement so that you don't have to worry about whether or not you have a relationship with the individual person you trust, for example, in the system of standards of weights and measures. And so I think that this is one of the things that we're seeing happening here is as people are dealing with larger amounts of data, they can no longer use face-to-face trust as a reasonable proxy because they can't know all of the people who have produced the data. And so they can't say things like, yep, Steve's a great guy. He really has his eyes on his students. He produces high quality work. I trust his stuff. And so instead, they have to look at properties and features of the data. And that's a very different way of knowing the data

than it is knowing who produced it, what conditions they produced it under. It also involves a different set of skills to be able to evaluate the sort of shape and parameters of the data rather than look at, for example, whether or not somebody used appropriate experimental controls. So in my view, one of the things we're seeing shifting is the trust in people to trust the numbers, and with that comes the skillsets that are appropriate for trust in people to the skillsets that are appropriate for trust in numbers.

Ranjit Singh (00:17:17):

Kristin, want to jump in?

Kristin Branson (00:17:20):

So I guess maybe there are two things related to what else Nicole was saying with respect to the amount of data and how much you trust particular types of data. So one is this question of how good your dataset actually has to be, how accurate your dataset actually has to be to be useful to be something that will add information to your training dataset. I mean, I think that's something that people in machine learning are actually quite interested in. There's a lot of these datasets that these big models are trained on are actually quite noisy and that the amount of data still makes up for that amount of noise. And then I guess the other thing that I go back to is I go back to, because my training is as a machine learning researcher, I actually try to go back to the math when I think about this of, what are the assumptions that are made by machine learning algorithms and how does the kind of noise in the data or the distribution of the data that we're getting, how well does that reflect the assumptions that machine learning is based on?

And so there are ways of thinking about some of these problems. So in terms of how much, so this would be more on the output side as opposed to the input side that Nicole was talking about. But one of the things that we've been thinking a lot about is this idea of, okay, I've seen this kind of approach work a hundred times in very similar problems to my problem, and so I'm going to trust because I saw that evidence that it's going to work again for my type of problem. And so this is the basic assumption that all machine learning algorithms are based on that you can think of this, if you have something that you think, is this algorithm able to tell whether my coin flip is going to be heads or tails? And you see it a hundred times properly predict whether it's heads or tails, you're going to with high probability, know that that is actually predicting what's going on. And so that's the basic assumption that all machine learning research is based on this kind of generalization principle. And so that's the type of thing that we are interested in thinking about for science problems. And the space of science problems is gigantic. And so how

do you know, okay, well, I've seen a hundred problems like my own problem. Well, I don't know what like my own problem means. And then you don't know how to measure whether it worked or not. So that's one of the places that our research is trying to focus on is finding the different axes of science problems to chunk it up into pieces where we can measure how well it's actually working on very specific types of problems and then give people confidence that it will work on their problem as well without them having to do a lot of evaluation.

Ranjit Singh (00:20:26):

Got it.

Kristin Branson (00:20:29):

Sorry, that was maybe more mathy than it should have been, but I still struggle to not think in math.

Lisa Messeri (00:20:36):

Oh, we love math. This is so really interesting. I was just thinking through some of the more recent work that Molly Crockett and I have been looking at, which has been more particularly in the human sciences and in the fields of cognitive psychology and experimental psychology where, Nicole, this goes back to one of the points you made about one of the big frustrations that scientists always have is with the human and how messy and unpredictable and terrible we are getting in the way of the particular or the universal with all of our particularities. And so it's perhaps not surprising, but still stunning, I think, that one field where LLMs are being seen as having a real value add to the sciences is in the human sciences and the question of whether LLMs are sufficiently human-like enough that they can be proxies for human subjects and human subjects research, such that you then never have to actually bother getting a human into a lab and you're instead have this really great scalable model of the human.

And this relates to some of the themes that both of you were bringing up because in our analysis of these AI surrogates as we came to call them, part of what we ended up kind of arguing and focusing on is how scientists come to feel confident that they can in fact substitute a LLM for a human in their research. And of course, this is done through replication. The question is, do LLMs perform in the same way that human subjects perform on classic economic games, for example, like the dictator game or any number of these studies that are used as proxies for something like generosity in the case of the dictator game. And sure enough, LLMs act just like humans on these games. So hey, we have these human substitutes, but of course what's missing in that picture is a

narrowed sense of the human that has already come to pass over the last 30 to 40 years of experimental psychology. So all these other techniques and strategies for controlling the human in the laboratory, which came from first controlling our bodies using chin rests or other kind of apparatuses, then actually controlling participants by putting them onto a computer where they're only pointing and clicking because that's even easier to control. And then given that that happened using platforms like MTurk to then crowdsource a more diverse study population, so by the time you get LLMs, well, of course an LLM can act like a human participant because the human participant has become someone who interfaces on a chat interface like MTurk or points and clicks. So the kind of ability for an LLM to replicate actually shows us this intense narrowing of what the human has become in this particular science. So we're also finding that claims of replication can be really illuminating for different disciplines to see the assumptions that have kind of become baked into the methods and practices over a longer historical period of time.

Ranjit Singh (00:23:51):

Absolutely. Also, because it seems to me that especially in fields where these methods are not particularly well established, a lot of the current effort seems to also be directed towards first showcasing the value of these methods and having publications where some of the work that is well established, can you replicate the same set of results using generative AI models, which then kind of shows that because we have found the same results, again, using this new method, we can establish it as a method that we can use. It kind of borrows on what Kristin was saying, that do we know these methods work? They have worked traditionally on certain things. Can we show that they work on these things that we already know about? And then finally we get to a point where we can say, can we know something new using these methods?

So there is this whole journey of how we do that work of verification or setting up the bounds of what it means to make a method legitimate. And I see that a lot of the work that is happening here focuses on that particular way of, one, figuring out what do we know, and what is our expertise? How do we translate that into something that is happening in the field? And then finally, can we bring the field back to these systems where we can actually translate that into data, especially in the ways in which you described the human subjects of this process, where can you emulate people? In many ways, the way in which Nicole describes her work, it becomes about, can people who are doing these experiments in automated labs replicate the kind of work that scientists are doing? So in many ways, what I'm trying to invite a little bit of conversation on here is that this other element of this whole journey is this question of where does expertise lie now, and how do we describe what expertise looks like, and how is it changing, and

what does it mean for scientists to then claim expertise on a particular domain of knowledge? We can either do that through claims to knowledge or we can claim that through basically thinking through these are the methods that we know really well, and all of these things are changing. So I'm just wondering about how you're thinking through this changing domain of how we connect this issue of tacit knowledge to the issue of expertise and then open up the conversation around what expertise in the current moment looks like. Nicole.

Nicole Nelson (00:26:24):

All right. Me first, I am happy to. This is a good, juicy question I would say. So maybe one thing to put on the table is some vocabulary from Harry Collins that I've been thinking with a lot too, where he distinguishes between contributory expertise, the ability to actually do the thing and interactional expertise, which is the ability to actually understand the thing enough that you can talk about it even if you couldn't physically do these things. And I think that this is something that the physical AI turn really challenges because doing the thing is no longer the source of your expertise in this world. And even where doing the thing is done by some combination of humans and agents, what it means to understand whether it's done well changes quite a bit.

So I'll give you just one concrete example to hang onto that has come up in my field work. Liquid handling is still a big problem for automated systems that are bench science wet lab kind of systems. We have robotic liquid handlers, they work sometimes, but they fail in others. And so when you're thinking about what failure to pipette out a certain amount of liquid looks like in a traditional wet lab scenario, you have a lot of somatic or kind of feedback that comes from somatic tacit knowledge where you can sort of see if the amount of liquid that comes out is not the same as in the rest of the wells of the plate, or you can hear what it sounds like when you're aspirating a bubble. But when that's happening with the robotic liquid handler, you can no longer get any of those kinds of sensory traces. And so instead, the thing that you have to read is a pressure graph of what it looks like when the robotic liquid handler tries to force out that liquid. And if it's compressing compressible air rather than incompressible liquid, you'll be able to see a weird trace and know, oh, there was a bubble in the line that didn't go well. So that's a fundamentally different type of expertise that takes us out of somatic land and it also takes us out of that contributory land and puts us into the world of something more like relational tacit knowledge or interactional expertise where many people could read that pressure graph and they don't necessarily have to have any experience pipetting. So it does change the constellation of people who could become involved in trying to adjudicate whether or not this experiment has been well done because they no longer have had to do the experiment themselves in order to make that

expertise claim. So I see that as potentially quite a big and consequential shift.

Ranjit Singh (00:28:56):

Lisa, would you like to go next?

Lisa Messeri (00:28:57):

Sure. Yeah, that's a fabulous example, Nicole. I'm just so excited for your ethnography and cannot wait to get my greedy little hands on it. It sounds just so good. I'm going to zoom us out a little bit, if you don't mind, with this question of expertise. And the conversation that we're having about science and AI necessarily has to happen in the framework of the big tech industry that is of course pushing these tools into all of our hands and making it seem as though our future careers, lives, et cetera, are dependent on us adapting or becoming obsolete. And at the same time that they are trying to sweetheart us into thinking that AI tools and products are good for us, they're also undermining this thing of expertise that we all really value within our careers and industries. And I'm just going to kind of mark one technique that I have found really interesting coming from big tech discourse in which I see expertise being belittled, which is an increasing use of the language of taste when it comes to these companies and spokespeople for these companies talking about what it means to work well with an AI product.

So I was particularly thinking about this when listening to an interview with one of the co-founders of Anthropic saying how their own hiring of programmers has really shifted that rather than looking for the person who has the best intellectual pedigree or the best training, they're looking for someone who has quote unquote the best taste. And when pushed further by the interviewer, he kind of clarified that he was using taste as a way that really was substituting for experience, which is also a substitute for expertise. And I just found this in many ways really offensive to reduce the kind of hard-earned expertise that we get by being in laboratories and through the study to this kind of more intuitive gut feeling of taste. So I just kind of want to maybe mark that that is a language that, of course, those of us in STS and anthropology can do a lot with taste and get very voir dilian about it and stuff. But there is something about that language that's starting to percolate that is really kind of disheartening when it comes to this question of expertise that I also just want us to have in the ecosystem more broadly that we're talking about.

Ranjit Singh (00:31:41):

Kristin, bringing you in.

Kristin Branson (00:31:43):

Yeah. I mean, this question has come up a lot with a lot of the conversations I've had because I spend a lot of my time programming, a lot of my colleagues spend a lot of their time programming, and what is the role of coding bots in programming going forward? And so at the level of moving around water or maybe reading a particular type of, sorry, moving around liquid or reading out a particular gauge is the syntax of coding. And that's something that I do find it very useful to use these coding bots for. There's an argument parsing of a Python program or something. I don't remember how to do that. It's very useful to just ask your coding bots to do that part for you. But I do think that this higher level of expertise and how to properly frame a program to make it something that you can, it's modular, you can add to all of the things that-- So I'm trained as a computer scientist, how you're properly trained as a computer scientist, those are really important. And I don't understand how people are using coding bots without knowing those things, also not knowing how to evaluate whether your code is working well or not. But it's a discussion I have a lot with my colleagues because there is another take on this that some of these things, well, they just won't be important going forward because it will just be coding bots talking to coding bots who are just writing this so you don't need good code anymore or something like that. And it's an interesting question. I would love it if someone could help me make the arguments about what are specifically the types of expertise that someone who has been programming for 40 years or something like that has that is not yet captured by a coding bot and whether we can actually try to ask whether coding bots are capturing that information to have particular evaluation metrics for those types of things.

Lisa Messeri (00:33:43):

And I know Nicole wants to jump in, but that is exactly what taste is doing is it's saying you don't have to. It's like creating a workaround to avoid answering that question in a very strategic and undermining way.

Nicole Nelson (00:33:58):

Yeah. I mean, Lisa and Kristin, what you're saying is really resonating with what we have seen in our fieldwork, which is that when we look at a lot of people that are showing up in these AI for science spaces, there are certainly some people who are enthusiastic about it, and they're generally people that are coming from the computer sciences background, but the people who are coming from wet lab backgrounds, they're there

with a little bit more of a sense of reluctance of they feel as though they're going to be left behind, and so they feel as though they've got to skill up basically in these new tools or else they're going to be left behind. And what's ironic about that is they're kind of creating the very future that they're afraid of in their actions in re the adoption of these technologies because I think one of the really important to note things that is happening here at the intersection of the space here with the big tech companies is a large scale explication of scientist tacit knowledge where they're being encouraged to take those skills that have traditionally made them the people that they are, and to then explicate it in a machine readable form such that it can be taken up by other actors.

And I see that Bart Penders has dropped into the chat here, and Bart and I and some co-authors in Germany have a preprint that has come out recently and will come out in Science's Culture, where we talk about the risks of explicating all of this expertise and tacit knowledge in terms of making it readily available for uptake by existing powerful actors like big tech companies. And so I don't have good predictions about exactly what might be done with all of scientists' explicated tacit knowledge, but the fact that all of this knowledge that has traditionally lived in the bodies of people is becoming machine readable and available to me seems like quite a big shift and one that positions powerful actors to be able to uptake and benefit from that knowledge.

Ranjit Singh (00:35:51):

That makes a lot of sense. And also it opens up this question of this way in which we think about why to a certain extent some of this work that is happening seems to focus or turn our attention towards what are these ways of knowing which allow us to think through. There's a moment of translation that happens in the way in which you know things versus how a machine knows things. And partly the challenge that both Kristin and Nicole are drawing out is this issue of how do you do that work of translation in a way that is accessible and remains, for the lack of a better word, epistemically defensible in the sense that you can defend the ways in which that translation is happening, and it becomes a new way of producing knowledge in itself. And that to a certain extent then makes expertise more than a question of tastee in a way. And it becomes a part of how do you know rather than what do you like in a way, because it seems to me that expertise and taste can be differentiated in this particular way of thinking about, this is not just about, this is my preference for something. It is fundamentally this question of, is this representative of something that we can actually hold true as we are thinking about the next steps of research and what that might look like in a way.

And that kind of brings me to the audience questions here, and we have a lot of them.

And so I'm going to basically ask the first one that we got. It's fundamentally focused on this question of epistemological indifference, which basically, to quickly translate it, it basically means that when you get an output from a large language model, the system produces statements without any relation to truth as truths. In a way, it doesn't really rely on the fact that, can you distinguish between something that is true versus something that it falls in a very epistemological sense? So how do we think about that particular form of simulation that allows us to basically operate and work with a tool that doesn't particularly care about whether an output is grounded in truth and still be able to actually produce knowledge that is disciplinarily sound, and what does that look like in a way? And Kristin, I'm going to bring in with you on this one because I think you are the closest to the kind of ways in which you're thinking about this problem.

Kristin Branson (00:38:33):

Yeah. So the whole point of machine learning is it's going to statistically accurately do things, whatever you've measured it to do. So I did some internships at NASA when I was a grad student and an undergrad, and there was no machine learning being done at controlling spacecraft for NASA because it wasn't okay, at least at the time, it was not okay to be correct 99% of the time and incorrect 1% of the time. But machine learning is inherently built on being correct some fraction of the time with it being okay for it to be incorrect some small fraction of the time. So I think that's to some extent going to be a thing about machine learning going forward. It's going to make mistakes all the time, or sorry, some fraction of the time it's going to make mistakes. It's just the way that it's built. So I guess that's something I am kind of okay with.

One of the things that we try to do is make the downstream analyses that come after the machine learning, something that is robust to that kind of mistake. So there's humans make mistakes also with, and let's say you're talking about bird recognition again, which is something we can quantify what a mistake means. It's okay if we're just trying to count how many of this type of bird are in this area versus that other area if we have a 1% mistake rate. So I guess that's kind of one of the things that I'm actually okay with as long as we have a sense of what that error rate is. And I think that's kind of the thing that's missing right now, again, is this kind of evaluation of how well these machine learning methods are working for particular types of problems. And I do think that that's something that scientists can really contribute to is proper measurement of this because I think as Lisa was saying, a lot of this is being pushed by companies who want you to buy their products and they're going to want to convince you that this is working.

And I feel like I, as a scientist, one of my roles is to try to measure whether that's

actually the case or not.

Ranjit Singh (00:40:46):

Right. It's also interesting that to some extent, if there's a task that both machines and people find really hard to do, it seems like consistency matters more than correctness in a way, because then you know that you're consistently wrong or off by a certain way. Does that sound right?

Kristin Branson (00:41:09):

It is comforting as a scientist to know what types of mistakes are being made and understand why they're making those mistakes. But if you kind of, again, me following through the math, sometimes that doesn't matter. And sometimes biased mistakes are actually worse than purely random mistakes in terms of certain types of analyses you might be doing downstream if you're doing statistical analyses later on. But yes, I do a particular type of science that maybe is different from all uses of machine learning.

Nicole Nelson (00:41:43):

Yeah. And maybe I want to jump in here to note that I think one important element to throw in this conversation is that algorithms will do these tasks differently and that that really matters because there are some areas where prediction is what you need or execution of the task is what you need. But coming from my earlier work where I was looking at animal models and these questions were live, but in a totally different form, sure, you can make a mouse do a thing, but the question is, is what's happening in the biology, in the brain of the mouse actually changing in the same way, such that then you can do the mechanistic perturbation work of then trying to figure out whether or not this type of change in the brain is going to help you find a drug that you can use to modulate that downstream somehow.

And so I think this is a crucially important thing to know is that not all tasks in science are necessarily predictive tasks. I was just at a really lovely talk this morning by Stanford scholar named Xiaochang Li, and she was talking about this shift in natural language processing from trying to model the actual processes that humans use to sort of forgetting about what the human does, whatever, and finding their own machinic way of doing it. When machines fly, they don't flap their wings like a bird, they soar like a plane. And that's great if what you need to do is get from one place to another. But if you need to understand the actual process, that's a huge mismatch, and I feel like that's an error that we can't afford to make.

Kristin Branson (00:43:11):

So that actually is one of the main focuses of our research. So there's two ways of using machine learning. One is to automate things that you just need predictions for. And usually those things are things that humans can already do and we just want machines to do it faster or automatically. And then this other part is doing these superhuman tasks of we got a bunch of data of a mouse doing something, and we want to understand that data. And so this is a thing that my lab works on is trying to build machine learning models of animals with the hope that we can discover something new. But I think a lot of this comes down to this question of what are you thinking? Are you thinking about the limitations of the algorithm that you're using while you're doing this? So whether that's you're doing something for prediction and you know that there's going to be a 5% error rate, or are you using this for hypothesis generation and you realize that it is hypothesis generation and not actually saying something true about mice when you've just made a machine learning model of a mouse, but it's hopefully a hypothesis that you can then go and test in a real mouse.

Ranjit Singh (00:44:23):

All right. So I'm going to pick on a question which has been upvoted quite a bit. This is a request to the attendees. If they find a question that they feel like they really want an answer to, please support them. And this question directly relates to what Lisa was talking about with human simulation and thinking about simulating human responses in a way. But this one is kind of slightly different. And it's also really related to the recent story about Anthropic's AI interview study automating 81,000 interviews in one go. So it's kind of simulating the interviewer rather than the interviewee, which is also an interesting flip on this question. So how do you think about how AI would affect qualitative methods and analysis given these studies and how this imagination of how AI interviewers are being imagined as one of the things that you can easily do now?

Lisa Messeri (00:45:18):

Yeah. So there's a team out of Stanford that a year or so ago had launched this thing called the AI Agent Bank. And it was this idea that they were going to create a thousand agents for researchers that would be the results of not a generic human, but a very specific human that an AI agent would have interviewed for two hours, that that human would have taken a bunch of surveys and uploaded their demographic information. And then they would have been ground truthed by playing a number of these economic behavioral games and showing that this generic agent or this generative agent then replicated.

So I've already gotten to stew on that problem when I saw the Anthropic 81,000 interviewees thing going forward. And so when I looked at that original paper, and I haven't yet looked too closely at the Anthropic study other than I scrolled quickly through their data vis and was like, I hate these people. Sorry. What I noticed in that initial paper, which was written by computer scientists, none of whom I think have ever done interviews, was that they had this really interesting metric of the words spoken by the interviewer and the words spoken by the interviewee. And it was roughly fifty-fifty. Now, as someone who has interviewed a lot of people as an anthropologist, I know that that proportion is way off, that part of what you're trying to do is to elicit someone to tell their story and limit the kind of guiding questions that you are asking. And I'm sure a lot of that is like bloviating sycophancy is like filling up a lot of the word counts on the part of the interviewee agent, but I just find it really hard to believe that the people who are validating these models and like these interviews have the tacit knowledge and expertise to really understand the purpose of a qualitative interview.

And then secondarily, the purpose of analyzing qualitative data. The Anthropic study took 81,000 models and then said, this is the largest qualitative research ever done. And I'm sorry, no, it's not. It is the largest quantitative analysis done, it's quantitative analysis done on a large language set, but that is not the same as qualitative research. So there is just a large misunderstanding of what qualitative research is, and that's our bad, right? That's like anthropologists and social scientists and those of us who value qualitative methods. We've done a horrible job of branding the kind of knowledge that we produce and why that knowledge is different from and complimentary to quantitative data. So I think in the same way that Kristin gave a really great call to action for lab scientists in terms of science needs to be there for in terms of validating these models and really thinking about how we can say whether these models are actually doing the thing they say they are. I think as qualitative researchers, there's also a role we can play in really explaining the point of qualitative research and valuing it.

Nicole Nelson (00:48:38):

I mean, Lisa said that so well, and I'm just going to jump in with the smallest anecdote on this, which was me training a bunch of scientists who were interested in mixed methods research on different kinds of qualitative data analysis software. And the fact that it was called data analysis software gave them an impression of what this software would do that was totally different. And so I walked them through some QDA software options and they're like, but this doesn't do analysis. This is basically like Zotero. It just stores your data in a structured way. And I'm like, yeah, that's right. You do the analysis. And that distinction between the software doing the analysis versus you doing the analysis was just completely a mismatch, cultural mismatch in terms of what these

scientists expected to be able to do with a QDA piece of software.

Ranjit Singh (00:49:27):

Fair enough. So we have one more question that have got a lot of votes, so I'm going to ask it. Can we talk a little bit about the tension between the discourses of inclusion, democratizing science through technology and the anti-intellectual discourses around expertise?

Nicole, you kind of brought up this point, so I'm going to start with you, but broadly, I think the question here is also about what does it mean to actually think about being able to speak in a way that allows for us to sound more scientific in a way. And this is also, to some extent, this challenge of there's a way in which language models emulate language that allow you to become a lot more versed in the discourse of a particular discipline a lot more easily than it was possible earlier. And that kind of opens up a lot of these different questions around the kind of publication, the volumes of publications that people are getting. It's created a whole set of issues around what does it mean for people to actually have a published accounts of their work. So in many ways, this question is also about what does it mean to actually being able to talk in a particular disciplinary language versus being able to contribute to a particular discipline and the differences between them, Nicole?

Nicole Nelson (00:50:44):

Yeah, I'm happy to start on that question of participation and democratizing participation because particularly for cloud labs, which are the infrastructure that I have been studying as the entities that we are pinning our hopes on for creating AI-ready data, the democratizing discourse is very strong, that we wouldn't have to be physically located next to all of this research infrastructure, nor would we even have to have to have the training to use these particular scientific instruments. All you would need is the ability to program in your sequence of experimental steps, and then anyone anywhere in the world could execute it. And so for example, the NSF's recent solicitation to build a national network of programmable cloud labs, one of the elements that they specifically include in there is they want these programmable cloud lab centers to partner with places like historically Black universities, our [unclear] universities, even community colleges, with the idea that people are then going to have access to this high-end scientific equipment in a very democratizing kind of way.

Now, I find myself deeply skeptical of this because I'm enough of a historian to look at what past technologies has done to know that basically when excess capacity becomes

available, it's not usually the people who didn't have access before that get to capture it. It gets captured pretty quickly by other people. And so for example, I hear a lot of scientists saying things like, wow, this will be a great future because we'll get to program our experiments and then we can go outside and touch grass while somebody else does these experiments for us. And so first of all, I want to say, who is the somebody else? Because in a lot of these scenarios, it's actually substantially de-skilled workers. And then the second question is, do you really think that you're going to get to go outside and touch grass? Because if you've read Ruth Shortscohen (sp?), more work for scientists to me seems like it's going to happen where the demands on you for increased productivity are just absolutely going to skyrocket. So infrastructurally, I think that the capacity for democratization is there, but we have to remember that we put these infrastructures in existing systems of social inequality that I think are really going to limit us from actually realizing that kind of potential.

Ranjit Singh (00:52:55):

Kristin, you want to come into this publications question in a way and thinking more deeply about ...

Kristin Branson (00:53:01):

Yeah, I mean, I think machine learning itself has become ... I mean, it's actually quite remarkable. Before there were things like PyTorch, for instance, you needed to know a lot of math and a lot of programming to actually be able to do machine learning. And so that actually has been, I think, one of the main things that's driven this kind of deep learning revolution is that it's much more accessible to people who don't have PhDs in machine learning to be able to do machine learning research. And overall, I actually think that's been really good. So you do see a lot of people who you can just pick up PyTorch and just try out something really, really complicated in 15 minutes. And I find that really amazing to some extent that you can have a lot more people trying things and feeling empowered to try these things.

I think the thing to Nicole's point that's happened at the same time is there's much less machine learning research being done in academia now, and it's almost exclusively being done... A lot of it, it's being done in these companies. And so there are two things that have been happening at the same time, which is the computational requirements for these techniques is such that you have to be at a company to be able to do it, or you have to be part of a really big group to publish something that is incremental compared to everything else in the amount of time that you have for you. And I guess that's the thing that's been changing quite a bit in machine learning is how you publish in machine

learning, how you get a paper into NeurIPS or ICML. You could in the past spend a year or two years working on a paper, being sure that it's actually correct, and it's doing something that is going to push the field forward. And that's just not the publishing model anymore. You really do have to publish within a few months. Otherwise what you've done has become obsolete, or somebody else has done something similar to it. And so I liked the point Lisa made during the introduction about the amount of publications, and just I don't know how to deal with the field of machine learning publishing right now. As an example, the International Conference on Machine Learning conference, the number of submitted papers went from 12,000 last year to 25,000 this year. And I mean, you just don't know how to deal with these kinds of numbers. So it's kind of like we don't know how to read the research anymore. We don't know how to publish. The good thing about the machine learning community is that they are very thoughtful in some ways. If we're going to involve LLMs and making good decisions about papers, how do we do that in a reasonable way where it's not hurting things? I mean, not that we know how to do that, but at least they're trying to explore ideas and think about new ways of doing things, but we are currently in a big mess of machine learning publications right now.

Ranjit Singh (00:56:22):

Fair enough.

Lisa Messeri (00:56:24):

And of course, it's a mess of its own making because no doubt there's some substantial fraction of that that is because of LLM assistance that allows shitty ideas to be readable at a faster pace. We're all facing that in all of our fields. I'll just return, I know we're nearly at time, but I just want to return back to that tension between democratization and inclusion. And this is an incredibly challenging problem. And I think one of the ways it will be productive for us as a community to think about it is by, again, making sure we're looking at it not only at our institutions, but at the larger ecosystem of knowledge, production, research, funding, et cetera, which extends outside of the university to the corporation, to politics, et cetera. Nicole already started to give us a way of thinking at that scale about some of these questions.

And what I would just kind of add is when a lot of us talk about democratization of ideas and inclusive projects, we are talking about creating a system which brings people in to be part of a knowledge community. And we are particularly focused on these tools which can lead to democratization and inclusion about being an augmentation to our existing systems and structures and our values. When we hear questions of

democratization and availability coming from maybe industry, certain subsets of our current kind of US political regimes, these are not about using these tools to augment. They're about using the tools to replace, to get rid of something like the university, to get rid of something like the research lab, such that everything lives in kind of a computerized, controllable platform. So we don't need to accept the language of democratization that is kind of coming down to us that we are asked to inherit because it does align with our values. And we just might need to be more strategic about the language we use and really being really upfront that at the bare minimum, when we talk about the goals of these tools, we are talking about them emphasizing our commitment to augmentation and delegitimizing these tools as engines of replacement.

Ranjit Singh (00:58:50):

Fantastic. That was a really fantastic final thought. And we have a minute left. So Kristin, final thought on this conversation.

Kristin Branson (00:59:03):

Why don't I give my minute to someone else? I really do appreciate as someone who's in machine learning, having these kinds of conversations, because so much of the time we're focused on the myopic, how do I make this method better? And it's lovely to hear all of your thoughts.

Ranjit Singh (00:59:20):

Fantastic. Nicole.

Nicole Nelson (00:59:23):

For me, final thought is I think it's really important not to just be looking at LLMs and not just to be looking at outputs, but to be looking at the infrastructure that feeds all of these things because I think that is arguably where the bigger rearrangements are happening. The LLMs produce really flashy, cool things. We feel the sense of the uncanny when we work with them, but for me, what's important is what's going on in the whole labor structure of science behind the scenes that are changing in order to feed into these processes. And that's the thing that we should really be focused on.

Ranjit Singh (00:59:52):

On that note, thank you so much for being the part of this panel and attending it. We still have about more than a hundred people in this panel, which is fantastic. Thank you so much.