AI and You

Transcript

Guest: Melanie Mitchell, part 2

Episode 143

First Aired: Monday, March 13, 2023

Hello, and welcome to episode 143! Today, my guest is Melanie Mitchell, Professor at the Santa Fe Institute. Her current research focuses on conceptual abstraction, analogy-making, and visual recognition in artificial intelligence systems.

Melanie originated the Santa Fe Institute's Complexity Explorer platform, which offers online courses and other educational resources related to the field of complex systems. Her online course "Introduction to Complexity" has been taken by over 25,000 students, and is one of Course Central's "top fifty online courses of all time".

Last time we talked about how she was enticed into the field of AI by encountering the work of Douglas Hofstadter, author of the famous book *Gödel, Escher, Bach: An Eternal Golden Braid*, and went to work with him, which led to a lot of research into the true nature of machine intelligence and which prompted her to write the recent book *Artificial Intelligence: A Guide for Thinking Humans*. There's a link to the book in the show notes and transcript.

You'll hear mention of the Chinese Room in this interview, and for the sake of people who've heard me explain that a dozen times already and want to get on with it, I'll refer you to one of the more recent explanations, with Rob Sawyer in episode 108.

Let's get back to the interview with Melanie Mitchell.

You describe in the book a system you were working on for creating or perceiving analogies, I believe. Can you tell us about that?

So, in my work with Hofstadter, my project was to build a system that could make analogies in a very idealized domain. We were looking at analogies between strings of letters like if ABC changes to ABD, what does PPQQRR change to? And to try and make an analogous change. And this is meant to represent analogy-making in the real world. Of course, it was a very stripped-down version of it. But the system was meant to be more general. So I built a system called Copycat, which was able to make analogies in this domain, and very specifically, it explained its reasoning. So more recently, a group of researchers tried these same analogy problems on GPT-3, and it was able to do an impressive number of them, but it also made mistakes. But it couldn't explain either when it got it right or wrong what its reasoning was. So there's something in it that's able to see abstract relations in some cases, but not always. But there's no transparency so that's definitely a problem with these systems.

So the example you have where your program would look at analogies, like if you have a letter sequence ABC and then you have another one that's BCD, what do you expect the next letter to be in each case? And this is about us forming rules in our mind. These are like intelligence test

questions, SAT questions. If we think about analogies like then another section of the test would be, “man is to woman as king is to...,” then there’s a program called Word2Vec that can figure those out from a big corpus of information because now we’re looking at experienced gain from general knowledge.

I’ll just say that it can do it in some cases, but not in the majority of such cases.

Now, if we feed these things tests, like someone gave ChatGPT I think it was the SAT test and it scored at the 52nd percentile. If we focus on the things it got wrong and said, “Well, it’s not perfect,” are we missing some of the impact of this if it’s already scoring on a level that wouldn’t cause someone to be kicked out of school?

Yeah, I’m actually writing something about this right now. People have given ChatGPT parts of the bar exam, parts of MBA exams, parts of medical exams, and it’s done sort of passively okay on them. So this is something that’s been happening in AI for a long time. We have these intelligence tests or these standardized tests; we give them to our programs. But the question is, how do we interpret the results? If it does really well on a particular exam, that exam was designed for humans and makes assumptions that if a human can solve this question correctly, they’ll be able to generalize. You know, they have absorbed the general idea behind the question. But in the case of ChatGPT, that assumption might not be true. So just giving it a test that was designed for humans and having it perform at a certain level doesn’t mean that it can do the same things that you would assume humans could do at that same level. And I actually tested this out. There was a question on one of the MBA exams that someone had given it and it had gotten the right answer and they said, “Wow, it really understands this concept.” But I then reworded the whole problem, made a completely isomorphic problem. You solve it exactly the same way, but with different wording, and it got it wrong. So there’s something going on there that’s different from what humans are doing, it’s not able to generalize robustly the way humans often can. I mean, humans sometimes don’t either and that shows that some of these tests are not testing the abilities that we want them to test.

I’ve experienced some of that brittleness myself. Some people have published wonderful results with ChatGPT that I’ve attempted to reproduce and that hasn’t happened. Sometimes all I need to do is hit the “regenerate answer” button and it comes up with a better one, which is intriguing. I want to quote - you’ll know where this is going - but the ending of your book, where you say, “The impacts of AI will continue to grow for all of us. I hope that this book has helped you as a thinking human to get a sense of the current state of this burgeoning discipline, including its many unsolved problems, the potential risks and benefits of its technologies, and the scientific and philosophical questions it raises for understanding our own human intelligence.” That sentence, of course, doesn’t mention AI; that was in the previous sentence. And then you go on to say, “And if any computers are reading this, tell me what ‘it’ refers to in the previous sentence and you’re welcome to join in the discussion.” So you know what’s coming next. I put the quote into ChatGPT, and I said, “What does the word ‘it’ refer to in the last sentence of the quote?” And it said, “It refers to AI (artificial intelligence)” in case I didn’t

know. So do you feel like you're at the point where you can have a discussion with ChatGPT that will be useful?

That's funny. So yeah, of course, I was sort of joking in that last sentence referring to the Winograd schemas. And you'd be amazed by how many people have written to me to tell me they've put that into ChatGPT and it's gotten it right. So I think it's worth considering that we could discuss it with ChatGPT, but I don't think that ChatGPT is thinking in the same way as the thinking humans that I refer to in the title of my book. But it certainly can mimic some aspects of thought very well.

I wonder whether some of us might be throwing out the baby with the bath water here. And I'll tell you why. There are conversations I don't have with ChatGPT for the same reason that you've just said, because I know there's nothing on the inside, and that it's only giving an imitation of self-introspection. Someone with less knowledge of AI than me had an extended conversation with ChatGPT where it produced a very convincing imitation of introspection. He was asking it about the nature of consciousness. Who are you? What are you? How do you do this? How do you think? And so forth. And it held him to an enthralling discussion for some time that he found useful. And so my thought about that is that if this was done as a Turing Test, if I didn't know that I was talking to a machine, I would be inclined to ask more questions and be held in that conversation longer by ChatGPT, than I am when I know that it's a machine. And I wonder if I'm losing something as a result of that, at the very least the possibility of holding that conversation long enough to find out more about what it really is doing. What do you think?

Yeah, that's an interesting thought. People have been talking to chatbots since the early days when ELIZA was created in the 1960s and have been kind of projecting understanding or consciousness onto them. A lot of people who talked to ELIZA found it very helpful. And in the same way, when we're talking to ChatGPT, in some sense, we're talking to a big collection of humanity that's been absorbed by ChatGPT, a big collection of the thinking of humanity that people have written down. And it's been absorbed by ChatGPT and it's now expressing things that are plausible continuations of the prompts. So it's a very interesting look into human thinking in a much more collective sense. That being said, yes, it can absolutely sound like you're talking to a human, although I think if it goes on for a very long time, you'll sort of see some non-human-like qualities to some of the things that it says. And interestingly, the Turing Test as it's been carried out, as you mentioned, the Loebner Prize, it's been very short periods where somebody talks to a machine or another human for five minutes or something. But if you stretch it out to say four hours, it's still not clear to me, and I don't think that ChatGPT would pass a Turing Test under those conditions.

I wonder what Turing would say for how long. I mean, he did give a time, I forget what it was. I think he said five minutes, didn't he?

He predicted that within 50 years, machines would be able to convince the judge within a five-minute period.

Seventy percent of the time, I think.

Seventy percent of the time. And yeah, it turns out that it's not that hard to do that. Perhaps saying something more about humans than about machines.

So in your research and teaching now about these areas, can you tell us what you're focusing on and does it explore or open up the direction more of where the technology is going or who we are as humans?

So, one of the things I'm working on now is another idealized domain for analogy-making that was created by Francois Chollet from Google called the Abstraction and Reasoning Corpus. And these are visual analogy problems using kind of constrained grids where a grid is transformed into another grid using some rule. And then you have to do the same transformation to yet another grid. And they're deceptively simple; the domain looks simple, and yet it involves all of these concepts that we humans have about objects and how objects interact with each other and geometry and a lot of spatial knowledge, things that in psychology have been called core concepts. And right now, there's no program that exists including any of these language models that can solve these problems.

Give me an example.

An example of a problem?

Yes.

Oh, well, imagine that you have a grid of colored pixels, and you have a square that changes to a pentagon that's made up of those pixels. I show you that demonstration, and then I show you another demonstration, a triangle changes to a square. Now I give you another one and I have a pentagon on it. What does that change to?

Hexagon.

Yeah. So you increase the number of sides by one. So I've only given you two sort of training examples, if you will. So it's very much few-shot learning. But it's a visual problem. So obviously language models can't do visual problems. Actually, you can give them characters that represent the pixels and the transformation and people have tried that and they can't do it. So it's a very subtle domain and it really shows how we are able to abstract concepts from very few examples and then apply them to some new problem. And there's just a lot of different problems here. But the idea here is that it's kind of the opposite of the approach with ChatGPT. There, you start with vast amounts of language that humans have written, and the system learns from that. That's very different from how babies learn, right? Babies don't learn by starting with language. They start by interacting with the world of objects and faces and other people and actively try to learn from things that they themselves are curious about. And then they acquire language. So the idea here is let's start with something more like what a baby is doing and try to have these core concepts emerge from a very restricted world in which you learn concepts and you abstract them. So now there's a competition for people working on this domain and a lot of people are getting interested, but it's still something that language models and their related visual versions can't do.

And I think it's going to be hard for these systems to be able to do this because there's not enough training examples. It's really a few-shot learning idea.

The problem that you give it has not got many examples, but our current AI models aren't trained on that kind of data. Yes, we have things that have been fed billions of images, but for the purpose of turning it into language, not visual concepts. So I feel like that hasn't been attempted yet. And in the same way that we haven't done a lot with audio or tactile data which would be of considerable benefit to robots. But I think you've identified actually a good way of making this test because one of the problems in computer science in trying to form these sort of AI discriminators, things like Winograd schemas and so forth, is that we need to come up with something that's testable, repeatable, measurable. And by its very definition, that ends up with being something that is numeric, computable. And that bar is eventually cleared by a computer, at which point we say, "Okay, forget it. We'll have to use a different way of telling whether it's AI." Which was what McCarthy was getting at when he said, "You tell me anything that you want to do - *X* - and I'll build a machine that can do *X*, and then you'll have to come up with another thing that you think only humans can do." And here you are, you've got something that'll be quite a while. Although I do wonder if the same sort of focus and resources were put behind giving AI visual data that would support that kind of research, how long it would be before they got there?

Yeah, that's an empirical question and I think one that should be studied. So that's one of the things that my group is interested in.

What else is your group doing?

We're looking at how humans solve these kinds of concept formation problems and we're looking at the interaction of language; how people describe what they're doing. And so we're trying to see if you can go from language to solving these kinds of conceptual problems, or if you have to go in the reverse order - solve the conceptual problems first, and then be able to describe them in language. So we're doing both psychological studies using humans and also doing these same kinds of things with machines.

It's an interesting philosophical paradox there actually. If you can explain how a human does something like that, then you can build a machine to do it. And so implicit in this search for that is the belief that it will be a computable thing, is it not? Wouldn't, say, John Searle, disagree with that and say, you won't be able to explain it because then the Chinese Room, which we've mentioned several times on the show, wouldn't work as his demonstration?

So this question about computability, I think, is a bit of a red herring because there's no proof that humans can do things that are uncomputable. The Chinese Room is more a thought kind of experiment to say, "This behavior of this human who's using this lookup table just seems like a conscious entity that would pass the Turing Test." But Searle sort of says it's obviously not intelligent and that's kind of his argument.

I think Searle is quite transparent that he just doesn't think computers can ever be intelligent and the Chinese Room is just one of those things. I almost feel like he's pranking us in a way, like "chew on this" because the Wikipedia page for it has something like 32 different rebuttals of it and it still doesn't stop him.

No. It's an intuition. It's kind of a *reductio ad absurdum*. He says, "Okay, now imagine that your computer's made from toilet paper and rocks or something. Clearly, that can't be intelligent," but where does that "clearly" come from? I don't see his argument as being based on any evidence. It's really an intuition.

Well, we're reaching the end of our time here. And so I'd like to ask for your thoughts about where you see your field of interest going in the future, first of all?

So one thing that's happened in addition to these large language models, which have kind of taken over discussion in AI, another thing that's going on is that more and more people in AI are looking to cognitive science for inspiration, in particular to developmental psychology, you know, how kids learn, how kids think. There's a program funded by DARPA called Foundations of Machine Common Sense, which is specifically to get developmental psychologists together with AI folks and to have machines do the kind of development that kids do. And I think that's going to be a really interesting thing to look at in the future. The other thing that I think is going to be big is multimodal systems. That is systems that are not just language, or not just generating an image, but systems that can deal with both image and language and also video and virtual reality and so on in which the language is more grounded in things outside of this language system itself. So I'm excited about all of these things. I think these are all going to create big buzz in AI and generate interesting demos and tools. But I wrote a paper called "Why AI is Harder Than We Think," and it has to do with why our intelligence is more complex than we think. And it has been since the beginning of the field. It still is. That's why I don't think that something like human-level intelligent AI is going to happen anytime in the near future.

Doesn't that actually argue against your actual research of trying to figure out how humans do think if on the other hand, you're saying that it's a really hard problem. You're making your own research sound like a very hard thing to, or maybe impossible to accomplish.

No. I don't say I think it's impossible, I just think it's hard. And there's nothing wrong with working on hard problems. But I think it's important to temper our expectations that we're going to have full-blown AI in the near future, that it's just going to magically come out of these systems like we have today.

And talking about learning how humans learn as babies, we had Alison Gopnick on the show back in episodes 96 and 97 talking about that. So do you have another book in the works? That you think will happen.

I'm thinking about it.

Enough time has elapsed that you have forgotten the pain, right?

Yes, exactly. It's like childbirth.

Yes. Exactly the analogy I used in mine despite the fact that I don't have that reference point. So how should people who want to learn more about you and what you've done and will be doing find out about that? We will have a link to the book, of course. But if you want to just remind us of that and where else they can find that information.

So my web page is melaniemitchell.me and you can find everything that I've written basically on there, including a new blog that I've started, and take a look at that.

Fantastic. Well, everyone, the book is *Artificial Intelligence: A Guide for Thinking Humans*. Melanie Mitchell, thank you very much for coming on the show.

Thank you so much. It's been a lot of fun.

That's the end of the interview. It's interesting how now the Turing Test has been all but sidelined. Not that long ago it was held up as the zenith of computer ability at imitating or replicating humans; even getting special attention as the McGuffin, if you will, in the movie *ex Machina*. We expected the Turing Test to be passed with a bang and instead it happened with a whimper, just like, now we just take it for granted that something like ChatGPT can do that but we don't find it terribly significant any more.

Some years ago, I included Hofstadter as saying what you've heard in this interview, when he saw Deep Blue beat Garry Kasparov: "My God I used to think that playing chess required thinking, now I realize that it doesn't," and as you've heard, that quote was really about saying not that humans who played chess weren't thinking, but that machines can do the same without replicating the human thought processes that are required for a human to do that. And I said those some years ago, that I wanted to be the first to say , "My God I used to think that passing the Turing Test required thinking now I realize that it doesn't." Well here we are. If we still had some thing like the Loebner Prize as an arbiter to prove it, then it would be perhaps more significant but they seem to have taken an indefinite hiatus, but I don't think there's a question that chat GPT would pass the Turing Test by the original and other reasonable definitions now. And it's just another example of the principle John McCarthy described even more years ago as, "When we do it we stop calling it AI," as in when we decide that when computers can do *X* it will be real AI, real general intelligence, and then later on when they actually do that now we say, "No it's not really that kind of intelligence, it's just a clever simulation, it's machine vision or something else." And no, we don't have AGI yet, but it is amazing how much we are able to peck away at the edges of that.

I notice how ChatGPT is coming into a lot of the conversations now as a kind of representative or exemplar in many situations that we used to talk about more theoretically, but now we can use ChatGPT to focus the topic on something that's happening now and what it means and where it might go.

In today's news ripped from the headlines about AI, researchers from IBM, MIT, and Harvard have created a benchmark for evaluating an AI model's core psychological reasoning ability – what you and I call "common sense" - that will enable them to build and test AI models that reason and learn about other minds the same way humans do. They call it AGENT, for Action-Goal-Efficiency-coNstraint-uTility. (Taking the N from the third letter of Constraint and the T from the second letter of Utility all reminds me of another AGENT – *Marvels Agents of S.H.I.E.L.D.,* which in their first episode someone outlined to

another character that it stood for “Strategic Homeland Intervention, Enforcement, and Logistics Division” and asked, “What does that mean to you?” and they responded, “It means that someone really wanted this thing to spell S.H.I.E.L.D.”)

Well, AGENT was inspired by experiments that probe cognitive development in young children. It is a large-scale dataset of 3D animations of an agent moving under various physical constraints and interacting with various objects. What that means is that they show, say, an animation of a 3-D world where an “agent” (that might look like a blue cube – the important thing is that it can move), is in a setting where it is separated from an object by a solid wall. This is all very primitive 3-D animation. The agent moves toward the object and jumps over the wall to reach it. The test videos show the agent, object, and wall in the same positions, but now in the next video, the wall now has a doorway. So now that’s the reference video, and now there are two more videos. In one of the videos, the agent moves through the doorway to reach the object. In the other one, the agent jumps over the wall to reach the object. And to humans, one of thee is expected – that you would go through a doorway – and the other is a surprise, in that you would jump over a wall when there’s a perfectly good door there.

This is surprising because jumping is not the most efficient way to reach the goal. Their model learns to label outcomes as expected or surprising in the same way that humans do, and this is taken as an indication of learning “common sense.” It might seem a long way from the sort of things that ChatGPT is already doing with language, but remember that ChatGPT is just reproducing patterns of language that it has discovered, whereas this is genuine learning about real world behavior.

If you’re enjoying this podcast, please remember to give it a 5-star review, and a like and a comment. Even the most successful podcasters ask for this regularly because they have learned that all of us depend upon that to reach new audiences.

Next week, my guest will be Elizabeth Croft, Vice-President Academic and Provost of the University of Victoria, British Columbia, and researcher specializing in how humans and robots interact. That’s next week, on *AI and You.*

Until then, remember: no matter how much computers learn how to do, it’s how we come together as *humans* that matters.

http://aiandyou.net