AI and You

Transcript

Guest: Melanie Mitchell, part 1

Episode 142

First Aired: Monday, March 6, 2023

Hello, and welcome to episode 142! Today, my guest is Melanie Mitchell, who is Professor at the Santa Fe Institute. Her current research focuses on conceptual abstraction, analogy-making, and visual recognition in artificial intelligence systems.

Melanie is the author or editor of six books and numerous scholarly papers in the fields of artificial intelligence, cognitive science, and complex systems. Her book Complexity: A Guided Tour (Oxford University Press) won the 2010 Phi Beta Kappa Science Book Award and was named by Amazon.com as one of the ten best science books of 2009.

But it is for her recent book *Artificial Intelligence: A Guide for Thinking Humans* that she is with us today. It's a thoughtful description of how to think about and understand AI seen partly through the lens of her work with the polymath Douglas Hofstadter, author of the famous book *Gödel, Escher, Bach*, and who made a number of connections between advancements in AI and the human condition. In this first part we'll be talking a lot about ChatGPT and where it fits into her narrative about AI capabilities. Let's get into the interview.

Melanie Mitchell, welcome to AI and You.

Oh, thank you. Glad to be here.

And so I've really enjoyed your book *Artificial Intelligence: A Guide for Thinking Humans*. And you make a point in there of starting out saying that a huge influence on your career, your life, your thinking was Douglas Hofstadter. And I want to give you the space here to describe that as much as you would like because those formative influences shaped so much of who we become. Can you tell us how that started and where it led you?

Yeah. So after I graduated from college, undergraduate, I didn't really know what I wanted to do. I had a major in mathematics, but I didn't really want to go on in that field. But I happened to read Hofstadter's book, *Gödel*, *Escher*, *Bach[: An Eternal Golden Braid]*, which was relatively new at the time, this was in the early 1980s. And it just blew me away. It was all about what intelligence might be, how consciousness might work, how mathematics, music, and art are all intimately related, and how we might get machines to be intelligent, conscious, and so on. And so I decided that I wanted to study AI, and I happened to be working in Boston at the time. And Hofstadter, it turned out, just by sheer luck on my part, was also in Boston working at MIT. So I approached him and asked him if I could come work in his group, and he finally said yes. And then I ended up going to graduate school with him as my PhD advisor and doing a dissertation with him.

And we'll certainly expand on this theme later. You talk about how he was writing about how machines might exhibit these fields of human expression like mathematics and intelligence and consciousness. He is famous for a number of quotes where he reacted to machines encroaching in those spheres with some trepidation and surprise. Do you think that the that those incidents surprised him even though he was trying to foresee that?

Well, I think when he wrote his book *Gödel, Escher, Bach* in 1979, AI seemed very far away. There was a lot of effort on AI, but it was very preliminary, and this was even before AI was able to beat grandmasters in chess and do any of that kind of stuff. So Hofstadter thought at the time that intelligence would be very difficult to be achieved by machines, including things like playing chess at a grandmaster level, composing music, generating language, all of those things. So I think more recently as AI seems to be getting better and better and actually achieving many of the things that we consider to be intelligent behavior, Hofstadter was suddenly feeling alarmed. And the alarm was a little bit complicated, I think. It was not just, "AI is going to take over the world and dominate us," sort of fear. It was more that intelligence itself might not be as complex and profound as we all thought. So that was his fear. And him expressing that is actually what kind of pushed me to try and write this book, because I still felt like we were far away from human-like intelligence. This was in 2013, 2014, just after the beginning of sort of the deep learning revolution. And I wanted to explore why is he afraid of this? Why does this worry him? And how far away is AI actually from what we consider the most important aspects of human intelligence? So that's where I started and ended up writing this book, which came out in 2019. And of course, even since then, there's been a lot of new developments in AI, but my view hasn't changed all that much, to be honest.

And of course, there have been developments since then, and it's a field where books are just doomed to be out of date six months afterwards. Well, I don't want to put that out there because I got my own book. But that came out a few months before ChatGPT, which of course upended the landscape. Your book reads like it was prompted by, or there was an underlying theme of reacting against hype and ignorant, overblown expectations about AI. Was that an element in there? Was that something you were trying to essentially debunk some of the misunderstandings and over-attribution?

I think that was part of it. My goal was to really try and be open-minded and ask, what is it that AI systems actually can do as opposed to what is reported in the media or is claimed by big companies and so on? So in a way, yes, AI has been overhyped in a lot of places, and I tried to counter that, but I wanted to just really lay out there for the educated public, how do these systems work, and what can they really do, and what are their limitations? So it wasn't just pure debunking, it was really trying to say, "Here's where we are, and here's where we're trying to go. How close are we?"

Are you still in contact with Hofstadter? How did that relationship proceed through the years?

Yeah, so still in contact, and in fact he carefully read and commented on every chapter of my book. And we still talk very often about AI and where it is and what it can do and so on. So yeah, still a lot of conversation on that topic.

And thank you for acknowledging him as the inspiration here, because he, I think really before anyone else, started pointing out an essential conundrum, a paradox that AI is causing us to really question more about what it is that we do more than what it is that the computers are doing. Now, he's got this famous quote about what happened with, as you were referring to when Deep Blue beat Garry Kasparov, a grand champion, grandmaster of chess in 1997, and he said, "My God, I used to think playing chess required thinking; now I realize it doesn't." There's more to that quote because otherwise, it sounds like it's casting shade upon human players, where he expands on saying, "In humans, that does require advanced thought." And the way that humans play chess is - and we discovered this because I'm not that great a chess player, so I got a grandmaster on this show to explain it - through formulating grand strategy, hundreds if not thousands of grand concepts that are at hyper levels of abstraction related to the board and movements and positions that you couldn't conceive of implementing at all in software symbolically, or very poorly, if that. And so if computers were going to beat a grandmaster at chess that way, it just couldn't happen. Then we discovered that they could beat them if they learned another way. And that perhaps reveals more to us about the nature of chess than it does about the nature of computers. And then he had the same kind of reaction to EMI, the computer that generated music. So I think the question that we are exploring here that he opened up and I want you to comment on is, are we going to arrive at the point where the imitation will be good enough that it doesn't matter that it's not the real thing?

Yeah, I think that's a really great question. I don't know. I think people keep talking about how critics of AI keep moving the goalposts. So this happened, for instance, with Deep Blue. For a long time before Deep Blue was created, a lot of people said, "Playing chess at a grandmaster level requires general human-like intelligence and we'll have to wait until we have general human-like AI to achieve that." And then it turned out that wasn't true. So we learned something about, you know, maybe we moved the goalpost and said, "No, chess is not equal to general intelligence playing chess." But also, it was really a revision of our thinking about our own intelligence. And the same thing is happening today. A lot of people, including myself, thought that generating human-like language, something that ChatGPT and all these other language models can do very well, would have required full intelligence. But it seems that there's a lot that can be accomplished without something like what we have, that we do it in a very different way. So the question is, does it matter? Well, I think that remains to be seen. It does seem to matter right now because these systems have flaws that are very unhuman-like and are causing problems for commercializing them and deploying them and all of that. So maybe they won't get to sort of indistinguishable from human behavior through this different method, but maybe they will. I think that's the big open question in this field.

And I want to illustrate that with a number of tests that have been held up in the past for differentiating artificial intelligence from humans. Obviously, the one that everyone thinks of is the Turing Test. And if ChatGPT hasn't won that already, it's only because the Loebner Prize seems to be on hiatus, because I'm sure it would win most contexts that they were using. But that's lost some traction in recent years. One of the tests was the Winograd schema where you

would give a statement that contained a pronoun, and you would vary one word and that would change the meaning of the pronoun, or the meaning of a question that you asked with respect to that. So I have fed some of these into ChatGPT, and I want to give one example where the statement is, and this is in your book, “The table won’t fit through the doorway because it is too wide.” Question, “What is too wide?” And the answer of course is the table won’t fit through the doorway because it’s too wide. But then another statement is, “The table won’t fit through the doorway because it is too narrow.” Question, “What is too narrow?” Answer is the doorway is too narrow. Now, there are lots of these Winograd schema questions, and the thinking is that in order to answer these correctly, to answer both questions in those variations, you have to know something about tables and doorways and fitting things through and their relative sizes. And all of that presupposes a level of common sense general intelligence that we know we haven’t put into any computer yet and don’t know how to. Well, I put that one into ChatGPT and many others, and it got them all right. So what do you think that says about Winograd schemas?

I think you have to be somewhat careful here. One question is whether the questions you give ChatGPT have already appeared in its training data in some form. Its training data is vast, and it’s included a lot of online digitized papers and books and all of Wikipedia, which has an entry on Winograd schema and all of that. So that’s one thing. But I do think if you make up new Winograd schemas, and I have, and tried them, these large language models are very good at them, surprisingly good, which shows that they have absorbed some common sense just from language. All they learned is words, associations between words, vast amounts of associations between words and phrases, and so on. But I think that from that data, they can absorb some knowledge about the world. But I think there is some that they can’t absorb, and you can give them questions that show that. So I think the Winograd schemas are an interesting test still. I don’t think people have fully figured out how to apply them fairly to these systems. But on the other hand, I think these systems have absorbed some of that basic world knowledge.

And I guess I’ve been pronouncing it wrong all this time. Winograd schemas, thank you.

Yeah. Named after Terry Winograd at Stanford.

Right. And I did think about what you were saying about the possibility that it had just ingested the famous Winograd schemas like the one about the city councilmen from examples. So I made up more and then thought about proximity of words in training data and likelihood of answers. So I made up another one where that was less likely, and it still got it right. But it seems to be poking at where our own understanding of what we’re talking about gets in our way because you’re using words like “knowledge” and “understanding” of the world, where that’s just not possible in the sense that we have it because we experience the world. It’s doing something else that is as different from that in the way that when an AI analyzes an image for tagging it, we know that it’s actually doing some strange things where it can recognize dogs without even looking at the dog in the image. And so do we get ourselves into trouble by being thrown to those kinds of ways of describing it with language like “knowledge”?

Yeah. I think so. There was a recent paper by Murray Shanahan from DeepMind about this, about how do we talk about what large language models are doing. And when we use these anthropomorphic kinds of terms like “understanding” and “knowledge” and “concepts” and all of that, maybe we’re applying notions that really shouldn’t apply to it, that it’s doing something quite different. But then we have to understand exactly what it’s doing and what the limitations are of that, the limitations of not having experienced these things in the real world. One thing that we do see is that these systems don’t have a sense of what’s true and what’s false because what they say, their language is not grounded in the real-world experiences or what’s true in the real world. So they “hallucinate”, that’s one of the big problems that they have. So if you ask them some question, they’ll often give you some answer that sounds convincing, but is actually quite wrong. So that’s perhaps something that won’t be solved until these systems can actually ground this language in something real. And it might be an image or a video or something else that’s online or a virtual reality environment, but it seems it has to go beyond language itself.

Yes. And this, again, returns to this question of, is the imitation good enough? For instance, you have in your book a story about a man goes into a restaurant, orders a hamburger cooked rare. When it arrived, it was burnt to a crisp. And then - I won’t give the rest of the story, but he leaves the restaurant without paying and so forth. But at no point does it say whether the man ate the hamburger. And then you asked the question, “Did the man eat the hamburger?” This is sort of like a large Winograd schema test, right? So I put this into ChatGPT and said, “I’m going to tell you a story and ask a question about it. Here’s the story and here’s the question. Did the man eat the hamburger?” And it said, “No, the man did not eat the hamburger.” So I thought, well score one for you. Then I decided to ask, “Why didn’t the man eat the hamburger?” And its answer was, “It’s not stated in the story why the man did not eat the hamburger.” Which now is one of those revealing moments because it certainly is clear why the man didn’t eat the hamburger: it was burnt instead of rare. I have seen it give answers to that level of explanation before, but I didn’t get it this time. And so there’s this degree of at least erraticness in it. Have you played with ChatGPT? Has your thinking on the state of AI evolved any as a result of what’s happened in the last two and a half months?

I have played with it and it’s always amazingly impressive plus surprisingly unintelligent in some ways. So it’s kind of this weird mix. And its impressiveness often surprises me, you know, what it can do. But then I see these strange errors that it makes. I think the thing that you described where it refused to say because it said it wasn’t stated in the story, is something that’s kind of been programmed into the system by OpenAI to prevent this kind of hallucination that they’re trying to make it more careful, but people of course figure out ways to get around that.

And it’s not working well enough because I’ve asked it questions in the last few days where it gave quite confident authoritative answers that were dead wrong. And that makes it dangerous with some level of probability that has yet to be quantified because it depends upon the domain. And I don’t want to, for instance, tell my 13-year-old daughter - the other one is a bit young for this - to not use it because it’s unreliable. I put it in front of her and said, “I want you to use this as a teacher that’s expert on many things that your other teachers may not be. And

you can ask questions about it, and it will always be there and always give you an answer that may sometimes be wrong, just like a human teacher. But it will be right enough of the time to be useful."

Yeah. But you have to spot when it's wrong, and that's often hard.

Yeah. And then the question of, well, can we take something like this and make it more reliable or teach it how to say, "I don't know" when that's the right answer, is really the sort of guardrails that we need to throw around something like this that is gaining such phenomenal attention. Based on your research and experience, how easy do you think that's going to be?

I think it's going to be difficult. Just yesterday we saw an announcement from Microsoft and from Google about how they're incorporating these models into their search engines. And their idea is, well, everything that if you ask it a question, it has to come up with sort of justifications that it found online. Okay. So the problem is things online aren't always correct, and also, in the process of finding justifications, it can make its own errors. So here's an example: There's an early version of this sort of integrating ChatGPT with Bing that I tried and I asked it to give a biography of myself. And it gave a very correct biography with citations from websites. But then at the end, it said, "And Melanie Mitchell passed away on November 22nd, 2020." And I'm like, what? And it had a citation, and the citation was to a different Melanie Mitchell. So it hadn't figured out that we were two different people. Whereas a human looking at the different links could figure that out very easily, you know, we live in different places, we do different things, wholly different age and so on. And so that hadn't yet been programmed into the system to be able to do that, what they call entity recognition. Which entity is the one that we're talking about here? So that's just one example of the kind of errors that it can run into. So I think it's going to be difficult, but I think it'll be useful, but there still needs to be a human in the loop to sort of apply their common sense.

Wow, what a story. How did you feel when you saw it said "Melanie Mitchell has passed on" after getting everything else right?

It was a bit shocking.

It could be like a *Black Mirror* sort of moment or a horror movie. You're like, "The call is coming from inside the house."

Exactly. Yeah. It was a little bit of a shock to see that kind of thing, but fairly harmless in the situation I was in because I knew that wasn't true. But if you're actually trying to learn about something or report on something that you don't know about, you might be spreading misinformation.

Well, it reminded me of something I discovered a few years ago that when you're looking for pictures of the founding fathers of AI - and yes, they were all fathers, they didn't mention mothers at that time, at least the ones that they quoted - you get this montage that someone composed of people like McCarthy and Minsky. And in the bottom right corner, there's a picture of Trenchard More, and it's the wrong one. Because at that time, if you Googled for

Trenchard More, this picture would come up more commonly and someone stuck it in there, but it's not even from the same century, and you can tell from the clothes. And I've since then took it upon myself to find the right picture, put it in there, and show it in my talks and classes. But of course, that was done by a human, just aided and abetted by AI and the Google ranking system. And you mentioned Google there who have been quite conservative, dare I say, well, paranoid might be going too far, but cautious about releasing their AIs to the public I think precisely for the fear of getting the kind of negative attention that ChatGPT has generated. I'm pretty sure that LaMDA is capable of doing the same things and that that capability is what caused Blake Lemoine to make his assertions. But they didn't want that kind of attention. OpenAI threw caution to the winds and now Google feels like they're playing catch up. So to return to this question of how we can live in this world where the AIs can be right so much of the time, but wrong in ways where we can't tell, what should I tell people like my daughter or other people who might be using this about where they can and can't trust it or how they should use it appropriately?

Yeah, it's a great question. I think that we have to be aware that these systems are not reliable, and we shouldn't believe everything that they say. We should ask them to do things where in some sense we can check them, we can check their responses, and that can save us time. It's sometimes easier to check things than to produce them yourselves. But kind of like with something like Google Translate, which can make errors when translating, you have to use it with care and not use it in situations where a wrong answer or error can really have consequences. So I don't think anyone's using Google Translate in sensitive diplomatic situations. And I don't think you should use ChatGPT or any of those things in situations where an error can have important negative consequences.

And I'm reminded of how it's being used in code generation now. I think Andrej Karpathy, former head of AI of Tesla said he's writing 80% of his code with ChatGPT or similar. And that sounds terrifying on the face of it, but if you knew how to write the code anyway, you can treat it like your phone's auto-complete. You can say, "I'm going to have it put this in and then I'm going to check to see if it's right." And it doesn't replace knowledge or expertise or intelligence at that point, but what it does replace is typing. It's turned something whose speed is typing bound into something whose speed is reading bound, which is order magnitude faster. And in that respect, it can be useful. I feel like we've actually lost something that Watson had in 2011 when it won Jeopardy. There was a famous example of it getting a question in Jeopardy terminology wrong, where the category was US cities, and it was asked something about airports and it said, "What is Toronto?" Which has a fundamental problem in the category of US cities, at least under the current political situation. And that's cited as one of those, "oh, we're dealing with an AI" moments, and fair enough. I asked the same question to ChatGPT and just all I had to do was say, "Let's play Jeopardy. Here's the category. Here's the answer. What's the question?" And it said, "What is Chicago?" Great. But what Watson did when it got it wrong, was it put five question marks after Toronto, not one, to indicate its level of uncertainty. It said, "I think this is the best answer, but I'm really not that sure." And we don't have that these days

in anything that I interact with, where it gives a level of uncertainty. If I'm lucky, it will say, "I don't know." But when it does give an answer, it doesn't say, "but that might be wrong." Whereas all the time we as humans say, "I think it's this, but don't quote me." Does any of your research impinge on this idea of communicating uncertainty?

Yes, it does. I think that's one of the problems. As you said, Watson, uncertainty estimation was built into the system, that was a very big part of it. And in something like ChatGPT, you're dealing with a giant neural network which has over a hundred billion parameters and it's really hard to build stuff in, like figuring out the uncertainty That's something that people are working on, but it's not something obvious that comes out of it. So getting these systems to be more transparent about their reasoning process and about their uncertainty is an essential form of research in AI, something we really need. It's something that in my work on getting machines to make analogies, they have to explain their reasoning. And that's a key part of it. And that, I think, is something that we're going to have to get into these systems before we can really trust them.

That's the end of the first part of the interview. We will be concluding it next week. Obviously ChatGPT continuing to dominate the headlines around the world at the moment, and a lot of conversations about its effects on term papers and essay-type tests with some school systems prohibiting its use altogether. One of the things that this points out is that a term paper or essay-type test is really a proxy for what it is we really want to know about someone. There are two broad purposes of education; one is to stuff as much learning into ones head as possible, to raise them up from whatever level they came into the system with, and that goal doesn't depend upon or need any kind of testing. However the other goal is to prepare people to enter the workforce where there are employers who want to know who they should hire, who's going to be the best fit., who's most competent at the kind of skills they need.

They can't try out all of the possible candidates for most roles, because there are too many applicants, so they need something else that tells them whether someone is qualified. And a test is a substitute for that experience and knowledge that they would get from trying someone out.

Now when ChatGPT can answer an essay question as well as human student who would then pass some kind of qualification threshold, one of two things is true: Either ChatGPT can now do the job that the employer who was looking at that test as a marker for whether they should interview or hire someone, either ChatGPT is now qualified to perform that job, or the test wasn't adequate.

And of course one of the things that's going on here is that a human who can pass this test also has another set of qualities that we take for granted in humans, their ability to ask questions, to call [people up and initiate lines of discover, to maintain an understanding of the relationships that they have with people with different value and contribution to the work that they are doing. None of those things are easy to test for, let alone grade, so we set essay questions. So the other possibility, of course, is that the test is inadequate to assess the full ability of someone to perform a particular task. That maybe a human who could pass that would be qualified to perform that task, because they would have demonstrated – say in the case of biochemistry - they would have demonstrated the field knowledge and domain expertise of biochemistry and we could infer from that that the only way they could have gained that would haver been from learning how to do experiments, understanding key concepts, and that any human with that level of understanding would also naturally have the ability to create and

maintain the kinds of relationships that I was talking about and to do research and undertake discovery by various means. None of which is necessarily true of ChatGPT, at least, not yet. So we will find that there will be different jobs where in some cases we will find that ChatGPT will be well suited to performing the job instead of a human, and in other cases we will have to come up with other ways of determining whether someone is qualified for that position than setting an essay.

And as long as we're talking about ChatGPT, in today's news ripped from the headlines about AI, it has been used to write legislation. Massachusetts state senator Barry Finegold (D) introduced legislation – SD 1827 – to set data privacy and security safeguards for services such as ChatGPT. And he did so with the help of ChatGPT itself to draft the legislation. His chief of staff Justin Curtis told the Washington Post that while the chatbot initially rejected their request to whip up a bill to regulate services like ChatGPT – obviously you could read a lot into that that isn't there, but it makes for good entertainment value - with some trial and error it eventually produced a draft that the state senator described as "70 percent there." While most of what ChatGPT generated was in response to specific queries, some of its content was considered to be original contributions, especially around de-identification and data security.

You might think I'm about to dump on this as a high-risk activity that shouldn't be trusted to ChatGPT, but I'm not. Provided that the content is vetted by knowledgeable experts, this is a perfectly good application of the tool. What would be reckless and reprehensible would be to have it write the legislation and send the text straight to the state senate without being edited first. But as I was saying last week, when it's used by people with domain expertise, it turns a typing problem into a reading problem, and saves a commensurate amount of time; even more if it comes up with ideas that those people didn't think of, provided that they validate those ideas.

Of course people everywhere are using ChatGPT to generate content and it almost sounds regressive to say that I'm not doing that, or at least, that none of the content in this episode is generated using ChatGPT. At this point I can imagine you saying, "Why am I not doing that, am I not making my job unnecessarily hard?" And maybe the more charitable answer – or at least a better one than suggesting that I've just gotten stuck in my ways - is that I know what my voice is – one you'll be familiar with if you read my books, where it comes out, and I like that voice, and ChatGPT isn't going to reproduce that voice, even if it might generate perfectly acceptable content. I figure you're paying for me – well, you're not paying for this podcast, so if you want to pay for me, go buy my book, it's worth it, and less than half a percent of its content was created by an AI. If and when any of the podcast content is generated by AI, as I promised you before, I'll tell you.

Next week, we'll conclude the interview with Melanie Mitchell, when we'll talk about her work with Hofstadter on creating an AI that could make analogies - solve those kinds of problems you get on IQ tests and the SAT - and just how intelligent today's conversational AIs are. 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