Talk Dec. 21, 2022

An A.I. Pioneer on What We Should Really Fear

Artificial intelligence stirs our highest ambitions and deepest fears like few other technologies. It's as if every gleaming and Promethean promise of machines able to perform tasks at speeds and with skills of which we can only dream carries with it a countervailing nightmare of human displacement and obsolescence. But despite recent A.I. breakthroughs in previously human-dominated realms of language and visual art — the prose compositions of the GPT-3 language model and visual creations of the DALL-E 2 system have drawn intense interest — our gravest concerns should probably be tempered. At least that's according to the computer scientist Yejin Choi, a 2022 recipient of the prestigious MacArthur "genius" grant who has been doing groundbreaking research on developing common sense and ethical reasoning in A.I. "There is a bit of hype around A.I. potential, as well as A.I. fear," admits Choi, who is 45. Which isn't to say the story of humans and A.I. will be without its surprises. "It has the feeling of adventure," Choi says about her work. "You're exploring this unknown territory. You see something unexpected, and then you feel like, I want to find out what else is out there!"

What are the biggest misconceptions people still have about A.I.? They make hasty generalizations. "Oh,

GPT-3[1] can write this wonderful blog article. Maybe GPT-4 will be a New York Times Magazine editor." [Laughs.] I don't think it can replace anybody there because it doesn't have a true understanding about the political backdrop and so cannot really write something relevant for readers. Then there's the concerns about A.I. sentience. There are always people who believe in something that doesn't make sense. People believe in tarot cards. People believe in conspiracy theories. So of course there will be people who believe in A.I. being sentient.

I know this is maybe the most clichéd possible question to ask you, but I'm going to ask it anyway: Will humans ever create sentient artificial intelligence? I might change my mind, but currently I am skeptical. I can see that some people might have that impression, but when you work so close to A.I., you see a lot of limitations. That's the problem. From a distance, it looks like, oh, my God! Up close, I see all the flaws. Whenever there's a lot of patterns, a lot of data, A.I. is very good at processing that — certain things like the game of Go or chess. But humans have this tendency to believe that if A.I. can do something smart like translation or chess, then it must be really good at all the easy stuff too. The truth is, what's easy for machines can be hard for humans and vice versa. You'd be surprised how A.I. struggles with basic common sense. It's crazy.

Can you explain what "common sense" means in the context of teaching it to A.I.? A way of describing it is that common sense is the dark matter of intelligence. Normal matter is what we see, what we can interact with. We thought for a long time that that's what was there in the physical world — and just that. It turns out that's only 5 percent of the universe. Ninety-five percent is dark matter and dark energy, but it's invisible and not directly measurable. We know it exists, because if it doesn't, then the normal matter doesn't make sense. So we know it's there, and we know there's a lot of it. We're coming to that realization with common sense. It's the unspoken, implicit knowledge that you and I have. It's so obvious that we often don't talk about it. For example, how many eyes does a horse have? Two. We don't talk about it, but everyone knows it. We don't know the exact fraction of knowledge that you and I have that we didn't talk about — but still know — but my speculation is that there's a lot.

Let me give you another example: You and I know birds can fly, and we know penguins generally cannot. So A.I. researchers thought, we can code this up: Birds usually fly, except for penguins. But in fact, exceptions are the challenge for common-sense rules. Newborn baby birds cannot fly, birds covered in oil cannot fly, birds who are injured cannot fly, birds in a cage cannot fly. The point being, exceptions are not exceptional, and you and I can think of them even though nobody told us. It's a fascinating capability, and it's not so easy for A.I.

You sort of skeptically referred to GPT-3 earlier. Do you think it's not impressive? I'm a big fan of GPT-3, but at the same time I feel that some people make it bigger than it is. Some people say that maybe the Turing test[2] has already been passed. I disagree because, yeah, maybe it looks as though it may have been passed based on one best performance of GPT-3. But if you look at the average performance, it's so far from robust human intelligence. We should look at the average case. Because when you pick one best performance, that's actually human intelligence doing the hard work of selection. The other thing is, although the advancements are exciting in many ways, there are so many things it cannot do well. But people do make that hasty generalization: Because it can do something sometimes really well, then maybe A.G.I.[3] is around the corner. There's no reason to believe so.

Yejin Choi leading a research seminar in September at the Paul G. Allen School of Computer Science & Engineering at the University of Washington. John D. and Catherine T. MacArthur Foundation

So what's most exciting to you right now about your work in A.I.? I'm excited about value pluralism, the fact that value is not singular. Another way to put it is that there's no universal truth. A lot of people feel uncomfortable about this. As scientists, we're trained to be very precise and strive for one truth. Now I'm thinking, well, there's no universal truth — can birds fly or not? Or social and cultural norms: Is it OK to leave a closet door open? Some tidy person might think, always close it. I'm not tidy, so I might keep it open. But if the closet is temperature-controlled for some reason, then I will keep it closed; if the closet is in someone else's house, I'll

probably behave. These rules basically cannot be written down as universal truths, because when applied in your context versus in my context, that truth will have to be bent. Moral rules: There must be some moral truth, you know? Don't kill people, for example. But what if it's a mercy killing? Then what?

Yeah, this is something I don't understand. How could you possibly teach A.I. to make moral decisions when almost every rule or truth has exceptions? A.I. should learn exactly that: There are cases that are more clean-cut, and then there are cases that are more discretionary. It should learn uncertainty and distribution of opinions. Let me ease your discomfort here a little by making a case through the language model and A.I. The way to train A.I. there is to

predict which word comes next.[4] So, given a past context, which word comes next? There's no one universal truth about which word comes next. Sometimes there is only one word that could possibly come, but almost always there are multiple words. There's this uncertainty, and yet that training turns out to be powerful because when you look at things more globally, A.I. does learn through statistical distribution the best word to use, the distribution of the reasonable words that could come next. I think moral decision-making can be done like that as well. Instead of making binary, clean-cut decisions, it should sometimes make decisions based on *This looks really bad.* Or you have your position, but it understands that, well, half the country thinks otherwise.

Is the ultimate hope that A.I. could someday make ethical decisions that might be sort of neutral or even contrary to its designers' potentially unethical goals — like an A.I. designed for use by social media companies that could decide not to exploit children's privacy? Or is there just always going to be some person or private interest on the back end tipping the ethical-value scale? The former is what we wish to aspire to achieve. The latter is what actually inevitably happens. In fact,

Delphi[5] is left-leaning in this regard because many of the crowd workers who do annotation for us are a little bit left-leaning. Both the left and right can be unhappy about this, because for people on the left Delphi is not left enough, and for people on the right it's potentially not inclusive enough. But Delphi was just a first shot. There's a lot of work to be done, and I believe that if we can somehow solve value pluralism for A.I., that would be really exciting. To have A.I. values not be one systematic thing but rather something that has multidimensions just like a group of humans.

What would it look like to "solve" value pluralism? I am thinking about that these days, and I don't have clear-cut answers. I don't know what "solving" should look like, but what I mean to say for the purpose of this conversation is that A.I. should respect value pluralism and the diversity of people's values, as opposed to enforcing some normalized moral framework onto everybody.

Could it be that if humans are in situations where we're relying on A.I. to make moral decisions then we've already screwed up? Isn't morality something we probably shouldn't be outsourcing in the first place? You're touching on a common — sorry to be blunt — misunderstanding that people seem to have about the Delphi model we made. It's a Q. and A. model. We made it clear, we thought, that this is not for people to take moral advice from. This is more of a first step to test what A.I. can or cannot do. My primary motivation was that A.I. does need to learn moral decision-making in order to be able to interact with humans in

a safer and more respectful way. So that, for example, A.I. shouldn't suggest humans do dangerous things, especially children, or A.I. shouldn't generate statements that are potentially racist and sexist, or when somebody says the Holocaust never existed, A.I. shouldn't agree. It needs to understand human values broadly as opposed to just knowing whether a particular keyword tends to be associated with racism or not. A.I. should never be a universal authority of anything but rather be aware of diverse viewpoints that humans have, understand where they disagree and then be able to avoid the obviously bad cases.

Like

the Nick Bostrom paper clip example,[6] which I know is maybe alarmist. But is an example like that concerning? No, but that's why I am working on research like Delphi and social norms, because it *is* a concern if you deploy stupid A.I. to optimize for one thing. That's more of a human error than an A.I. error. But that's why human norms and values become important as background knowledge for A.I. Some people naïvely think if we teach A.I. "Don't kill people while maximizing paper-clip production," that will take care of it. But the machine might then kill all the plants. That's why it also needs common sense. It's common sense not to kill all the plants in order to preserve human lives; it's common sense not to go with extreme, degenerative solutions.

What about a lighter example, like A.I. and humor? Comedy is so much about the unexpected, and if A.I. mostly learns by analyzing previous examples, does that mean humor is going to be especially hard for it to understand? Some humor is very repetitive, and A.I. understands it. But, like, New Yorker cartoon captions? We have

a new paper about that.[7] Basically, even the fanciest A.I. today cannot really decipher what's going on in New Yorker captions.

To be fair, neither can a lot of people. [Laughs.] Yeah, that's true. We found, by the way, that we researchers sometimes don't understand these jokes in New Yorker captions. It's hard. But we'll keep researching.

Opening illustration: Source photograph from the John D. and Catherine T. MacArthur Foundation

This interview has been edited and condensed from two conversations.

David Marchese is a staff writer for the magazine and writes the Talk column.