Unconfuse Me with Bill Gates

EPISODE 05: Yejin Choi
Date aired: November 16, 2023

YEJIN CHOI: Current AI, the fact that it's so opaque, and nobody knows what's going on under the hood - that's just not healthy.

BILL GATES: I'm lucky to be able to talk to experts – getting people who've got different backgrounds, different perspectives – that stimulates my thinking. I feel privileged that I get access to people who can do that. I call that 'getting unconfused.'

[music]

Welcome to *Unconfuse Me*. I'm Bill Gates.

[music]

My guest today is Dr. Yejin Choi. She's a Computer Science Professor at the University of Washington, Senior Resource Manager at the Allen Institute for AI, and recipient of a MacArthur Fellowship. She does amazing work on AI training systems, including looking at natural language and common sense. She gave a great TED Talk this year, entitled, "Why AI is Incredibly Smart, and Shockingly Stupid."

Welcome, Yejin.

YEJIN CHOI: Thank you so much, Bill. I'm so excited to be here.

BILL GATES: Are you surprised at the advances that have come in the last several years?

YEJIN CHOI: Oh, yes, definitely. I didn't imagine it would become this impressive.

BILL GATES: What's strange to me, is that we create these models, but we don't really understand how the knowledge is encoded. To see what's in there, it's almost like a black box, although we see the innards, and so understanding why it does so well, or so poorly, we're still pretty naive.

YEJIN CHOI: One thing I'm really excited about is our lack of understanding on both types of intelligence, artificial and human intelligence. It really opens new intellectual problems. There's something odd about how these large language models, that we often call LLMs, acquire knowledge in such an opaque way. It can perform some tests extremely well, while surprising us with silly mistakes somewhere else.

BILL GATES: It's been interesting that, even when it makes mistakes, sometimes if you just change the prompt a little bit, then all of a sudden, even that boundary is somewhat fuzzy, as people play around.

YEJIN CHOI: Totally. Quote-unquote "prompt engineering" became a bit of a black art where some people say that you have to really motivate the transformers in the way that you motivate humans. One custom instruction that I found online was supposed to be about how you first tell LLM's "you are brilliant at reasoning, you really think carefully," then somehow the performance is better, which is quite fascinating. But I find two very divisive reactions to the different results that you can get from prompt engineering. On one side, there are people who tend to focus primarily on the success case. So long as there is one answer that is correct, it means the transformers, or LLMs, do know the correct answer; it's your fault that you didn't ask nicely enough. Whereas there is the other side, the people who tend to focus a lot more on the failure cases, therefore nothing works. Both are some sort of extremes. The answer may be somewhere in between, but this does reveal surprising aspects of this thing. Why? Why does it make these kinds of mistakes at all?

BILL GATES: We saw a dramatic improvement from the models the size of GPT-3 going up to the size of ChatGPT-4. I thought of 3 as kind of a funny toy, almost like a random sentence generator that I wrote 30 years ago. It was better than that, but I didn't see it as that useful. I was shocked that ChatGPT-4 used in the right way can be pretty powerful. If we go up in scale, say another factor of 10 or 20 above GPT-4, will that be a dramatic improvement, or a very modest improvement? I guess it's pretty unclear.

YEJIN CHOI: Good question, Bill. I honestly don't know what to think about it. There's uncertainty, is what I'm trying to say. I feel there's a high chance that we'll be surprised again, by an increase in capabilities. And then we will also be really surprised by some strange failure modes. More and more, I suspect that the evaluation will become harder, because people tend to have a bias towards believing the success case. We do have cognitive biases in the way that we interact with these machines. They are more likely to be adapted to those familiar cases, but then when you really start trusting it, it might betray you with unexpected failures. Interesting time, really.

BILL GATES: One domain that is almost counterintuitive that it's not as good at is mathematics. You almost have to laugh that something like a simple Sudoku puzzle is one of the things that it can't figure out, whereas even humans can do that.

YEJIN CHOI: Yes, it's like reasoning in general, that humans are capable of, that these ChatGPT are not as reliable right now. The reaction to that in the current scientific community, it's a bit divisive. On one hand, that people might believe that with more scale, the problems will all go away. Then there's the other camp who tend to believe that, wait a minute, there's a fundamental limit to it, and there should be better, different ways of doing it that are much more efficient. I tend to believe the latter. Anything that requires a symbolic reasoning can be a little bit brittle. Anything that requires a factual knowledge can be brittle. It's not a surprise when you actually look at the simple equation that we optimize for training these larger language models because, really, there's no reason why suddenly such capability should pop out.

BILL GATES: I wonder if the future architecture may have more of a self-understanding of reusing knowledge in a much richer way than just this forward-chaining set of multiplications.

YEJIN CHOI: Yes, right now the transformers, like GPT-4, can look at such a large amount of context. It's able to remember so many words as spoken just now. Whereas humans, you and I, we both have a very small working memory. The moment we hear new sentences from each other, we kind of forget exactly what you said earlier, but we remember the abstract of it. We have this amazing capability of abstracting away instantaneously and have such a small working memory, whereas right now GPT-4 has enormous working memory, so much bigger than us. But I think that's actually the bottleneck, in some sense, hurting the way that it's learning, because it's just relying on the patterns, a surface of patterns overlay, as opposed to trying to abstract away the true concepts underneath any text.

BILL GATES: One of the areas that the Gates Foundation would love to see is a kind of math tutor. There's a question, do you need a big, big, big, big model to do that? Because if you make it so big, then our ability to know how it behaves, it's hard to test. We're hoping that one of the more medium-size models that mostly learn math textbooks, and won't have such a broad knowledge of the world, we're hoping that that will let us do quality assurance.

YEJIN CHOI: In academia, there's actually a lot of such effort going on, but without a lot of compute. Including my own work that tries to develop a special model. Usually, the smaller models cannot win over ChatGPT in all dimensions, but if you have a target task, like a math tutoring, I do believe that definitely, not only you can close the gap with larger models, you can actually surpass the larger model's capability by specializing on it. This is totally doable, and I believe in it.

BILL GATES: Certainly for something like drug discovery, knowing English isn't necessary. It's kind of weird, these models are so big that very few people get to probe them or change them in some way. And yet, in the world of Computer Science, the majority of everything that was ever invented was invented in universities. To not have this in a form that people can play around with, and take a hundred different approaches to play around with, we have to find some way to fix that, to let universities be pushing these things, and looking inside these things.

YEJIN CHOI: I couldn't agree with you more. It cannot be very healthy to see this concentration of powers so that the major AI is only held by a few tech companies, and nobody knows what's going on under the hood. That's just not healthy. Especially when it is extremely likely that there is a moderate size solution that is open, that people can investigate and better understand and even better control, actually. Because if you open it, it's so much easier for you to adapt it into your custom use cases, compared to the current way of using GPT-4, which all that you can do is sort of a prompt engineering, and then hope that it understood what you meant. The math tutoring case seems to be the case, where the language models have seen a lot of educational material already out there online. So, that probably is, indeed, much more around the corner, because it has seen a lot of data. Whereas the drug discovery, now the challenge is for AI to come up with something new that doesn't exist yet. I suspect that that's a different type of a challenge for AI, because now it truly needs to reason more in a symbolic manner that is grounded in knowledge, as opposed to, 'oh, there's a bunch of the sequences, and let's predict what comes next and get lucky.' That's inspiring for me to think about, the different types of challenges and what it might take in order to push things to the next level.

I think that's basically the future. I am excited to see a lot more open-source effort, really catching up rapidly right now, the fact that it's just so opaque. Current learning is unbelievably brute force, which I don't think is the correct way of doing intelligence. There must be a better solution. And for that, we have to open it. In order to be able to really promote better science around it, we need to open it. We don't have to open the largest or best one, however, because even if you open it, it's not like academic people can do anything with it. If GPT-4 is open for me, there's no compute for me to run all those!

BILL GATES: I think to deal with the complexity and the accuracy you probably want to build these things from scratch.

YEJIN CHOI: I believe, with a bit more effort, something like that could be built. And with that wishful thought, I'm also working toward that sort of a system where we might have a little bit more explainable, descriptive knowledge that we can give to the machine to really, truly learn and memorize. Then when it does make mistakes, being able to control the machine through, 'Oh, what kind of knowledge are you assuming for that kind of answer?,' and being able to provide, 'Oh, you know, your assumption is wrong that way. From here on, learn this knowledge.' Those kind of problems unlock really exciting new types of machine learning problems, where you need to be able to unlearn, not just learn, but unlearn the incorrect knowledge, and then be able to revise over that in the way that humans also are able to. Whereas right now, everything, like you said, is a bit too black box. But I do think that with effort, that this sort of technology could happen.

BILL GATES: Someday maybe we'll understand how knowledge is represented in the human brain. It's one of the great mysteries of how evolution did that. Let's say we figure out both the software and the real brain, do you think we'll end up seeing that there are similar algorithms underlying how they work?

YEJIN CHOI: Oh, good question. What do you think?

BILL GATES: I think there are aspects, like visual recognition, where we can see that, as you go up, and you're trying to go to higher-level representations, that some of the same mistakes that the human visual system makes, weirdly appear in these systems. So that at least suggests that there's a common way. Evolution was sort of trying out different approaches. So it may be that there's this one fundamental approach that we see a glimpse of in software that evolution "discovered" and managed to use. It's the greatest miracle that humans' reasoning capability is so phenomenal.

YEJIN CHOI: Yes, totally. Evolution somehow figured out the algorithm behind our amazing learning capabilities, but we humans haven't figured out the AI version of it yet. I suspect that there's definitely a better algorithm out there that we haven't discovered. It's just right now, there's a bit too much focus on, 'let's make things larger,' and everybody's trying to do that. Whereas there may be really a better solution, an alternative solution that's waiting to be found, but there's just not enough of an attention there. Because people tend to think, 'Oh, it's not going to work.'

Let's go back to Microsoft and the very first personal computer, because when that first came out, it was really super-exciting and amazing. Then every single year, there's a better computer and smaller computer, a faster computer, and it becomes better and better. Similarly, when we look at phones, rockets, cars – the very first invention is never the optimal solution. There's always a better solution. I do think that there's a better solution, it's just that right now, there's way too much emphasis on the bigger the better.

BILL GATES: I do think, like in a math tutor case, though, the downside of a mistake can be pretty modest. And I think we are seeing that we should be able, give us two or three years, to create something there, that is pretty profound for engaging learners in a way that's motivating and at the right level for them. That'll be a pioneering test, that is not the same as relying on it for dangerous decisions.

YEJIN CHOI: I totally agree.

BILL GATES: Are you worried that things could go too fast, and almost have humans ignore the control and the misuse? The sense of purpose of humans, if we're sort of dumb, compared to the AI, I'm more worried about that now than I was a few years ago.

YEJIN CHOI: Even I get a bit of uneasy feeling, if hypothetically, suddenly, AGI does arrive and it's all around better than us. How are we supposed to think about that? Are they going to replace all of us and we just go vacation all the time? That sounds really boring. Although that thought experiment is quite interesting, even if that doesn't happen, I worry that AI is impacting human life a lot already. And it will do so even more in the coming years. It seems that, unless we put the right kinds of efforts, trying to understand where the limitations and capabilities are, and then try to develop both the policy but also other AI techniques that can better control this impact on humans – if we don't put in enough effort, this could be disastrous. If we're not ready for it, it could be very hard on us. I'm at least optimistic that more and more people worry about this, and then there's a lot more conversation going on, so I hope that it's a sign of people doing more actions around it. But yes, it's a concern.

BILL GATES: I thought that we would get the super-capable kind of blue-collar robots way before this reading and writing thing became at least somewhat possible. The inversion that we don't know how to pick parts out of a box, but we know how to rewrite the Pledge of Allegiance the way Donald Trump would write it. Those two tasks, the robot task I thought of as much easier, and so it would come first.

YEJIN CHOI: That's a really sharp observation, Bill, and there's actually a thing about it, which is Moravec's paradox, which is that the perceptual tasks that look seemingly easier for humans are actually much harder for AI, compared to say, a chess game, which is harder for us, which is actually easier for AI. In fact, that inversion happens in other ways as well. I'm currently proposing this thought, a generative AI paradox, where it might be that somehow generative capabilities are stronger than the understanding capabilities, which again, may be a little bit inversed version of how humans tend to be able to understand amazing novels, but we find it harder to write. And again, paintings we can appreciate without being able to generate those great paintings. Whereas right now, it looks as if these capabilities are a little bit reversed. Because when you look at DALL-E 2, DALL-E 3, it's able to generate amazing images, but then there's no amazing current AI that truly understands the image content in a way that surprises us. They are lagging behind, weirdly enough, so it might be that between generation and understanding capabilities, there's something interestingly reversed about it.

BILL GATES: But it's almost a paradox that in the near term, the risk is that we overuse it, like take advice from it, and it would be wrong. In the long run, maybe the fear is that it's too good. In your talk, you expressed that: because it's such a different kind of intelligence, it's both the "smartest" by some definitions, and the "dumbest," like in medical applications. My foundation would love to have the equivalent of a doctor for poor people who can never get access to that expertise. But how do we test that? How cautious do we need to be when we have a hard time characterizing what we've got here?

YEJIN CHOI: Part of me wonders whether that hypothetical, AGI-like capability, if it did exist, and if it's so good, can it actually answer some of the hardest questions that humanity faces like climate change? Again, some people disagree, what is it doing? And can AI really help answer those kinds of questions in such a satisfactory, such a high quality, reliable manner? If AGI really truly comes, I don't know. Is it actually going to be good enough for that kind of purpose? That relates to your wish about doctors. We somehow need to create these AI technologies that can benefit humanity better, but are they actually going to be super-reliable? How much of a gap will there be? I think that's very uncertain right now. We want to believe that it's around the corner, in some sense, especially those technologies that can be really beneficial for humanity.

BILL GATES: In my twenties, I definitely thought, like for language translation, that there would just be a set of processing steps. This is a noun, this is a verb, and that it would be an explicit piece of logic. When Google found that their logic approach, which was a pretty large team, hundreds, was just beaten by their neural net approach – that was the beginning of this mind-blowing thing. So yes, we are often naive, particularly about what it takes to match human capability.

YEJIN CHOI: I don't know for sure whether we're really around the corner, or we are just opening the can of a lot of curious, fundamental questions about intelligence, and it might turn out to be that it's a lot messier than we expected. It's a lot harder than we expected. Then building really reliable, trustworthy AI turns out to be harder than we thought. I'm not necessarily saying that that is truth, either. We just don't know how far or close we are.

BILL GATES: Do you see a problem where the commercial applications of this and the money going into it is a gold rush, even making the Internet gold rush seem modest? Would that possibly drain people out of academia, who are doing the important work, or do you see that happening somewhat?

YEJIN CHOI: Unfortunately, there's a leak from academia to industry. But actually, there's a bigger concern for me. Whether they're in industry or in academia, I do worry that a lot of people feel a bit hopeless, in the way that there's really strong messages dominating the field, which is that scale is all you need, and GPT-5, 6, 7, will be even more amazing. There's maybe nothing one can do about it. There's a bit too much currently shifted towards the prompt engineering as the main research focus. I genuinely worry about that, everybody doing the same thing, can that be good? I do hope that people explore what happens with the bigger scale out of curiosity, but the fact that there's so much emphasis, and all the companies, major companies, now they feel like they need to catch up with ChatGPT. I hear from many friends that there's a lot of this internal refocus, reprioritization, which is totally understandable, but if this is a global phenomenon, that's not healthy at all. We need to put more research effort around safeguarding AI and building alternative methods that are more compute efficient, and therefore also less carbon footprint.

BILL GATES: We need to bring math and maybe even physics people, but certainly math people. I feel lucky that I was a mathematician and then did computer science, because these models are very mathematical. Just being a programmer isn't really the training you need for this stuff.

YEJIN CHOI: And currently, brute force at scale is the way to go, but there may be an alternative, where sometimes these smaller models, the specialized models do learn on a lot more specialized data, and the data is actually the key. And that data can be not just more data, but it's better data, high quality data. Sometimes the data that was really designed to teach you that particular mathematical concept, for example. When you think about humans also, nobody learns very well just by reading random Web data. We tend to learn better when there's a great textbook and tutorial. Similarly, I do think that this is about how to transfer knowledge or information in the most efficient way. That's another reason, for me, why I believe that the smaller model or modest-size model could have a major edge. But that requires innovation about how to get that information, alternatively.

[music]

BILL GATES: I've got a turntable here and I asked you to bring in a record album.

YEJIN CHOI: This music, it's called "Virtual Insanity." Very relevant to our current conversation, but I used to listen to this when I used to work for you.

BILL GATES: Oh, wow.

YEJIN CHOI: Yes, here in Redmond. This was before I did a PhD. Before coming here, I was excited to learn about this Microsoft programming language package called the MFC. I don't know if it rings a bell to you.

BILL GATES: Sure, yes.

YEJIN CHOI: I self-taught that, because it wasn't really a part of the curriculum, per se.

[music – "Virtual Insanity" by Jamiroquai]

YEJIN CHOI: Somehow, I found the development job. I used to listen to this. The genre is like acid jazz, but it's not really jazz. It's like a modern variety of it, and I believe these are like maybe UK.

BILL GATES: "Virtual Insanity," wow.

YEJIN CHOI: Right now, it is virtual insanity. [*laughs*]

BILL GATES: It's kind of like jazz and rap. Next thing we know, we'll have AIs not only making the tunes, but the lyrics as well.

[music fades]

BILL GATES: What are some of the ways you're most enthused about that AI can help us improve the world?

YEJIN CHOI: My wishful thought is AI to really better understand humans more than humans ourselves do. I think that's fundamentally a reason why there's a lot of conflict. There's a lot of disagreement, and I'm hoping that we can use AI as a tool to better reflect about ourselves, and then be able to communicate to each other better, and coexist together more peacefully.

BILL GATES: I completely agree with that. It's kind of scary, that we seem to be more polarized. Other technologies gave us hydrogen bombs and bioterrorist pathogens. It's just a dream, because the AI is not there yet, but if it could help us understand each other and maybe reverse this trend towards polarization, that would be an incredible favor to the world. A lot of people worry about AI safety, that it doesn't take over the world, but at the same time, maybe it can improve and reduce conflict, and improve understanding. That's worth working on.

YEJIN CHOI: Yes.

BILL GATES: Well, thank you, Yejin, for taking time. It was a fascinating conversation, and it's going to be interesting to see where it all goes.

YEJIN CHOI: Likewise, thank you so much for having me here.

[music]

BILL GATES: *Unconfuse Me* is a production of The Gates Notes. Special thanks to my guest today, Yejin Choi.

YEJIN CHOI: To be honest, I never imagined to give a TED talk. I just don't have that kind of personality. But I got the arm twisted to do that, because basically, the recruiting person told me that otherwise, it's going to be just a lot of tech CEOs, who are also men.

BILL GATES: Ah! [*laughs*]

YEJIN CHOI: That was motivating enough. She clicked the right button on me.