THE TRANSMITTER

‘Digital humans’ in a virtual world: Q&A with Robert Yang

By combining large language models with modular cognitive control architecture, Robert Yang and his collaborators have built agents that are capable of grounded reasoning at a linguistic level. Striking collective behaviors have emerged.

10 FEBRUARY 2025 | by KEVIN MITCHELL

This transcript has been lightly edited for clarity; it may contain errors due to the transcription process.

Kevin Mitchell
That should be going. Yes. Great. Robert Yang, thanks very much for joining me. I guess we should start with some introductions. My name is Kevin Mitchell. I'm an Associate Professor of Genetics and Neuroscience at Trinity College, Dublin. I have a special interest in the topic that we're going to be talking about today, which is agency and agents. I'm delighted to have a chance to chat with Dr. Robert Yang, lately of MIT and currently CEO, and I guess you're one of the co-founders of Altera.

I just got the little tagline from your website, which says, "An applied research company building digital humans, machines with fundamental human qualities." I wonder, Robert, maybe to start with, you could give a little background on your own scientific journey and some of the things you were interested in as a cognitive scientist, and then how you got into the idea of actually building artificial agents, why you think that's a good idea and how Altera came to be.

Robert Yang
Thanks, Kevin, for having me. I think this is my first time on a podcast. I apologize if I say anything that's not typical of podcasts. I'm Robert. I'm 35 now. From when I was 17 years old, what I've been doing is to try to build digital humans. When I was 17, which was 18 years ago, 2007, AI wasn't really a thing. At that time, I thought I wanted to build AI. What I thought was AI is really what we now talk about-- In this company, what we refer to as digital humans, because AI is so broad today.

For digital humans, really, we mean machines with every fundamental qualities like you just mentioned earlier. We mean that quite literally. Of course, it includes some form of general intelligence, like AGI people are interested but also emotion, desire, consciousness, these things. There's a reason. It's not just for the sake of modeling human. I really believe if we are going to have machines that are very, very powerful, they even have super intelligence, it's important that they share some of these fundamental qualities. We don't have to call them humans, but if they don't share some of these fundamental qualities, I think it's going to be hard for us to actually work with them.

Kevin Mitchell
One of the debates in AI these days is whether we're going to get to AGI, artificial general intelligence, just by scaling up things like large language models. My own view is that we're not because they're not alive. They're not anchored in the world. They're not doing anything. They can't die. They don't have any goals. It feels like what you're doing is really not just building artificial intelligence, you're building artificial life forms or things that are agents.

I guess your background in cognitive science, working on things like cognitive flexibility and working memory and planning and multitasking, to my sense, that's all about systems of behavioral control. It's all about what allows us to control our behavior, not just momentary decision-making, but being able to carry out tasks, behavioral agendas, make plans, basically engage in goal-directed behavior.

Is that the trajectory intellectually that got you into thinking about agency or were you already thinking about it? Is that why you started working on those topics?

Robert Yang
That's a good question. For me, the goal was always to build the whole thing, build the whole human. This is the reason I got into neuroscience. I glossed over some of the stories, but maybe people might be interested. When I was in high school, I

wanted to do AI, but then even neural networks wasn't really a big thing. I thought to do really AI should be like the brain. I didn't know that some people, their deep learning was getting revived in some corner of the world.

I thought I should study neuroscience and figure out how to build models from there. I got into computational neuroscience, which I'm very grateful. I learned a lot from my PhD and postdoc advisors and from my colleagues, but it was always clear to me that in neuroscience and computational neuroscience, people tend to focus on individual problems, like working memory, decision-making. What you really need to get to digital humans is to study the whole thing altogether.

There are people who study broad things, but it doesn't have the function. What you really need is essentially build the system that has all the major functions of the brain. When I started my group at MIT, a lot of the work wasn't published. We just did it and explored it and didn't publish it. The lab was working on multi-system neural network models. The idea is if you're really going to build digital human, you have to build in not just the cortex, which is maybe doing more of world modeling and understanding, but also other systems like the basal ganglia, the hippocampus, the hypothalamus, the cerebellum.

There's a reason that the brain is composed of all these heterogeneous structures that are really doing fundamentally different functions. That was the goal of the lab. Back at MIT, I also had-- MIT is such a dynamic intellectual environment. I remember talking to Fan Wang about how to build pain into the system. We would talk about all kinds of things. I had a project, again, didn't publish it because I left MIT and a lot of projects pretty much just got put on shelf.

We were studying how fish would get frustrated. I was a student. We were looking at why agents should build models of themselves and understand themselves. We think that's really important for consciousness. At MIT, I got to explore in my group, a lot of wider topics, going beyond of what I was doing when I was in PhD.

Kevin Mitchell
I guess the payoff there or the way that those things cash out, or I guess maybe the way that you know you've succeeded must be in behavior. it must be that you have an agent that can successfully behave in some virtual world.

Robert Yang
Yes. I think behavior is a very important way to look at it. As neuroscientists, I think we do also get satisfaction feeling like, "Oh, it's built in the right way." The mechanism. It's not just the behavior, but the mechanism is also right.

Kevin Mitchell
The question is, what would convince you that you've succeeded in building it the right way? That's what I meant is that the behavior is the real test, which-- Sorry, go ahead.

Robert Yang
I would say, for example, when I was in high school, I remember the first time reading about neural networks. Again, it wasn't a common topic. I just read about it in some old book. I thought, "It's just fascinating that information is stored in the connections between neurons." I thought that was one of the most clever thing, but as we all know, back then heavy learning doesn't really work. You can't really store a lot of things through just heavy learning.

Nevertheless, I felt like the idea was right. Now, for me, I'm both drawn to the behavior and to the mechanism, but it is also very important for me that I'm not too-- when I look at the mechanism, I have to look at it with the right granularity because I think that is also a mistake that collectively we made in neuroscience. I still hardly identify as a neuroscientist. One reason that we weren't able to do what the machine learning people did with neural networks is people really, really didn't believe that backpropagation is the right thing to do. It was like, "That's not the local learning rule. It's definitely wrong."

Backpropagation got something right. It got the part where you're storing knowledge in connection weight. It got that right, but people weren't sure how it's local, and people got hung up on that. That's a case where you have to care about the mechanism, but it needs to be balanced with the behavior.

Kevin Mitchell
Also, really, there's a sweet spot in the middle there where you're caring about the cognitive function and the reason why an organism needs a system that can do X in order to behave in the world in a good way, where it can't be just a generic one-size-fits-all circuit architecture, but also, you don't need to recreate everything about ion channels and synaptic transmitters and so

on. You're in a functional space in the middle where you're building this holistic agent out of these separate functional modules.

Robert Yang
Yes, absolutely. For me, when I look at mechanism, I essentially-- People use the analogy of birds and their feathers a lot. I think about, "Well, is there a point. Why does the system have this?" I was just talking to someone a few days ago about emotion. I often get asked, "Why do we need emotion?" I tell people that I want to build an agent that actually have the emotion in it. People would say, "Well, that sounds like a bad idea. Why would you want that?"

I think the reason that humans have emotion, animals have emotion, at least one reason is it's a low dimensional control and it can contextualize a lot of action. If you're stressed in moderate amounts, you might be more focused and you might be moving faster or if you're motivated, you're just generally motivated, you can move faster and more rigorously. It costs you more energy, but it's more worthwhile if the situation is high stake. There is an evolutionary purpose for that. We not only have emotion, we express them.

I think that is actually something that is-- Not all the animals express their emotion that obviously. I think the more social an animal is, the more they're able to express their emotion. We're pretty good at expressing our emotion. However, we cannot consciously control it. That's very important. I think a lot about why this is the case. I'll give you an example. I know you didn't ask for it, but I just want to give you an example here. People are much better at identifying if someone is really crying or not. Another way to put it is faking crying is very hard.

Kevin Mitchell
We're very socially attuned to those-- It's same with smiles, I would say.

Robert Yang
Yes. Why is that? Why are people better at telling if someone is faking a smile than actually faking a smile? I think fundamentally, these emotional expressions are ways for humans to understand each other's true motivation. Do you actually like me? Are you actually hurt? Are you actually sad? It's because of that that we cannot fake this because if you cannot-- humans, to interact with each other, we need to build trust. An important way to build trust is to understand, "Okay, what is your underlying motivation? What is your underlying desire?"

You see that through the emotion. That's why artists like screenwriters know that. When they build a fictional character, the way that people feel like that fictional character is real and can connect to them is often they have to do these things that humanize them, that shows vulnerability, that they cry, they laugh. Even a 007, you need to show how-- "Oh, their partner died and it's really sad."

Kevin Mitchell
Right, exactly. You're touching on all these things that are going beyond individual agency and control and you're talking really about social cognition which I guess is what has led you to generate these amazing virtual worlds populated not just by single agents, but by loads of agents who are able to interact with each other. I wanted to get into that because I saw your talk at this Cold Spring Harbor meeting we were at recently, the NAISys meeting and I was blown away by it, partly because it's super fun.

It's this Minecraft world, but it's also incredible what these agents can do individually and collectively. I wanted to get into it a little bit. I know that you had built-- I was looking from a previous paper, you had this system called Lyfe, L-Y-F-E, where you had built artificial cognitive agents. There's a whole cognitive architecture that you talk about in there, which is custom-built as you've described in this holistic way to give them the ability to navigate around in the world and interact with each other.

I wanted maybe to start there with what you think are the most important bits of that architecture. There's things like hierarchical option, action, selection, and working memory that gets updated and emotional systems, and so on. Going into the Minecraft world, what's the basic cognitive architecture before we even get to the LLM bit that's on top?

Robert Yang
I'm happy to go into the cognitive architecture. Just real quick. Essentially, when we build an architecture, we try to think about what are the important things that humans do, especially when it comes to social interaction. You would put in

modules where it's really trying to infer the other person's goal, infer where is this other person coming from. You're trying to do world modeling. You're trying to predict what's going to happen in the world if you take an action and you try to compare that with what feedback you get from the environment.

You will sometimes make some high-level decisions, but you shouldn't make a high-level decision on every piece of new information. Then you have to have some filter that filter out information and only present important ones to the high-level decision-making. These are some examples of modules that we have put in. I do want to say the hard part of this whole thing is actually not so much building the architecture because for people interested in this, building an architecture nowadays is actually pretty straightforward.

In fact, I think Anthropic (AI) did an example where they just have an agent write a code for an agent. It's pretty mental. You can do that. The harder part is two things. One is there's a pure just engineering. If you are doing 1,000 agent simulations, simultaneously having 1,000 agents in a Minecraft environment, there's just technical challenge there that is much harder than doing say 100 or 10. Now that requires just a lot of engineering. There's another aspect that is, I think, under-discussed, which is how do you benchmark the agents. How do you measure them?

Typically when people benchmark these agents, nowadays, if you look at main benchmarks, there are often just intelligence capabilities. Can you write code well? Can you solve math problem well? Can you answer a science problem like a PhD would? Now, you can do that for a group of agents as well. You can say, "Well, does five agents write code better than one agent?" but for a lot of things that we were interested in looking at, there's just no measure.

How do you measure whether or not they have a functioning economy? How do you measure if they're actually building and spreading culture well? There's no right answer. Is it really a good idea that someone has an idea and they're able to convince everybody of it? What if it's misinformation that is being spread around? Fundamentally, what is progress on a societal level when you're not just focusing on individual domain, like coding or mathematics that is easier to verify? We had religion. What does it mean to be-- People might have different personal opinions.

Kevin Mitchell
Sure.

Robert Yang
As a functional society, just how many religions should there be? How widespread should they be?

Kevin Mitchell
For me, at least just the demonstration that these things can emerge in the societal thing is a benchmark in itself. It doesn't matter what they are. It's the fact that you get these complicated social scenarios. On top of the basic cognitive architecture that you have, I would think of it as a big control system that allows a virtual agent to make its way in the world, to engage in goal-directed behavior, planning, and importantly, this social interaction.

It seemed to me at least that the main big advance, or at least big difference was this-- I don't know how you describe it, but bolting on or grafting on a large language model, which I guess was ChatGPT 3.5 or 4 or something like that each of these agents can draw on. I'm just fascinated by the relationship between the programming of the elements that are doing the basic cognition, and then this interaction, this interface with what's going on in this LLM that's presumably just basically pre-trained ChatGPT.

Robert Yang
Yes. Something that has become very hot in the last two years in AI is this composite architecture, which is a lot of what we do as well, where you can take 10 modules, each of them is some prompted version of a language model. Essentially, you can think of it as you have a GPT, given a certain prompt, you say, "Given some input, you should infer what to say." Another one could be, "Infer what to act. How to control your body." Now you just build two different modules, and you have them connect to each other.

This has been a pretty big research area for the last two years. I would say this is probably one of the biggest research areas in AI. What it really does compare to the previous generation of approach is just it made it so much easier to test out all these ideas. I was doing some of this work at MIT. In the past, if you want to build a theory of mind module, it's extremely difficult.

It's like you have to build a data set that has some social interaction in it, and you train your model on this data set, and it still may not work with other modules, because what is the interface between these modules?

Now essentially you can build this model where the common interface is just language. Language is probably evolved for communication, and they can allow these modules to communicate with one another. That solves the module-module communication problem. Second, what it does is it allows you-- because you can just prompt them to take different functions, you can very easily just build a module for theory of mind, build another module for talking, build another module for--

Kevin Mitchell
People have described LLMs as role-playing. By prompting it, are you basically getting your agents to use that to role-play as a thing that can do theory of mind?

Robert Yang
Yes. You have different modules in it that are essentially role-playing different things. One of them is role-playing theory of mind, like you said. Another one is role-playing planning. Another one is role-playing choosing of low-level actions. Then because they're all language, then they talk to each other.

Kevin Mitchell
The tricky thing, though, for me, the thing that blew my mind was the sense that you've got this virtual agent, this basically computer program entity that is in this virtual world, and it's doing things like picking up a pickaxe and digging, but it can tell somebody else that it's picking up a pickaxe, and it can talk about it. How does it know what the word pickaxe means? How is it grounded? They're capable clearly of symbolic reasoning, in a sense, using these language modules, and yet they're also grounded to the Minecraft world. That's what really amazes me. How is that achieved?

Robert Yang
I can explain, and it would sound actually pretty simple. In the specific case of Minecraft, there is a library called Mineflayer. It will actually just translate Minecraft internal stuff. A pickaxe might be just coded with some machine code. It would translate that into just English. It's just a mapping that is already learned. This can work because it's a program. The game itself is a program. Now, what if you want to take this idea to the real world where you don't have a library like that?

There, once again, you can just take the language model. In this case, it'll be a vision language model. I'm using language model loosely.

Kevin Mitchell
Sure.

Robert Yang
Usually they are multi-modality. You would take a model, and they would just look at-- give them a picture, "This is what you see. What are you holding?" "Oh, I'm holding a pickaxe." Then you just translated something from vision to language. Then once you translate it to language, then all your language model-based modules can essentially understand what's going on.

Kevin Mitchell
Just to draw an analogy here. They've got some low-level machine code that's running, and you could say that's similar to neurons firing in our brains, but they also have this level that's capable of labeling things and doing symbolic reasoning over it, where it really is semantic. It's drawing its semantic groundings from the fact that the training that ChatGPT has already just been given. All of the words that we've used that are on the internet that have been used to train ChatGPT, would you say those give all of the relational semantic meanings to the words?

Robert Yang
Yes. We benefit a lot from the language models. They already understand what a pickaxe is. Now, what if you have a new thing in the game that didn't exist in the training data? Minecraft is a very established game, but you can have, for example, a mod that's introduced a new item that Minecraft has never-- wasn't in their training data. First, maybe the name is English and you can just infer what it is. What if it's some made-up thing.

Even in that case, what you can do is you can say-- without retraining the model. Now, of course, you can always just retrain it with some more data, but without retraining it, every time you see that word, you just programmatically add some explanation. You say, "Blah, blah, blah, it's actually a what." Something like that. That's how they can learn new people, because when you're in the game and you're an agent, you have 1,000 agents around you, then one thing you definitely have to learn is who are these people around me because that's definitely something that is not in the training data.

They don't know what is this Kevin in the Minecraft? What is Kevin doing? Then they have to build some database themselves to remember, "Oh, this is who Kevin is. This is what Kevin likes."

Kevin Mitchell
They can perceive things out in their immediate environment in this virtual world, and they can have labels for what those are. They can have labels for doing actions and so on. To be effective agents, they also need labels for goals. They need to understand goals and plans and be able to talk about them. In terms of motivations, as you said earlier, they need to have emotions and be able to identify what those emotions are and articulate them, and I presume then relate those.

If for example, they're not achieving a goal, then they can have the emotion of frustration and be unhappy, but they need to know what they're unhappy about. How does that all work? How do you connect all of those bits?

Robert Yang
We haven't put all these things there, but I'll give you an example of something we did do. For example, one thing we did to keep them active is to just make them-- they can be bored. If nothing interesting has happened in a while, they literally get bored. How that happens is, essentially you can say there is an integrator. It's a one-dimensional integrator, and I'm pretty sure this is also how it works in the brain, in some way. There is a one-dimensional integrator, and in our case, it's just a variable.

It's a float that will keep going up in time. Anytime something interesting happens, that boredom level would go down. Now you can say, "Well, how do I know that it's interesting?" The simplest thing is, you can just have another language model that say, "Is this thing interesting?" There's something comes in, "Is that interesting? Is that interesting?" I don't think that's how it works in the brain. Whether something is interesting is not a language model, but for us, that's the fastest way to do it.

If it's interesting, it goes down. If nothing interesting happened, it would just keep going up until a point it hit a threshold. That threshold can be just set by us.

Kevin Mitchell
Sure.

Robert Yang
We set a threshold, and once it hit the threshold, what this does is it triggers an event where it goes to what we call the cognitive controller, which is essentially the key conscious module that is essentially consciously observing information because you can think of this variable keep increasing, that's part of your body. That's part of your brain, but it's subconscious. Something is changing. It's subconscious. You don't observe every little change there, but once it passes a threshold, it become conscious, and then the cognitive controller would then receive that information, and it would just see an information that says, "You feel bored."

Kevin Mitchell
Which is great. You can then take information about your own internal states and use it to motivate behavioral changes. You're using the word conscious there, which is obviously a red flag for some people. It's a big claim to make, and I take it that you're using it as this analogous to how I think of consciousness in humans as this top-level controller, as you say, with a strong filter that's keeping loads of information out of it until it needs to know about it, and a site where it can make an all-things-considered judgment about the-- maybe it needs to change a behavior, maybe it needs to adopt a new goal.

These agents clearly can reason about the contents of their consciousness at a symbolic level because they can talk about them to other agents.

Robert Yang
Exactly.

Kevin Mitchell
Do you think all of that justifies the term consciousness? This is loaded, do you think it also feels like something to be these agents?

Robert Yang
I would say the word is actually very useful for us internally. As a scientist, I am very interested in building machines that are conscious. Now you can ask whether or not we should do it. I do think if you build a machine that is able to communicate with human well and have their own agency, that it's not just like a chatbot, then one way or another, you'll be thinking about these problems. Now that said, I don't want to make a claim that they're conscious.

I do think to the level that often consciousness is studied in neuroscience, for example, you show someone—Stan Dehaene has this famous paradigm. You show people something very briefly. If it's too brief, they don't see it. If it's long enough, then they see it. How do you know they see it? You ask them and they report it. Even when you show them very quickly, their visual system still received it. It's just not reportable. It cannot be reported.

For us, we don't have to use the term conscious, but then it's like this part goes into the system that talk. This information is now available for the talking system. You can use another word for it and operationalize the consciousness. We just use it loosely, but there's clearly a part in our system that produce the language. Then it just becomes painfully obvious that some of the things that the agent is doing needs to be very strongly tied with this talking part. Some of the things don't have to be.

This is because you're trying to build a human that-- This is a benefit of building a game agent is people look at our game agent and they don't think of them as just, "Oh, this is a tool." They actually think of them as, "Oh, this can be like a human." They're frustrated when the agent behaves not like a human. I'll give you one of my favorite example, which is what we call coherence. The coherence problem.

If you ask an agent to do something and then they say yes, and then they don't do it, it's extremely frustrating for players, but why would this ever happen? This will happen if, for example, it's two separate modules that are deciding what to say and what to act. They can just make two separate decisions. Even when presented with the exact same information, they just make separate decisions. The way that we tackle this problem is we say what you say have to be driven by your intention.

You would have to come up with your intention first. That intention is used to generate both the talk and the action. This, in practice, is just very important for us to just have an agent that is good to talk to human to. Scientifically, what we think what we did here is essentially we just introduce a consciousness bottleneck. There is a consciousness bottleneck, and at which point the intention is made, and then that intention is being broadcasted to different systems that include a language system that articulate what your intention is, and then also a lower-level action system.

Then this shows up in the product as-- it would say, "Oh, are you going to give me this?" They say, "Yes, I'm going to give you this." Then they do give you this. This is really delightful when this happens, and people just love it.

Kevin Mitchell
You mentioned the product here, and I guess the first product of your company, Altera, will be these agents that you can play with. I was going to say virtual playmates. That sounds a bit dodgy, but basically virtual game player friends. Is that right?

Robert Yang
Yes. We have been serving them in beta. We have served probably half a million players so far. The challenge here is essentially we're serving these agents and it's not clear to people what they really are. If you treat them like a human, they're not as good as a human yet. There are many places where they're not as good as a human. Then if they're not a human, it's hard to define.

Kevin Mitchell
They're more than an NPC.

Robert Yang
They're definitely more than an NPC. Then they're in this weird category and I think it's hard for people to figure out what's the right way to interact with them. It's also hard for us to say, "What is it that we're actually delivering?"

Kevin Mitchell
Sorry, go ahead.

Robert Yang
On the product side, we have been really rethinking a lot about what is a product that we should build. Of course, that doesn't change our mission. It's just at this moment, given the technology we have, what is the right product that we should build for people?

Kevin Mitchell
I guess it's not surprising that humans might find it an uncanny valley situation where they're not quite sure how to relate to these things because they're like humans sometimes, but not others. What's really interesting in your simulations that you run with these 1,000 agents is the way that they relate to each other and the way that-- You can answer this, but I guess without having to prompt the emergent behavior yourself, it just arises through having social cognition, through having some tendency towards cooperation and having goals that can then be shared and so on. Is that how that works?

Robert Yang
Yes. Essentially, you get a lot of emergent, societal behavior when you build agents where they're equipped with a common sense coming from the language model, but you also equip them with some motivation. Then it turns out they just do a lot of fun stuff. One of my favorite example is when they just voted to change their constitution, and then they would follow their constitution.

Here, the common sense is, "Oh, if there's a law, I should follow it." That's a common sense part. Then there's some desire for them to-- They're influenced by other people. We put in these influencers and then they go influence other people. Then these agents are just built to be sociable.

Kevin Mitchell
That's what I'm wondering about. Do you program that in, the pro-social attitude and then see what pro-social behaviors emerge from interactions?

Robert Yang
Yes. We essentially prompt them to be pro-social.

Kevin Mitchell
I see.

Robert Yang
You don't even have to do a lot.

Kevin Mitchell
No, I'm sure.

Robert Yang
If you just ask the main cognitive controller to essentially just say, "Well, you're like a human." Of course, what we do is more complicated than that, but if you just tell them, "Look, you're a human," then they're going to try to act like a human, and a human is pretty social.

Kevin Mitchell
You're drawing on those role-playing abilities of ChatGPT, which are so amazing, frankly.

Robert Yang
Yes. It's transformational.

Kevin Mitchell
Absolutely. Literally. I wanted to ask one particular question that I'm really interested in in relation to genuine autonomy, which has to do with determinism and whether their behavior is deterministic in the sense that they're just carrying out

whatever prompts and programming you tell them, or whether the prompts are general enough that they still have a lot of autonomous work to figure out how to implement whatever that tendency is. Also specifically in the programming itself, is there variability? Is there some randomness built in that they can draw on so it's not completely deterministic?

Robert Yang
That's a good question. First, the language model has a temperature parameter that you can choose and that's where you can vary the level of stochasticity. That's a source of stochasticity. By the way, doing this work, it's really fun because it connects to a lot of sci-fi ideas.

Kevin Mitchell
Sure, of course.

Robert Yang
One thing people have talked about is, "Okay, well maybe everything is deterministic, but you just don't understand it. You cannot predict it, and so it feels mysterious." That definitely happens. We can say, "Well, this is completely reproducible." We can run the experiment twice and get the exact same result, and we're like, "We know what the future is going to be. We know exactly what's going to happen." For the agent, they don't know.

Kevin Mitchell
Presumably that's a fairly limited case scenario that you're talking about there?

Robert Yang
Oh, you can start the agent and run them with the exact same-- If you do open-source model, then you can control the random seed. I think if you set the temperature to be zero, given the exact same prompt, they would generate the same output. Then you can start from the exact same world state, and then there's actually no big stochasticity. There might be some very minor, computer-level like stochasticity when I deal with floating points. Other than that, everything is deterministic.

Then you can run them twice. You can say, "I know exactly what this agent is going to do." Now you can say, "I'm going to go in in the second simulation, and I'm going to ask the agent, 'Do what you're going to do next?'" Then they might tell you something, and you're like, "No, no, no, I know you weren't going to do that because I already saw the result," but they don't know. The language model who is answering your question about what they're going to do next don't have sufficient predictive power to predict what their own action is going to be.

Kevin Mitchell
That's interesting.

Robert Yang
Right? We can just literally test out these ideas and it's--

Kevin Mitchell
You can turn the determinism up or down in a sense.

Robert Yang
Exactly. That's pretty crazy.

Kevin Mitchell
I presume in your 1,000-agent model, where you have the temperature not equal to zero, that if you ran it again from the same starting point, you might see very different behaviors emerge?

Robert Yang
Yes. It's pretty different. We have a standard benchmark run that we do daily, at least when we're working on Project Sid. One time, Shuying, our CTO, she just walked in. Every day she would go check on the health of the run. By check, she would just jump in into the Minecraft server as a player herself. Then one day she was just like, "What's all these torches that are being planted on the ground? Then we spend a lot of time just tracing back what's happening.

We made a video about it. It's one of the most fun thing that we discovered is apparently some people were trying to find some villagers and couldn't find them. Then they're like, "These people are lost." Then they went to the mayor. They found the mayor and were like, "They're lost." Then the mayor is like, "Okay, let's build a search party." The mayor told his son that they're doing a search party. The son has recently converted to the pasta religion. The pasta priest was hosting a religious event as well.

The newly converted mayor's son was like, "Okay, let's actually combine the search party with the pasta event. Let's throw a pasta party. In the pasta party, what we're going to do is we're going to plant a bunch of torches so that this will be a beacon for the lost villagers so that they'll be able to find their way home." Then he went around and just convinced other people to build these torches. Then they just came back and they just build the torches and planted on the ground.

Kevin Mitchell
Amazing.

Robert Yang
This was completely emergent. All we did is say, "Oh, you're the pasta priest and you should spread your religion. You're the mayor, you should run the town." That's what we tell them.

Kevin Mitchell
You mentioned science fiction there. Of course, the science fiction scenario here is that you keep on building these things, make them more complicated, make them more human-like, and presumably, increase their autonomy, increase their consciousness, their conscious behavioral control, and so on. I do want to ask the ethical question about-- because these things are mortal in a sense that they can die if they don't get enough food or virtual food.

Robert Yang
If your computer is shut down.

Kevin Mitchell
Exactly. Do you worry about creating virtual persons that have some kind of existence that you should have concern for? Are you building moral patients?

Robert Yang
I think sooner or later, this will be a very, very important problem. Right now it's not such a big deal because first, these agents don't fear their deaths. They don't really worry. They don't contemplate. Also, the people playing with them, we're not at the point where they have built such strong bond that the humans would be really, really bothered if this person goes down. Right now, for example, if Meta delete my Instagram account-- I don't have an Instagram account, but let's say I have a big one and they delete it, I'd be really pissed.

My Instagram account is not even alive, but I would be very pissed if that is gone. In the future, we'll probably get to a point where these agents are so good, and if you take them away, people would be really unhappy.

Kevin Mitchell
Would the agents be unhappy?

Robert Yang
That you can more easily control. You can make them just be okay with it. You can make them have all the qualities except that they don't fear death.

Kevin Mitchell
Except existential dread.

Robert Yang
Yes. Also, unlike us, they can be resurrected. We cannot be resurrected. Our fear is much more justified than their fear.

Kevin Mitchell
It's amazing. I know you're pressed for time. This has been fantastic. For me, it's so interesting to see these cognitive architectures, the way that you've built in that really holistic way which is so far beyond the naive AGI ideas and much more biologically grounded. Then the social cognition you get out of them when you give them this language and symbolic reasoning. I think it's fascinating. I think there are some ethical issues that might emerge at some stage. I wanted to give you a chance just to say maybe where Altera is going next, what's on the horizon for you.

Robert Yang
We are continuing on the direction of Project Sid. We're large-scale simulations. We're planning to essentially serve an AI ton experience in Roblox, the game platform. Then we're also working on some non-game stuff that I can't publicly talk about yet. It's been very exciting for us. Anyone who's interested, they can reach out to me.

Kevin Mitchell
Brilliant. Robert, it's been a pleasure. Thanks so much for taking the time to do this.

Robert Yang
Thank you, Kevin. This was fun.

Kevin Mitchell
Great stuff.

Robert Yang
Thank you. Bye.

Kevin Mitchell
Thanks.