MFS transcript 4

Continuity: How To Grow A Human, my Frankenstein Summer with Dr. Phillip Ball episode four, Helping It Learn.

Dr Philip Ball: It was a hot summer when I visited. And when it rains, it really rains still. I had work to do I headed up to Harvard to talk to a psychologist about how machines learn, because it seems that they really do learn and we don't know how to find out. We might need not just computer scientists. At psychologists in the 1933 Frankenstein movie, Boris Karloff as a creature was shown as having childlike intelligence, which was partly when coupled to his immense

strength made him so dangerous, but also what allowed us to feel some empathy for.

Passage reader: Um, get away with that.

Dr Philip Ball: I went to Boston to find out how new technologies in biomedicine and tissue engineering, genetics, and artificial intelligence might make it possible for us to create new human-like entities that think perhaps even live in ways like. In the last episode I was left pondering whether the best way to create a brain for such a creature might be through artificial intelligence and computer technology.

We're already going down that path, of course, with the AI of today, which is finding applications in all kinds of fields from advanced robotics and driverless vehicles to medical diagnosis and drug discovery. Some researchers hope that eventually AI might speed up the rate at which we can develop new medicines, like the vaccines we need to tackle.

COVID-19

some of these AI systems have superhuman capabilities, but only in a very limited sense. In other respects, they have less intelligence than a child or even a dog. And no one believes that they have any real sentience at all. What we don't have is machines that can truly reason, like. In fact, arguably they don't really reason at all.

I'm on my way to Harvard university to speak to psychologist, Tomer, Allman, who is using the tools of his trade to think about the possibility of making machines that think to understand what that could even mean. We need to have some understanding of what we're doing when we think, and in imperfect.

How we learn how to think. And as I discovered, there's plenty that we still don't know about that.

Tomer Ullman: There is this interest right now. Um, of course, the huge interest in building and trying to build machines that reason, like people are at least as intelligent as people, I should say. That's actually a different question between as intelligent, as people in reason, like people there's different paths there, right?

Some people think the way to do that is to resist. Anyway, but within that, there's, it's true that there's a lot of interest right now in building machines that do this in a childlike way. I don't think it's a new idea. It's true. That it's a resurging interest. Um, but this is certainly happened in the past.

Um, there's this quote that a lot of people like. Bring it around right now, which is, this goes back to before, you know, before the terms, artificial intelligence and cognitive development, or even around, or even cognitive science back, uh, Turing's Seminole paper on computing, machinery and intelligence, which everyone knows because that's where it introduced to the Turing test.

Right? Where we say, what is intelligence? What is a machine that can think, I don't know exactly, but let me introduce this test that, um, uh, if you can pass this, then we can say you're as intelligent as a human.

Dr Philip Ball: If you seen the movie blade runner, you know, about the Turing test, except look there, it's called the void comfy test.

And it appears right at the start of the film as the method used to distinguish between real humans and the artificial beings called replicants, which look in in many ways, act just like. The test was devised by the British mathematician, Alan Turing, and a paper in 1950, titled computing, machinery and intelligence.

He called it the imitation game. The way it works is that human question pose a series of questions to a human and an artificial intelligence, and tries to figure out from the answers, which is. It imagines machines that are able in our

interactions with humans to pass themselves off as human to and cheering, raise the implicit question.

If they can do this, do we have any right to deny that they can really think or to put it more provocatively, but we have to accept at least the possibility that a machine like this, rather than just being very good at imitating a person is actually conscious.

Tomer Ullman: But in that paper, he both introduces the idea, brings up some, um, potential objections to that idea, but he says he also speculates what would be a way to get a machine to pass this test.

And he says, well, it seems to me that we would want to build a child machine, right. We're not going to be able to build an old, the knowledge of the world. That would be very hard. And we don't know exactly how we're thinking, but maybe we can build a machine, but things like a child. And then we could teach it much as we teach a child.

So the idea is. One could maybe find other examples of literature or ideas of like machine children that you teach and things like that. But I think we can at least bring it back to Turing. And again, it's not just me, there's a lot of people saying that. And again, almost from a historical perspective, that idea of sort of had its ups and downs as much as I have its ups and downs.

Um, and I think it influenced AI, both influenced the way that we think about development and development influenced the way that we. About AI.

Dr Philip Ball: I asked Toma to explain how Alan tiering thought this machine learning would work.

Tomer Ullman: Um, if you look at the way the Turing talked about this machine child that he's proposing, he said, perhaps it's something like an empty notebook, much as you buy from the stationers.

And we fill it in with symbols and things like that, which is a very blank slate approach to the structure of the mind. Then it's also a very particular approach to the way that learning works. So it's very much like positive and negative reinforcement. Um, One of the things that touring I think helped to launch the Coleman of revolution, right?

The overthrow of behaviorism, the idea that we should think of the mind is more than just this input output mechanism that we're giving positive and negative input to. But we should think about as more like a computer. And it's okay to talk about the information processing that's going on in that computer about the programs that this mind might be running either if it's not exactly a Turing machine or this computer, but it is doing something like a computing.

That's the central in that. But that is sort of an overthrow of the story. Simple idea of learning of reinforcement learning and things like that, but it might be a little. Complicated than that. And I think it also gave this idea, this early idea to call them the development that, okay, maybe what development is, is learning different programs.

You start with this particular program, and then what you learn is you learn this new program and this program transformations happening again. This goes back to the seventies and things like that, but also in terms of just the capabilities of a computer, you might think of a development of development, along information processing lines.

Like you have more. Uh, you have faster computing speed, things like that. It's not exactly about having a better program or this radical transformation that you have program a moving into program B but more development is something like what you have, you know, better executive control that a computing power better.

You know, your bottlenecks are not quite the same and that can lead to improvements in many different ways. So they get the metaphor was. Useful, I think, or maybe not exactly metaphor. The influences I think were, were happening in both directions. And I think we're very

Dr Philip Ball: useful, but it's one thing for tiering to say that the way to make a smarter artificial intelligence is to build one that can learn like a child.

It's quite another thing to know how to do that after all, we don't really understand how a child learns. But one thing seems clear, a child is not just a blank slate with a head filled with squishy computing stuff, all ready to be filled up with knowledge.

Tomer Ullman: But again, there's this question of, okay, we want to build a machine that learns like a child.

What does that mean? And I've been to workshops where you have people who are sold on this idea from computer science, but don't necessarily know anything. Development. Um, so it's great that there's this enthusiasm, but it doesn't, it's not always accompanied with. Knowledge about cognitive development and they just speculate about like, well, you know, we want to build something buildings like that.

How do children learn? And they, you know, it's not free association time or, or they say things like, well, you know, we wanted to build a machine that learns like a child, you know, from scratch from nothing. And it's like, well, I, I don't think that's true. I think that there's a lot of arguments going around child development right now where even if you're, you know, you're very interested in building AI, you're very interested.

You're sold on this idea and not everyone is, but supposedly you're sold on this idea of building a machine that learns like a child and you say, okay, I'm so. And I'm willing to learn. Let's go to cognitive development people and see how the children learn. And what you'll end up with is probably a lot of arguments and people saying like, well, maybe it's like this.

And maybe it's like that. There is no one, um, dogma. There's no one agreed upon thing about how it is, how do children learn? Um, but I think that the questions that people are asking in development are very much the questions that people are asking in AI. You know, what's the state of knowledge that we start.

Um, for development, it's sort of empirically. We just want to know what's the state of law we start with for engineers. That's what's the state of knowledge we should start with. How much should we build it? And then, you know, for children again, it's how do we get. You know, what did we start with? How do we get more?

What's the learning algorithm. What's the learning process. What's the optimization function. And for engineers again, it's like what should be the optimization function? Um, with again, very fertile things I think, happening in both directions,

Dr Philip Ball: nearly a lot of debate and argument about how children learn.

But I asked Homer, if any clear themes seem to be emerging, that can be applied to how machines

Tomer Ullman: one example, for example, now a lot of people are interested in curiosity. Now it's a whole topic. This word is like being bandied about both in development and in AI. And you can formalize that in a way you can say, look, I have an AI, for example, that has a particular architecture and it has a loss function.

Oftentimes what you do. You notice you have some machine and you have a particular architecture and the sort of things that can learn, and you give it some sort of target, like a particular optimization technique or a particular learning rule or a loss function. And you say, for example, your loss function is to predict the next state of the world, right?

I'm not telling you exactly. It's something you need to distinguish cats from dogs. And I want it to be relatively unsupervised, right? What you need to do is to predict the next state of the world. Maybe that's how children do it. That's how I want my AI to do it. And then maybe we'll learn all sorts of things about the physics and psychology of the world and things like that.

Just from having two. The next frame from the given frame and it's incredibly. But it runs into a lot of stumbling blocks and roadblocks and things like that. And you can find yourself, you know, you know, I know exactly what's going to happen next for the next hundred years. Right. I'm staring at this wall.

Right. And you want to add in something else you need to add in on top of that, maybe some sort of curiosity, some notion of, I'm not just trying to predict the next state. I also need to be curious about wanting to predict for the next day, wanting to seek out the cases where I'm not actually predicting that well or.

And again, it's in some sense, an old idea and exploration, things like that. We own some exploration bonus, but people in developing them and saying this for a while, like children are curious, it's not enough that they want to just predict the next thing. There's something about the unknown, something about being curious, that's happening in addition to that.

So people in machine learning are now saying, okay, well, we can think of that as an addition to our loss function, to build machines that have this architecture and this loss function, but also this curiosity angle, something like that. And if

we do that, the now our machines are going to be better and all sorts of metrics and ways.

Dr Philip Ball: Toba went on to explain how much of, how we think seems to depend on the kinds of intuition that are hard to put into words, let alone to formalize and program into a computer. In particular, we tend to perceive and attribute goals and intentions to many of the behaviors we see

Tomer Ullman: in the same way that, you know, we look at a dog and we all know it's a dog and we agree it's a dog, but we don't know how we know it's adult.

There's some complex. Computation that's happening that we don't have direct access to, but we can study scientifically. There's a lot of judgments that we make that seem intuitive. That seem obvious that we say, of course they make sense that we don't have access to, like you say, oh yeah, bill is helping Jane.

It's like, how do you know, like how, what makes you say that judgment? And we could say a bunch of stuff about that, but it's about as elucidating and, you know, helpful as asking, how do you know it's a dog? And people say, well, maybe because of the ears and like, You don't know, you don't know about the convolutions that are happening.

And there was this one study in particular that I remember encountering the first year of my PhD and the seminar that was taught by Jeff Tannenbaum and Louis Velcade. She's one of the powerhouse of like a lot of the things that we're doing. We're doing it. But they brought out this Reese back then, it was a recent experiment of seeing, um, showing pre-verbal infants.

So you can't directly ask them, but you can measure their responses in different ways. Um, and you show them an agent, for example, climbing a hill, right. And failing them. So you need to imagine like a googly-eyed agent, like show you a video of this, but it goes like

and you get the sense as an adult that it's trying to get up there, but. And then one of two things happens. Uh, this is the protagonist let's say. And then one of two D Deuter agonists entered the scene like this, this big triangle. And it kind of pushes this agent up the hill and it gets to the top of the hill.

But like, or this big square come then of course your counterbalance, if it's a square triangle or whatever, the square comes in and picks up and pushes it to

the bottom line. And then you put it on the kid. They see this until we present the kids with an option. Do you want to perf you want to play with the triangle or you want to play with the square and the children overwhelming you choose the triangle or whatever it is that helped the agent help is a judgment that we, as an adults are making.

But the conclusion of that study that was published in nature is that children prefer helpers to hinder us. And I remember reading that. No on the second, I'm willing to buy this. I'm not arguing with the findings, but helping seems like this. How do we know that the children think that this is helping?

How do we know, how can we build a machine that would make that judgment? It seems like the children need to understand that this agent has. How do we know what a goal is? Just from the image. We need to maybe think that the other agents think that this agent has a goal and that they're trying to, what is helping in a model sense?

Like if I want to build an AI that has that understanding at the level of a 10 month old, how would you build that? And it was not at all clear to me how one would build that, but it turns out that there are models like that. Right? The translate that English sentence. Infants understand that other agents are helping one another, they prefer helpers to hinder us.

There are models that can try and capture that you can build into machines. And that's what I became very interested in saying, okay, if these are the right, and there were other people building these models, but if these are the right models, uh, is a true, how sophisticated are they? Where do they come from?

The children develop this sometime before 10 months. Is it earlier than that? Um, W are we born with this? And these are questions that, again, people in development have been asking for a long time in gathering data on. Part of the interest. And of course there's alternative accounts that are much simpler.

Like this is bad, this is good. Or when someone goes, that means that, you know, you associates in positive valence to that. And it's certainly the go-to multi-way thing for a lot of people that are, um, encounter this for the first time. Right. It's, it's easier to imagine, like a lot of simple visual perceptual features without having to say anything about mental state.

Goals and intentions and beliefs and things like that. Like this is bad, or if you're the shrinking distance between yourself and this thing indicates that that's

maybe a goal or things like that, you don't need to bring in a load of other fancier mental reasoning and things like that. But I don't think that that's the case.

I don't think that's true for what we would say, what we would call intuitive psychology. So the. Mental reasoning that would go into something like that scene I just described. I don't think that that's true for intuitive physics, um, which is, you know, just the fact that you know, that if I was at the wall more or less, what will happen, right.

You have a vague, maybe you can't predict at the pixel level, but you have. What would happen and even very, very young children to have a sense of like things shouldn't go through one another and they shouldn't disappear and appear and

Dr Philip Ball: things like that. So children develop these two capacities and intuition about physics or how objects behave and an intuition about psychology or how people behave that help them navigate the world and develop expectations about it.

They're really some of the basic building blocks of what we like to call common sense. Well, common. It may be, but it's frustratingly hard to build it into AI. How might we do that?

Tomer Ullman: There's a lot. We don't know about what children know. So I said it like, we now have this 20, 30 is the research. You can take it to the bank and things like that.

It's not the case. As soon as we tried to build these sort of models, you encounter a question. Well, here's a decision I need to make now in this model, um, what do children do here or what do they know? And people in the home, like, I don't know. That's a good question. Great. We should find out. Um, and that's been a very good experience in a way, because you're trying to build these models and then you go to people and development, and there's now this push happening at IBM and MIT and with work, um, at Harvard as well for trying to do.

The machines along these lines that would have common sense reasoning along the lines of that young infants have the young children have. Um, and I think that idea has caught on, but there are other people who are also interested in this idea and coming at it from a different direction. So for example, people at

Google deep mind, um, or Facebook, things like that, I would see that approach as being more, um, for example, over the last year, maybe even less than a year, they published a few different papers that are trying to pass these kinds of tasks.

Um, along the line, Young infants can pass, but they do it in a way that does not require what we would say are the representations that we think young children have. And when I say we, I don't mean all of development, like we in this particular group and we're open to being wrong, but we think that children do know something about agents and objects.

That's the interesting thing about what you were saying, because

Dr Philip Ball: this,

Tomer Ullman: it sounds as though this takes place at an earlier stage than people of

Dr Philip Ball: traditional. Thought

Tomer Ullman: of theory of mind evolves. So what's going on. Do they have

Dr Philip Ball: a sense of, there are other agents that do things yeah.

Tomer Ullman: And I can control. So this is we're going into argument land right now.

So, um, I want to be sure it's a minefield and I want to be sure I'm treading carefully. I have opinions and I'll give you my opinions, but I should stay there while they're based on evidence. There's evidence in different directions. Topic right now that people are arguing about. I think that one needs to distinguish between intuitive, psychology and theory of mind.

Um, and even within theory of mind, it's not an nothing or everything type thing. It's not that you don't have theory of mind. Suddenly you have theories much.

Dr Philip Ball: This idea called theory of mind is thought to be a key developmental milestone for. Loosely speaking it's when infants realized that other people or other beings in general have minds that are separate from theirs so that they might not know things that the child knows.

If the child sees you hide a marble inside one of three boxes while their mother is outside the room. And they realize when mum comes back in that she won't know which box the marble is in, then that child has developed a theory of mind that typically happens around the age of four. It's a crucial aspect of all human interaction, believing that other minds exist with motivations or knowledge and inclinations different from our own.

And if AI is going to communicate convincingly with us, it's fair to wonder whether it will need to develop a theory of mine to,

Tomer Ullman: but I think first of all, we need to, we need to distinguish that from intuitive psychology. So intuitive psychology is more like an ability or a phenomenon. It's hard to argue that.

I have that. It's more like it's, it's the thing that needs to be explained. So the very fact that you and I talk about other people's mental states is not objectionable, right? Um, the fact that we spend a lot of time doing it, the fact that we often agree, um, the fact that we can do it even for very impoverish stimuli, if you've ever seen the hider and similar experiments, people.

Dr Philip Ball: The hider and similar experiment is a classic in psychology. They were two psychologists who in 1944, made a short animated movie showing two triangles and a circle moving around, not a white background going in and out of it around a box with what looked like a door. That was it. Just abstract objects moving around.

But people who watched the animation were convinced that there was a story to it. The big triangle was bullying the little one and then threaten the circle. And then the little triangle helped the circle escape and so on. It was a classic example of how our minds construct narratives about what we see and deduce intentions, even in inanimate objects.

If you wanted to give a machine, a theory of mine, How might you give a bad shit?

Tomer Ullman: You might hear it as, uh, action understanding as in risk planning, you might hear it as in verse RL, you might hear it as Bayesian theory of mine. There's a bunch of different names or the name of utility calculus. Um, they captured different parts of a similar proposal with a proposal of something like the following.

So, um, suppose that I'm trying, I've read the literature and things like that. Fine. I'm trying to build up my machine now that can do this thing that we just talked about and see you take it out. Can see those little person climbing up the hill, trying to reason about their goals, beliefs, intentions, things like that.

How do I do that? Well, maybe I should think about the forward problem first. Um, I should think about if I had a goal, if I had to believe what actions should I take, people have solved that problem or not solve, but people have spent a lot of time on that problem and they've come up with many clever things.

This is planning. This is robotics, right? Robotics. We don't necessarily care about psychology or facing the forward problem. They say, I want to design a system, a machine, a robot, an entity. I want to give it a. Where the goal is specified as a utility function as a, some predicate that you need to fulfill something like that, some gold state.

And I want to give you the goal and I want to give you your beliefs in the sense of say a probability distribution over the state space of the world. And I need you to generate the set of actions. I don't want to have to specify the set of actions for you. I want to say robot gives me the cup of coffee on the robot.

No, don't do this. Right. And I don't want to have to say, you know, the actuators need to go exactly like that. That would be horrible and boring and brittle because if I then move this, then the robot will not do it. But that's the forward problems, you know, your input, belief and intention or goal or whatever outputs action.

Now, the inverse, I see. What is the set of goals and beliefs and intentions that generated that actions you conditioned on the action. What's the likely setting in this forward planning program. But that means that I need to have a planning person. For you, right? This is where it gets to be kind of like a theory, right?

I have this model of you that may or may not be right as something like a robot, like a planner. And if I input certain goals and beliefs, I have a prediction of what you'll do. And when I see what you do, I can then correct my program. But simulating you, that program may be very different from how you actually work.

It's maybe a useful approximation on the same way that I have an intuitive theory of physics. That is pretty good at predict. What will happen, but one shouldn't confuse that with the actual physics. Right? Um, my own sense of this

thing will what'll happen if I throw it at the wall, we could call it until there's certain computations and simulations, and then maybe even close to reality, but no one should say, but that's how things actually work.

There's an actual theory of physics there in a similar way. My theory of how you work, maybe. Maybe a good approximation, but there's a whole field out there of like how people actually make decisions, which seems to deviate from our theory, intuitive psychology. Right. So there's claims out there that maybe we shouldn't even think that people have beliefs or desires.

That's a useful. As we said, it's a theory. It's a useful approximate theory, but it's no more capturing the way that you work than Newtonian mechanics was actually capturing the way that reality works. That is all making sense.

Dr Philip Ball: Having a theory of mind, it's just one possible feature of an actual mind. And what I really wanted to find out from Toma is what he believed about the kinds of minds that AI systems might have.

What though is the measure of a model? How could we characterize and measure the feature that minds of any sword can have? What does the space of possible minds look like?

Tomer Ullman: Kind of zombie feel regret, can Goldfield regrets can a corporation fuel regrets can a human fuel regrets can a baby fuel regrets can a robot fuel regrets can have dolphin.

You get to the point to like for many of these entities where you might expect differences in mind, adult, child, human being. Smart animal non-smart animal gold angels robots rate them on all these things. And when you do basically a dimensionality reduction technique, where you try to say, okay, how many factors do I need to, to capture the variance in this thing you find that you need not one but two it's agency and experience, and you can find things all in all areas of this quarter.

Like human adults are high. Robots corporations and interestingly gold or high on agency, but low inexperience. Sure. Robots can plan, uh, robots can, uh, enact goals, robots, Microsoft can desire to achieve its goal in the next quarter, experiences a little more about emotion and things like that. There are many things that are high on experience of own agency, like human babies.

We're perfectly capable of saying they're experiencing the world in some really, you know, a valence way that we're happy to be happy to ascribe to them, maybe dolphins, but like. Agents. Right. And there's things that are local, both in everywhere.

Dr Philip Ball: This brings us back to the classic way of trying to deduce.

If there's a real sentient mind inside an AI, the Turing test Toma told me about an intriguing new wrinkle that he'd developed on that old idea,

Tomer Ullman: John McCoy. And I wrote this paper on, um, uh, minimal touring, which is, it's a bunch, it's about a bunch of things, but let me just give you the experiment. And then, um, And then the explanation for the experiment.

So experiment first, um, imagine that you were in the room with a smart rowboats and you need to distinguish yourself from the smart robot. There's a, it's a standard terrain test. The catch is that the judge does not have time for this nonsense. Um, and they're all going to have a conversation with you.

They want you to give one word to prove that you're human. They're going to ask the other contestant for one word to prove that the human, they can't see you. All that on the basis, that one word, the judge will decide. Who's the human, who's the robot, and we're going to kill the robot. So it's a high stakes situation.

Um, uh, you have to give a word from the standard English dictionary. Um, you both go to the word at the same time. The judge is smart and fair. They want the human to win. There's no catches in that sense. The robot is smart and wants to live, whatever that means. Um, so what word do you give you? She walked in expecting it to think about that, but.

Dr Philip Ball: Yeah. I mean, I think what immediately came to mind was emotive words like love.

Tomer Ullman: Very good. Um, do I survive? Well, it's interesting. So people, we asked a lot of people, this question, um, thousand something people, and there were a bunch of interesting things there. First of all, a lot of people like people did not say we didn't get a thousand words.

So we got a bunch of words that you can basically. Into kind of the same cluster. So a bunch of emotive words, anger and fear. Some of them, just the word

emotion, like the majority of not the majority, but the plurality, like the, by far, the most popular word was love. Um, next on the list was human and then you get a bunch of words, for example, that have to do with faith gold, Jesus mercy.

Um, you have a bunch of words that have to do with bodies' liver, bile brain, things like that. Do a bunch of us love to do with family. Now, where does that fall on experience or agency? I don't, I don't know. It doesn't really fall on that, nor would you necessarily expect it to, but you could think of this as it's not just about robots.

It's about there's this other agent and I'm trying to distinguish myself from this agent. So I need to think about like, what are the characteristics of this inner feeling of like this dominion, a demand like this high dimensional space. Agent and myself, both reside. I'm trying to now pick out some properties or words that pick out some properties that are close to me.

And not that agent, but also not so obvious that everyone would just guess this thing. Right. So you can imagine running this testimony, other things, right? Like how would you distinguish yourself from, you know, conservative and liberals or old and young or men and women the way. Turning to us, right? Well, inspiration for the Turing test.

And you can use that to basically try and think about what are people's intuitive theories, because in order to pass that you need to think about like, okay, what do I know about the robot? What would I expect the judge to know about the robot? Right. In order to pass this task, you need to have sort of an intuitive theory of the robot mind in a way as something that's maybe not about faith, not about emotion.

Um, these things, we ran full up to that where we had people play with judge and we told them like, here are, we took words that people actually said, oh, I didn't mention it. There was a whole category of like taboo words. Um, um, I'm not going to start swearing, but you can imagine. Right. Um, and if you also, interestingly food and especially greasy foods like hamburger, pizza cake, things like that.

Um, so. Side note when we asked. So we asked this online and we force people and we made sure that it was like, had to be a word from the English dictionary. When I ask people in person, they often try to break the rules as a way to like, I would go, or I would give two words or like, there's a way to write, but we, then

we asked people, we said, okay, we can't run all the words the people gave, but we're going to choose representatives.

So if we cluster the words and we say, okay, like here, all you know, here are the emotion words. Let's pick one hero what's left to do with Baldy. Um, I'm hoping I'm getting some of the details, right. But like, you know, one food word, when things like that, we chose 10 clusters, 10 words, and we gave to the judges combinations of these words, like you're, you're in front of, uh, these two contestants contestants.

They said, love contestant bees alive on them, said, you know which one is the robot? Which one is the human. And we can basically rank so we can rank the word. And you might think if this is all just sort of random, not, you might get cycles and things like a B2B B2C, it's a, but you've got actually a hierarchical graph.

If you just look at that, I should say at the bottom of this. So some people, my wife included said robot reasoning, but no robot would say robot. Um, but the judges think that's a particularly bad. So that's at the bottom of this list. So it's beaten only by human. So despite the fact that human was the second most popular one.

It's not a good word to say, which, you know, suggesting that people are not so running the right simulation of what the judges would say or things like that love is a good word in that. It's a lot of people say it and judges rank it pretty highly. Um, it's beaten by the taboo word. Um, but which is maybe not surprising.

I mean, there's a lot of things that, um, we don't necessarily generate the best solution, but once the solution is presented to us, we think, oh yeah, The best solution. Um, I don't know how much it would generalize to old type of words or what's going on the rights. I'm not, I'm not making hard claims

Dr Philip Ball: on this.

I love this experiment. It's so simple. Anyone can understand the rules and play the game. What I tweeted it after our conversation. It got a huge response. And just as tumor said, many people tried to break the rules or talk their way out of it. Sun even seemed offended at the thought of being put in such a stressful and perilous position.

And some of them swore, which I guess means they survive. But there was another aspect to the question of AI cognition that I wanted to explore with Toma. In the last episode, neuroscientists, Alan Jassen off explained just how important it is that our brains are a part of our bodies. Not some abstract disembark need intelligence doing computation in a.

How much might that matter for artificial intelligence, would it need to have a body to truly think like a human, uh, another aspect of this question that I have for various reasons we're interested

in is the, is

the function of embodiment in all of this, all the models that we run, one. Reflect the fact that we know that we have physical agency in the world that we can affect firms.

And we, of course, you know, this is clearly important for understanding animals because they have different kinds of embodiment, but also they are AI. So what does that fit in, in all these cases?

Tomer Ullman: I have opinions and I can give you my opinions. And some of these opinions are like, I think that they're pretty well based in like engineering and, and experiments and things like this.

I will say, I think it's important that, that we do have bodies. I think that being in different bodies will probably have an effect. I think that one can still remain. And the AI running a program that's successful and useful and things like that without necessarily having a balding. Cause it has a, uh, a physics cause it has a physics program and it has at least, and one could think of, for example, when I see you right now, sitting in that chair, I can recognize that you were an agent that you're in as a year human, like.

One operation that we can think about as calling a D rendering or the animating when I'm reconstructing from the, the pixel input, let's call it right? Like my eyes see pixels, but I'm reconstructing from that fact that you're a person sitting in that chair. And you have particular intent, all that stuff that I'm constructing like an internal representation of this, this thing.

What does that internal representation? Does it just include some simple features, like the fact that you have eyes, therefore you were an agent or things like that. Probably not. It probably includes a whole bunch about you. And I

think part of that representation is also maybe some rigid, roughly rigid Baldy representation, you know, and again, computer games, for example, if you're trying to animate something moving around, right.

If you think about this is a very useful metaphor. I think if you see some particular image in a computer game, how did we generate that image as engineers? We don't care about psychology. How do we generate that? Well, we have some understanding of. And objects. And one of the things that we need to use is something like a graph to capture the body.

Right? If we now want to move this memory note around, it's very useful to imagine that we have this representation with this graph knows the tools, how these things can work in the same way that we, we talked about, you know, the inverse of seeing actions and having to invert it and think about the actions that led to it.

I could see an image and think about the invert. Though, like the body representation that led to this particular thing. And I think that that's a very useful representation. I think that if you try to capture a certain actions and you try to say, for example, you know, these days it's very popular to show machines, for example, a thousand examples of people.

And saying, sitting, sitting, sitting side, standing, standing, sending students in extended now, and they will pick up on some visual features like the presence of the chair, not even a chair, like certain visual features that they can use to classify that this is sitting, but you can endow them with other representations that they may not pick up on their own.

If you just use a very simple, let's say bottom up algorithm, like say maybe there's a skeleton here at skeleton. Mr. Richard Baldy representation. And if you have that mid-level representation, you can now classify things without a thousand examples, with three examples of four examples. And that representation is much less about this particular pixel patch or this particular visual feature, but about the fact that the skeleton is in this particular pose, right?

That's that's I think where the classification is happening. And I think the children are much more like that when they learn. So they don't need a bazillion examples. Four or five or something like that. Um, but I think that's pointing to

some mid-level representations and within the middle of a presentation, I think representations of bodies are important.

But again, you could imagine some deep bodied, this embodied people, this embodied AI that still has a representation of bodies. Um, you might say it wouldn't come to that. If it didn't have one on its own, I don't know, but one could at least think it possible.

Dr Philip Ball: Um, well, 10 minutes being fantastic.

I felt like I got a good sense from Toba of what an AI might need. If it was going to get anywhere close to doing some of the things, our brains can a sense of intuitive physics of how the world works, how objects behave, how things move around. And a sense of intuitive psychology. The notion that other beings have goals, intentions, and desires that drive their actions, a model of how another mind might work, an AI might acquire, or at least refine these capabilities through experience.

So now the question is, can we make a machine intelligence like this? How far away are we from achieving it? That's what I'll explore in the next episode. When I go to find out what is going on in the AI labs of the computer giant, IBM,

Continuity: How To Grow A Human: my Frankenstein Summer is written and presented by Dr. Philip Ball and directed and edited by Keith English. This show is brought to you by Aurra Studios. Listen to the full series on Apple podcasts or wherever you get your podcasts.