A New Test to Measure Understanding in AI Models

https://mindmatters.ai/podcast/ep361

Robert J. Marks:

Greetings and welcome to Mind Matters News. We're talking to Dr. Georgios Mappouras about his new and interesting paper Turing Test 2.0, The General Intelligence Threshold, and we're going to continue our talking about this. The idea is whether or not AI is intelligent and the classical definitions of this come from the Turing Test. Last time we talked about why the Turing Test is probably insufficient. So Dr. Mappouras has come up with a new method called the Turing Test 2.0 for Measuring Intelligence, and that's what we're talking about. And the paper is called The Turing Test 2.0. Again, The General Intelligence Threshold, and we haven't dug too deep in here. So let me start out. George, by the way, welcome. Welcome back. It's good to have you back. Could you outline the three rules that define whether a test qualifies under the Turing Test 2.0?

Georgios Mappouras:

Right, so in the paper I define three rules for a test to be a valid Turing Test 2.0, it's like a framework that can help you, you can use it to generate such tests, right? So there's three rules. The first rule is that the system will have, we will transfer to the system through training or through hard coding, some amount of functional information. And from what we talked about in the previous episode, it's a group of functionalities. So things that we expect the system to be able to do. The second thing is that we have to give to the system information that maybe is non-functional. What do we mean by that? The system doesn't know yet how to use it.

Robert J. Marks:

This is what we call before the flash of genius. It has to come up with something that it hasn't been trained on.

Georgios Mappouras:

Exactly. Yeah. So some information that is not functional for the system, we have to know that the system doesn't know how to extract the functional, has not yet extracted the functionality from that information. So we have two sets. First rule, a set of functional information, it has some functionality. Second rule, a set of non-functional information. And then the third rule is that no other external training, external functionality has been imported to the system. So after that, we can, let's say, try to interact with the system and try to see if it can come with new functionality based on the information we gave it. To use example we used in the previous episode about the anecdote about, "Okay, we have the apple falling from the tree, you can extract Newton's theory out of it."

So the idea will be I give you enough information that we know it's enough to extract the new theory, the new functionality. Can you showcase that you can do that? And it is not just showcase randomly that maybe you spit out some random thing that happened to be true, but after that point you have to show that you can always reuse this knowledge. So an example that is easy to understand is that you might have, let's say an array and you might know how to find an element in that array. And then you ask the question, what if the array is sorted? And the system has to come with a binary search where it's a more efficient way to search an array if you know that it's sorted to find a specific element if it's inside the array.

And then if I ask it later, the next day or a year after, it has to know to reuse that knowledge, that knowledge is not lost, but it's able to reuse it and say, "Oh yeah, if the array is sorted, I can do it more efficiently. I learned that I can always apply this functionality from now on." So it's not a functionality that will showcase one randomly, but you have to consistently be able to show that, "Yes, I learned this new functionality and now I can use it."

Robert J. Marks:

So again, I kind of equate the idea of non-functional information with a flash of genius. And I think that that's how we can tell that human beings are creative because we've all had flash as a genius. We get a solution which pops into our head and we're not sure where it came from. And history is replete with people that have had flashes of genius. I think one was Tesla for the brushless motor. He said he's walking along the beach and he said like a flash of lightning, it came to him and he brushed off some stuff or brushed off some dirt and actually drew a schematic in the dirt the same one that he used when he published the paper. Friedrich Gauss said he woke up in the morning and said, yeah, the solution to this problem had been working on. And he said he had no idea where I came from.

And it also comes from the arts. People have flashes of genius, for example, of writing music. You hear people, for example, saying, "I don't know where I came from. I'm too afraid to examine it, because I'm afraid it might go away." So we all have these flashes of genius, this non-functional information which humans have access to. I think it is a key marker of intelligence. You mentioned last time, which I think was very interesting and true, that you see no path to achieving this goal. In fact, the attempts that have been made, AI writing better, AI writing better AI, writing better AI, I think it's been debunked and I don't think it's conclusive, but I think there's just lots of evidence. There's something called model collapse wherein that if you use AI to write better AI pretty soon, yeah, you're not going to get new information that it kind of becomes a blubbering idiot, but that's the source of another talk sometimes.

In your experiments, you gave a bunch of examples, why did you choose the task like drawing a clock at I think it was 6:30 or generating a hexagonal stop sign? By the way, when I took my driver's test, I missed that on the test it says, "How many signs do a stop sign have?" And I missed that. This was-

Georgios Mappouras:

Oh, it was an actual question.

Robert J. Marks:

It was. It was back when I was 16 and I missed it. It was the only one that I missed on the thing. I thought, "Well, it's a regular polygon, but I'm not sure," so I guessed wrong. So anyway, tell me about the 6:30 and the hexagonal stop sign and what does it reveal about large language models?

Georgios Mappouras:

Right, right. So this all again came kind of as almost a coincidence. In the previous episode we talked about the Chinese Room argument and how it shows that the person inside does not have real knowledge of what he's reading. And this came kind of almost like an accident. I was watching the videos explaining what is the Chinese Room argument, counter arguments. Then I see a random video where there's a person talking about a problem AI has, that it cannot draw accurately the time and the interpretation wise because the images on the internet, which typically is what AI uses as training data, mass data from the internet tend to have a specific picture of clocks or watches that point to what we call the ten two. So the clock hands typically are in the ten two position. And the reason they do that is

for cosmetic reasons. They assume that, not assume. It is considered that this is the best image if you want to showcase a clock or a watch.

Robert J. Marks:

Oh, that's right. You go into a clock store and if the clocks aren't running, they all have their hands at the same place.

Georgios Mappouras:

Exactly, yeah. And they are talking about in a funny way, "Oh this is funny. All the images are like that, AI cannot produce other images easily." And I was like, "Wait a second. That's interesting." So they were actually being very, let's say, very casual about it, but I was like, "This is very interesting why it cannot, this is something very simple." So then looking at what we already discussed, I said, "Wait a second, does AI know how a clock works?" Because if you know how a clock works, that's enough information to draw any picture. So then I went back to the AI and I followed the Turing Test route, the Turing Test 2.0 rules, and I said, "Okay, can you draw images? Yes, you can draw images." So this is the functionality already has.

"Can you draw images of a clock?" It can't. It gives you very nice images of clocks. It already has this functionality. This is the number one, my first rule for a Turing Test 2.0. So then I looked at non-functional information, meaning is that information that it's enough if you really understand it, to draw any clock?

Robert J. Marks:

Oh, that's a great example, yes.

Georgios Mappouras:

So I asked the AI, "How does a clock work?" They give a perfect description of how the clock works within minutes, seconds, everything. And even you can ask it, "Can you tell me where the hands of the clock are going to be if it's 6:30?" And they will describe it in detail accurately. They'll tell you, "Oh, the hour should point at six, a little bit past six actually." They're very accurate, "A little bit past six. And the minute hand will point at six because it's six times five instead." So I was like, "All right, you have all the information you need. You pretty much gave me the image, but in a text form. So then I'll ask you to generate that image that I described," and it fails. It gives you a random, 10 10 or something close by. And that shows that it doesn't have understanding, that it cannot extract the meaning of the things it's giving you. So if it doesn't have a meaning, how can you get functionality out of it?

The third rule, now this is where the third rule can break. The third rule says of the Turing Test 2.0 says there should be no external other external functionality come from. One can think that this problem is kind simple actually to fix. We just have to balance our dataset, the training set, give you more images of different times, label them correctly, train the model and fix this problem. The problem is easy to fix, yes, but then you're not testing for intelligence anymore because you broke the third rule. You gave it external information. So that's why it's a good test it's because of how the nature of the data they're used for these large models, they go to the internet and get what is out there pretty much.

Robert J. Marks:

I like this very much. Here's an idea that popped into my mind if all the literature in the world said that the world was flat, but it also knew things like, "Well, the North Star changes as you travel, and that

when ships disappear in the distance that they disappear from the bottom to the top." And it could take that information and say, "This idea of a flat earth, it doesn't line up with these sort of things." So it's that sort of correlation that will generate this non-functional information. So tell me about the hexagonal stop sign test that you did?

Georgios Mappouras:

Right. So after I noticed the behavior with 6:30 with the clock, I wanted to make sure first that this is not just something random I stumbled upon, but I can actually reproduce it with other images. So what was unique about the 6:30 that we have an image that is very, let's say, dominant in the data set. So I asked myself what other images we have that are dominant like that? And I started thinking about cases. Actually in the paper I give other examples too that I didn't really present in the paper, but I'm like, "You can try them out, if you want." And one of them that came to my mind was like the stop sign. So if you look at a stop sign, it has a very specific image, it has an octagon and it has a very specific, you have a red sign, white letters.

So I'm like, okay, the AI has seen that image many, many times. It is likely it hasn't seen some specific shapes. For example, what if it's a hexagonal? What if I wanted to do the same exact thing but this time in a hexagonal shape? And I noticed that AI had big difficulty doing that. When I tested different models, ChatGPT if you insist a lot, it was able to produce an image, but then if you reserve your chat, it would fail again and you have to try hard to make it produce the same image again, the correct image again. So it's what I said before is that if you really want to pass the Turing Test 2.0, you have to see that if you extract some knowledge, that knowledge doesn't go away, but you are able to apply it again. And there are other things like that that you can play with. For example, a triangular flat screen, right?

Robert J. Marks:

Oh, yes.

Georgios Mappouras:

Because flat screens typically have a specific shape or a driver license, not a license, a car license plate in weird shapes. So sometimes depending, because for example, you can think that, "Oh, there are different signs in different shapes." AI is very good at, let's say correlation, getting one thing, applying the other thing on top of it. So there is a niche there that you can start looking at these images that are very, very unique.

Robert J. Marks:

So this is very interesting because as a watcher of sitcoms was either on The Office or Arrested Development where they came up with this, I think it was The Office. They came up with this idea for a triangular screen. So I'm wondering if that exists, but if it did exist, certainly that the large language models didn't catch up on that. My experience, George, is that a lot of these large language models go and they put band-aids on things and stuff that didn't work last year, kind of worked this year. A classic example, I was informed and I checked it out and it was true, and it was Tom's mother has three kids, Snap, Crackle, and... Now if you know about Rice Krispies, their logo is Snap, Crackle and Pop. So yeah, Pop was the last one. Tom's mother had three kids, Snap, Crackle, and... Well of course it was Tom's mother, so it was Tom. It should have been Snap, Crackle and Tom. So this was something that was answered incorrectly a long time ago, but it was corrected.

Somebody went in and I don't think it was training. I think it was a human being that went in and corrected that. I suspect that if they do identify the problems that you're talking about, people at

OpenAI or Grok are going to come in and they're going to put a Band-Aid on this and fix it, but it's going to be a human intervention. It isn't an epiphany of the AI, is it?

Georgios Mappouras:

Exactly. This breaks actually the third rule of the Turing Test 2.0 because what will happen is that external sources, in this case humans, will come and give this functionality to AI. And that's why it's important to preserve the third rule. And that's why this test, yeah, they're nice, but very likely they'll go away as we add more functionality on AI systems.

Robert J. Marks:

Yeah. And I think a lot of that is being done by human beings behind the scene.

Georgios Mappouras:

Exactly.

Robert J. Marks:

They go in and they tune all the errors. In fact, if you go to some of these large language models and they make a mistake, you can say, "You made a mistake, blah, blah, blah." And they will come back and they'll go, "Oh, you're exactly right." And they'll fix it and the next time you won't see that same mistake again. So it is apparently self-correcting. I hope that it verifies it before it corrects, but it is self-verifying. And also I think that humans are that way.

One of the things that I tried a long time ago is to tell ChatGPT not to do something. I said, "Draw a picture of Times Square with no pink elephants or no pink rhinoceroses," I think it was. And I says, "No pink rhinoceroses. It didn't want..." It did not know how to not do that. So give me a picture of Times Square with a rhinoceros in it, a pink rhinoceros.

And I tried it on a bunch of other things, but they have fixed that then. I retried it, and I don't know if it came to the attention of the people that do these images, but I tried it just a few months ago and I said, "No pink rhinoceroses." And it recognized it. So somebody behind the scenes is dinking with this stuff, making it and more and more accessible. So that's interesting. So let me ask you about this. You talk about, I think it's, Chollet, C-H-O-L-L-E-T. How would that be pronounced? Do you know? Chollet's definition of intelligence? Francois Chollet?

Georgios Mappouras:

Oh yeah, I don't know how to pronounce his name.

Robert J. Marks:

Okay, let's go-

Georgios Mappouras:

It think it's French.

Robert J. Marks:

Let's go with American.

Georgios Mappouras:

I would say Chollet. Yeah, but I don't know.

Robert J. Marks:

Chollet. Chollet, okay.

Georgios Mappouras:

I don't know, so yeah.

Robert J. Marks:

I'm in Texas. We have this road here called P-O-G-U-E, which of course is French and it should be Pogue, but everybody calls it Pogue. It's Pogue Road. And there's another one, B-O-S-Q-E, which should be Bosqe I believe, but everybody calls it Bosqe.

Georgios Mappouras:

I'm the last person to judge about accents.

Robert J. Marks:

Okay, so we'll call it Chollet's definition, and we apologize to Francois. I got his first name, Francois Chollet about this. He was an employee of Google when he published this, and he proposed a definition of intelligence. And you talk about that in your paper. Could you unpack this? I found this as interesting, and I confess I had not heard of this definition of intelligence. I kind of like it.

Georgios Mappouras:

Yeah, yeah, yeah. Actually, when I first started digging around, I came across his paper and I found it very interesting. It actually has a whole benchmark, so it's more detailed as in it's a whole benchmark that you can actually apply to model and see how good it does. So the difference there is Chollet comes with a specific test, which a lot of people do, but his test's a little bit different than other researchers. And he says, "I have this type of test that are kind of unique," meaning it's kind like playing a game that you haven't really seen before.

And these games, if somebody Googles Chollet's work, they can go to the website, they can even play the games themselves. So it pops up on your web and you can actually do the test yourself to test yourself. And it's actually pretty much what I would call IQ tests. If anybody's familiar with IQ tests, they will probably get an idea of what I mean, where you have some patterns and then you have to predict the next one. So it gives you, but rather than being random numbers or whatever, let's say it's adjusted for computers. So you have pretty much, let's say a grid with each grid has some squares inside, and the squares have colors and positions in the grid. And it gives you, let's say three different variations where the squares in the grid from one variation to another, they change either in colors or in position.

And what you have to do is detect this pattern and predict the fourth grid. So that's why I'm saying it's like if anybody ever solved IQ tests, it is very similar. It's like IQ test adjusted for computers. That's the description I would give. And it's very nice because you test a computer if you can kind of reason in a way, that's how it feels like. And a lot of other people did other tests. The other tests that to be honest, why I like Chollet's is because other tests might be something like, for example, "Bill proposed, can you get inside a typical house and find the coffee machine and make coffee?" Okay, good. But why is that

intelligent, right? Chollet's has a little bit more close to intelligence because you're trying to reason and find something that needs you to what we call think, right?

What I don't like about this though, it's the major thing that I think it's common almost to all peer work in this field is that we compare in order to figure out if it's clever enough, if it reaches the general intelligence, we compare it to human results. And I'm like, "Okay, again, which human? The average human? Should we get the smartest?"

Robert J. Marks:

A guy that goes... Kindergarten kid, yeah.

Georgios Mappouras:

Exactly. And even a human that is not smart in this way might be smart in another way. I've seen it through my friends and family that somebody might be not as good in math, maybe very good at physics, maybe not good at physics, very good at painting, maybe not painting, maybe sports. And you can see the intelligence there, you can see how they innovate. So that's why I don't like this very specific, very tailored tests because, okay, it's a specific type of intelligence. And the other thing that, because I've taken tests like that before, IQ tests for different exams and stuff. If you take the test the first time, you're probably not going to do so well. The second time you're going to do better. The third time even better, Am I might becoming smarter or am I just training and learning the patterns? You see what I mean?

So if I'm learning the patterns, then sure, AI can do that. We know they can do that, but that's not what we're seeking to find if it can learn patterns. Yeah, AI has, the more you're training with this type of tests, it will do better eventually, probably better than humans. But yeah, I recognize patterns. No real intelligence, because if I test every time you take it, you can do a little bit better. Then it's just training. And this is my biggest, let's say objection to this test. It's a very nice test. But like the Turing Test, like the original Turing Test, I think it doesn't really measure intelligence. And that's my biggest, let's say, critique, is that why we test machines on things we are good at, right? And not the machine. For example, there was a time where the machines were not good at doing high precision math. They could only do a small precision. We actually used to use humans for high precision, right? Then they became better than us.

Robert J. Marks:

And they were known as computers, I think, right?

Georgios Mappouras:

Exactly. Yeah. So why don't we use that test? Computer is very good at it. What if I want the same action repeated precisely over and over again? Humans are terrible at it, machines are great at it. And that's exactly how I came with the Turing Test to 2.0. I was like, "In order to be able to get with a test, I have first have to define what does it mean to be intelligent? What does it mean to be general intelligent? What is this thing we're looking for?" And that's how, as we discussed in the previous episode, I came with this idea of what we call creativity. And I define it as being able to extract new knowledge out of existing information, information you already have. So it's not like you need to go get other new information that you don't have. I already gave you that information.

Can you extract new knowledge? What is new knowledge? So it means can you apply what you learned? Did you get new functionality out of it? And this is actually how good teachers test students like in an exam, what is a good exam is to show that what they learned, they can apply it in a different

environment. You can see that, okay, you learn, let's say The Fourier Transform. Okay, here is a problem. I'm not going to tell you that the Fourier Transform is needed, but you have to think about it, be like, "Oh, here, this is a good solution. Can I apply it?" If you applied it correctly, it means you have knowledge of it. So this is what it means to extract new knowledge out of it. And I like that because you can see actually how some teachers teach. Some teachers when they teach, they don't tell you the solution. They try to guide the students to the solution, and they will tell you, "Think about this." So for example, we talked about example, the binary search that works in a sorted array. And the teacher-

Robert J. Marks:

What kind of search, again?

Georgios Mappouras:

The binary search. When you try to search-

Robert J. Marks:

Binary search.

Georgios Mappouras:

An element in a sorted array, and you can do it faster than just simply checking each element individually. And maybe the students don't know about the binary search and the teacher starts to teach them and it says, "What about if the array is sorted," and the students don't know what to do and then tell them, "What if I pick a random element in the array? What can I learn from them?" Then the students be like, "Oh, random element. Oh, because it's sorted. What I'm looking for would either be on the right or the left of this element," and then they can kind of lead them, right? So they can extract themselves. And that's what I'm trying to find as intelligence, being able to, if I give you just enough information, can you extract the knowledge, new functionality, then you can consistently apply this new functionality from that point on you can apply it to similar problems.

Robert J. Marks:

I like that. That's very much. Yeah. Chollet I thought was very intriguing. I think in the early days of AI, one of the examples would be if you specifically taught a AI to play, for example, checkers on an eight-by-eight checkerboard, and this was pointed out I think first by Gary Smith at Pomona, and unless you program that generality into the checkers playing thing, if you gave it instead of an eight by eight checkerboard, you gave it a six by six checkerboard, it would have no idea how to generalize to it. So that's an example also. Okay. Very interesting. By the way, you mentioned Fourier, so I now feel justified in referencing my book, Handbook of Fourier Analysis & Its Applications in the podcast notes for those that want to learn about Fourier Transforms. You never hesitate to advertise your books. So I will go ahead and do that.

Well, this has been a fascinating talk and I appreciate it. We're going to have one more outing with George, and we're going to conclude this podcast now because we've been going on, I don't know, roughly a half hour. So we've been talking to Dr. Georgios Mappouras about his new interesting paper, the Turing Test 2, The General Intelligence Threshold, which I think is an intriguing new method of measuring whether or not AI is going to be intelligent. I would point out that George has mentioned that his test has not been achieved. He doesn't see a path to it. I'm a proponent of the Lovelace Test that hasn't been demonstrated, and I think for the Chollet that we just discussed, that hasn't been proven yet. In fact, I read that, and I don't know if this is true or not, that Chollet deliberately designed his

theory so that current AI approaches like deep learning and reinforcement learning would fail. And so that's kind of interesting that all of these methods for measuring intelligence are not yet here.

But if you believe in Ray Kurzweil, the singularity is right around the corner and it's been right around the corner for over 20 years. So we'll see what happens. Thanks for listening. Thanks George, appreciate that. We'll see you again real soon, and until next time on Mind Matters News, be of good cheer.

Announcer:

This has been Mind Matters News with your host Robert J. Marks. Explore more at MindMatters.ai, that's MindMatters.ai. Mind Matters News is directed and edited by Austin Egbert. The opinions expressed on this program are solely those of the speakers.