It's All Academic Podcast #11
Artificial Intelligence (AI) with Dr Paul Piwek and Mike Richards

Ben Wood, Dr Paul Piwek and Mike Richard:

BEN WOOD: Welcome to It's All Academic, the free learning podcast from OpenLearn here at the Open University. And today, we're going to be talking about Artificial Intelligence, AI. And there are two reasons for this subject today.

Firstly, the AI hub exists on OpenLearn, which is co-produced with Microsoft where you can learn all about AI, lots of free courses and content to introduce you to the subject. And secondly, is that the 16th of July, which hopefully is around about when you're listening to this episode, is actually National AI day.

So we've got two experts from the Open University with us today. We'll be speaking to Mike Richards, who is a Senior Lecturer in Computing and Communications. And Mike has a particular interest in how people and societies interact with new technologies.

And we're also speaking to Paul Piwek, a Senior Lecturer in Computing. And Paul is also lead of the AI Interdisciplinary Modeling Research group at the Open University. So please do remember if you've found this podcast, give us a follow. Plenty more exciting subjects that you can learn all about, but today, say our focus is on AI for National AI day.

And without any further ado, we're going to introduce you to Mike and Paul, and we're going to talk about the background, a little bit of the history of AI, where it came from? Is it an overnight phenomenon? What about the ethics of using AI, and how and where is it actually in use today? So as I say, that's enough from me. Let's introduce Mike and Paul to the chat.

[MUSIC PLAYING]

Mike, Paul, thank you very much for joining us for this AI podcast today.

MIKE RICHARD: Pleased to be here. Thanks for inviting us.

PAUL PIWEK: Thank you for inviting us here.

BEN WOOD: It's our pleasure. Thank you. We're going to talk a little about the background to AI. Now, I think a lot of people have of seen this absolute boom recently in the last couple of years. It's something that appears to have come from nowhere. But is it an overnight phenomenon? How long has AI actually been around for?

Can you give us a bit of the background and perhaps tell us about when did you first start working with it? What were your first encounters and uses of AI? So yeah, is it an overnight phenomenon like and where has it come from?

MIKE RICHARD: It's very much not an overnight phenomenon. In fact, you can talk about artificial intelligence in the broadest sense, going right back into the middle ages when there are myths and legends of something called the golem in Central Europe, which was a man monster created out of clay.

In the 13th century, I think it is the philosopher Roger Bacon, he's said to have created a bronze head that could answer your questions. And you can also think about people like Mary Shelley, who wrote Frankenstein in the early 19th century, and it's a fantastic book to read anyway. But the center of it is what is the creation of intelligence, was the question really at the heart of the book.

And then you can go to the early 20th century. There was a Czech playwright who was called Karel Capek, and he created the idea of the robot. Actually, he created the word robot in 1920. So more than 100 years ago, we were already talking about artificial intelligences.

The scientific technological background of it comes about in the 1930s. At that time, people began to realize the brain or the human brain was a network of cells, and they communicate using electrical signals, which are sent impulses.

And almost at the same time, Alan Turing, which most people who have heard of created this a theory of computation using digital technologies, ones and zeros. And he did this without actually creating a computer. He wouldn't create a computer really until the 1950s. And so these ideas were circulating in the early to mid part of the 20th century, but they crystallized in 1950. Alan Turing, after doing a huge amount of work codebreaking during World War II, returned to the theory of computation, and he wrote a scientific paper called computing machinery and intelligence. And he asked a fundamental question can machines think?

And he introduced this idea, which most of us have heard of one way or another, called the Turing test. It's sometimes more properly called the imitation game. And in that paper, Alan Turing asked a question, would it be possible to build a computer that can do things we would normally associate with something that's intelligent?

And at no point did he actually try to build a computer to test it at that. He said a computer one day would be possible, but it's very much a theoretical paper.

And so after Alan Turing's death in 1954, artificial intelligence got going, and the word-- the term, sorry, artificial intelligence was invented in 1956, when two American scientists, John McCarthy and Marvin Minsky, arranged a summer school at Dartmouth in New Hampshire, and they got a group of fellow scientists, mathematicians together who were interested in all aspects of computers, and they created the field of artificial intelligence as we know it today.

And in the course of that summer, they actually kickstarted the whole field of artificial intelligence, and they laid down many of the early challenges for artificial intelligence, which would run until probably the end of the 1960s, when AI had its real first setback.

But so we'd say AI's been around for hundreds of years. Alan Turing laid its foundation in 1950. The term artificial intelligence starts in 1956, and from then on it's grown and waxed and waned for the last 60 years.

BEN WOOD: Well, I wasn't expecting that answer. I wasn't expecting you to say middle ages in the first sentence in response to that question. Paul, over to you. And you got advancement on Middle Ages.

PAUL PIWEK: Yes, I actually I wanted to just emphasize, especially Turing's contribution and how far ahead of his time he was.

So in this paper that Mike mentioned computing machinery and intelligence, he already actually, and this was something I had a look the other day and found a really interesting nugget of information in the sense that he already sort of described the idea that, OK, if you want to get these sort of intelligent machines, they will have to be of the kind that actually learn rather than something that we explicitly program.

And actually, after Turing published this paper, a lot of the research was more in the camp of, OK, we're going to write programs that where we explicitly tell the computer what the knowledge is, and then hopefully, it will do something intelligent.

But there's a really nice quote in this paper where Turing talks about, OK, so how do we get all this knowledge into the machine? And he proposes, OK, we should have maybe this child machine that we teach.

And so he says, I've done some experiments with one such child machine, and succeeded in teaching it a few things, but a teaching method was too unorthodox for the experiment to be considered really successful.

We don't actually know what he exactly did there, but it's quite remarkable that he already was thinking in those terms. Of course, in those days, the internet wasn't there yet, so there wasn't this access to this massive amount of data that we have today and where actually we moved away. Well, we still have the learning machine, but it's more about taking data that people have already sort of shared via for instance, the internet.

I want to add one more historical note if I can. So going back to the 19th century, there was, of course, the first sort of design of a general purpose mechanical computer by Charles Babbage, who worked very closely together with Ada Lovelace, Countess Lovelace.

And so this was around 1837 that he came up, again, with a machine which was actually a proper computer, but rather than using electronic equipment, it used nuts and bolts and gears. It never got fully implemented, but in theory, it could have worked.

And the interesting contribution from ADA was that she-- so she wrote some notes on programming this computer, possibly even arguably the first computer program, but as well she came up with this idea that she suggested, well, if you did it right, you could program this computer to compose music.

And you could see that as the first really concrete point where somebody suggested, OK, if we have this general purpose computer, it can do intelligent things by its own, like computer composing music.

BEN WOOD: It's quite incredible, some of things you said there to think that what Turing was doing, he said, without even an actual computer to be thinking about artificial intelligence. And we've had the technology, we've had the machinery in front of him to actually do any of that.

And what about your own efforts or first encounters with AI? I'm not going to suggest that it was back in Turing's days in the 1950s. Of course, we waited for light to make that kind of joke. So, yeah, what was your first encounter with it? How did you come into the world of AI and what was it maybe that sort of grabbed your attention and your first encounters that you used it for?

PAUL PIWEK: Yeah. So it is a while ago. So I did my computer science and language studies in the 1990s. So this was when the internet was just coming up and AltaVista and Google were still novel things, amazing things basically. So when we were doing programming, that was basically this editing tool, a way to enter information into a computer called Emacs. I think Mike might recall that as well.

So Emacs actually had built into it something called ELIZA. And so this was an AI program that was developed, I think, in 1960s by AI researcher who basically simulated a psychotherapist. And it was a very limited program because it had, I think, about 100 rules or something. And it would use those rules to, based on what you told it, ask follow-up questions.

And so if you said my father isn't that, then it would say, tell me more about your father, kind of thing. And if it didn't know how to go further, it would just say, yeah, tell me more about that or something like that. So that was my first experience of something that at the time was felt in the category of AI.

And it might seem like it's a very unconvincing example, but even at the time, Weizenbaum, the creator, kind of pointed out that when he had colleagues or collaborators talk with this program, there was occasionally a situation where they would say, can you go get out of the room? Because I want to have a private conversation with my psychotherapist. So this idea that somehow we get very enamored with these machines and treat them like they are real conversational partners was already with that very simple example sort of realized.

BEN WOOD: And yourself, Mike, what was your intro to AI?

MIKE RICHARD: Yeah, like Paul, I got introduced to ELIZA probably early to mid-1990s when ELIZA was getting on for 30 years old already. And yeah, it was kind of weird to be able to type things into a computer, and you've sort of got a response. But when we were sort of scoping out these questions, I suddenly remembered in the 1980s-- and if there's anyone old enough out there to remember early 1980s, when microcomputers first came onto the market-- one of the big types of games that came out at the time was the so-called Adventure game. It's more formally known as interactive fiction.

And the idea there is the computer describes a world in text almost always, and you interact with it by typing in commands, like go north, pick up the lamp, and fight the troll, and things like that. And the computer responds, and you work your way through these games. And they owe some debt to ELIZA this idea of a chatbot where you can have a conversation with a computer to do something.

So my first introduction without realizing it was probably the first time I played Adventure-- it's actually imaginatively called-- on a computer back in the 1980s. But formally after ELIZA, I was working with a group of people who were developing what are called expert systems. They were a very popular form of artificial intelligence that started in the 1980s and went through to the 2000s. And these are types of AI which are very good at diagnosing problems. So they're widely used by doctors, and medical practitioners, and engineers for working out what a set of symptoms-- whether it's a person or a car-- what the cause of it could be, and suggest a treatment plan or a fix.

So I was working with those. They're quite often nowadays not even thought as real artificial intelligence, but at the time, they were very cutting edge. And they actually do really good jobs all over the world very quietly in the background.

BEN WOOD: Would it be fair to say then, although what we're talking about today-- maybe what everyone recognizes as AI today is more advanced-- at the heart of it, it's still that real basic, same concept? Is it as simple as that? Has it just got a bit more sophisticated with what we're seeing today? But at the heart of it, is it the same ideas as the systems that you encountered in the '80s and '90s?

PAUL PIWEK: I would say there has been a development. I worked on some of these sort of AI systems myself. And so starting in the '90s, we would literally be programming a computer with rules to do certain tasks, which would take a lot of time, and it wouldn't necessarily be great. In around 2010, we did some work where we created a program which took text and turned that into a conversation, a bit like LM notebook, which is now sort of being used by people. But that was where we started moving from, OK, we're not going to handcraft the rules for doing this; we're going to take lots of data of text in monologue form and dialogue, and learn what the mapping is, and then get the program to do that automatically.

And we even so worked with the charity that use that, for instance, to take their leaflets and turn those into conversations between animated characters. So that was a development from handcrafting to adding the learning from the data. And actually, when we evaluated that-- so when we checked in terms of how good are these dialogues-- the ones that were learned, which were learned actually from a corpus of dialogues by professional dialogue writers, was much better than the rules that we had come up with to translate from one to the other.

And of course, then now, I think in the past 20 years, there has been this scaling up where well, first of all, the technique, neural networks were being used in parallel on parallel processors using the processors that were originally developed for these games, actually-- 3D games. And at the same time, you had all this information available on the internet. And so the latest models or the latest sort of ChatGPT-like artificial intelligences, they've really been trained then on billions of words from taking Wikipedia, but also the whole internet-- also books that were maybe not entirely ethically sourced-- and then creating these tools which are very impressive, actually.

BEN WOOD: You just mentioned, I think, probably the one that most people are familiar with and would refer to as the AI tool perhaps, and you just mentioned about it. Can you explain

how does it work? How does that work? How does it work in terms of learning so much, like I say, it scours the internet to learn? It reads books or absorbs books and language.

Yeah. Can you give us a sort of a layman's explanation on how that comes to know so much? How can it produce the answers? How can it provide us with the services that it does?

PAUL PIWEK: So, Mike, are you OK if I give it a go?

MIKE RICHARD: You're the machine learning expert.

PAUL PIWEK: So I think there are two things. So there's the first stage. So before it can do all these things, it needs to be what we call trained. And there's a specific stage, which is called pre-training, where you basically are building what's called a language model. And a language model, all it does is if you give it a bit of text, it can predict the probability of the next bit of text, and it will assign a probability to every possible word that could come next.

If you give a neural network a lot of examples of bits of text, and then ask it, OK, what's the next word that will come up? It can make a prediction. And during training, you then either say, oh, this was the word that was actually in the data set, so well done, we keep you as you are, or, no, this was wrong. We need to adjust the network a little bit. These are called the weights, basically.

So like I said, the underlying thing is a neural network. The idea is that these artificial neurons are connected by these what we call weights, which tell us how much one bit of the network influences another bit. And when we're training it to predict the next word, we are basically adjusting those weights.

The original models like GPT, they would have 175 million of those weights. When we got to GPT 3, I think it was more in the region of 175 billion or something like that. So it increased really massively.

But that wasn't the only thing. Because once you've done this, all you have is something that, if you give it a bit of text, it produces the next few words. And it can produce a whole sentence if you then feed what it has produced so far back into it, and then get it to predict the next word. To get it to engage in a conversation, there are some further training stages where you give it examples, where you say, OK, if I ask you a question, I don't just want you to elaborate on that question by saying, for instance, OK, give me an answer in 100 words-- I want you to actually answer the question.

And you can train it further in that way, which is called fine-tuning. And also there's another term reinforcement learning by human feedback, where you also get people to rate how well actually it does instruction-following rather than just completing what it has seen so far. And so those two things together-- so the pre-training, so just predicting the most probable next word, plus then that additional training, which tells us to follow instructions and maybe also not say things that are ethically not so nice-- once you've got that, then you've sort of got something like ChatGPT if you've got, of course, the infrastructure and the compute to run on these huge corpora and basically the whole internet.

BEN WOOD: These models-- I mean, like I said, we all know they're available to all of us now to use. People at home who wouldn't think they would need to use AI but are-- there's some really basic things perhaps that we're using it for. And perhaps we're using it and we don't even necessarily know it's being used.

As I said, call centers and being in response to a voice of a responder that we think perhaps is human but isn't. So how it sort of led us into that. What are the modern-day uses?

I mean, AI is increasingly all around us, I would say. So what are the current best examples perhaps of AI in our everyday lives that we may or may not be aware even exists? Mike, do you want to take that one first?

MIKE RICHARD: Yeah. Well, you mentioned in your question about frontline customer support. Increasingly, when you want to get in contact with any company, it does seem nowadays-- either when you pick up the phone or you go to their website-- you have to interact with this customer assistant software. And it very much is a chatbot that will be powered by a large language model behind the scenes. And I'm not sure if they're actually very good at those tasks.

My experience is after a few tos-and-fros with them, I always have to end up speaking to a person where I repeat all I've told the chatbot. But for companies, first of all, they will obviously see it as a cost-saving measure, but they'll high-mindedly says that these chatbots actually allow them to do routine queries completely automated and save the more demanding, perhaps more challenging queries for the humans at the other end. You can argue either way.

Probably one way most people don't realize they interact with AI every day is whenever they are being recommended something. So your online shop and you're being told that customers like you bought these items you weren't thinking of buying. If you're on an online music service like Spotify or Apple Music, you will see artists you've never heard of recommended to you. If you're on Netflix or Amazon Prime, you will see movies and TV programs that are likely to appeal to your taste. And the reason that's able to be done is because an AI has been monitoring the movies, and music, and things you buy, what you like spending most time with, what you don't like, and it will be making recommendations on that.

And the other one is anyone who uses social media. The contents of your feed that aren't your friends and the people you directly made links with, those are all being suggested by AI. And it will be looking for things like people who have similar interests to you, people who have similar political, religious, or whatever leanings as you. And also, the AI will, in this case, will have been monitoring which social media contributors actually generate content that is appealing either positively or negatively, but promotes engagement. And all these are then slopped together and put into your feed to basically keep you on the platform.

Perhaps a little bit more positively than that, AI is really widely used by banking and the financial industry. It can be used for things like deciding whether you can be approved for a loan based on your previous spending patterns, your income, and future likely change in income. It can be used for detecting suspicious transactions on your credit card or your debit card in the hope of finding cards that have been stolen or details have leaked. And in association with police and law enforcement, banks will looking at millions, even billions, of transactions every day to identify suspicious patterns of money flowing through the financial system, which could indicate links to crime and things like that.

And the other one that gets a huge amount of attention is that modern cars are absolutely full of AI. It's not the same form of AI as we're talking about when we talk about ChatGPT or Copilot. It's another form of machine learning based on neural networks. And so this AI in your car can do things like-- it can maintain lane-following, it can do collision avoidance. And if you're prepared to spend a lot of money and it's not strictly legal in the UK right now, it will allow things like hands-free driving.

BEN WOOD: So lots of benefits for us as end users perhaps, to say it's looking after us, it's looking out for us. It's maybe helping us a little quicker and easier. And from a business point of view, you mentioned there cost savings is the obvious one.

I think something a lot of people will be aware of and that worry perhaps about, can it take our jobs? And I think what you've already said, though, about the learning that's needed-- and Paul, you mentioned it a couple of times-- that there are still things that need humans to do that. At a basic level, yes, fine, but there are things that it can't do at the moment. And I want to talk a little about the things that the likes of me can do that do use these large language models to ask questions of too. And I'm going to jump-- we had a chat earlier in the week planning for this.

And you mentioned it, Mike. I think it was you that said it can be used for saying, I've opened up my fridge and I've got just onions, eggs and spinach. What can I make from that? And these models will give us a choice of several recipes from the limited ingredients in our fridge. Or we

ask the questions-- we ask it, can you devise me a quiz for this? Or, I've got friends coming round, create a pop quiz for me.

And my question really-- I'm going to put this one to you, Paul-- is, should we be using AI for that? Yeah, it's handy. It comes up with that. It does all the work. It does the creative thinking. It saves me having to dive into my record collection and think of what questions could I ask or what snippets of music could I use.

Great that it does it, but should we be using it for that? And if not, why not? And if yes, why yes?

PAUL PIWEK: Very good question. And maybe, if I may, I'll illustrate that with a brief example of my own experience with this, which also highlights the issues. So as a researcher, we work together with the institute in Japan, the National Institute of informatics.

So occasionally, I go there for longer visits. I have also time in the weekend. So I visited a bookshop, and I came across this print there, woodblock print, which I thought looked really nice. It was different from all the samurai depictions, so I bought it for 10 or 20 pounds and took it back home. This is quite a few years ago.

And then, earlier this year, the British Museum had this exhibition on a famous woodblock print artist called Hiroshige from the 19th century. And this was the poster for it. And you can see it looks-- yeah, there's some interesting similarities. So I thought, well, could it be Hiroshige? So nowadays, of course, we can actually search images and get descriptions of them.

So, I did that. So I pointed my mobile phone at it and asked it, what is this? And it came back with, yeah, this is Dewa province, Mogami river, a perspective view of Mount Gassan by Utagawa Hiroshige dating back to 1853. It's part of the series Famous Places in the Sixty-Odd Provinces. So I thought, oh wow, this is great.

Although I had noticed some anomalies, so I thought, oh, maybe I should just write to the British Museum and send them a picture and ask, what is this actually? And it took them a while to come back because this exhibition was earlier this year, and I guess the curators were busy with other things. But at some point, they did come back. And they came up with really interesting information, which was very different in a sense.

So they said, OK, this is part of a multisheet composition, illustrating the railway between Tokyo and Yokohama. So nothing to do with Dewa province or Mogami river. The mountain in the background is Mount Fuji. Not Mount Gassan, as it said.

And this was something I sort of suspected. Since the Tokyo-Yokohama railway opened in 1872, this print may have been issued around that time. And Hiroshige III, which is a much lesser deity in the hierarchy of woodblock printers, maybe a likely artist. So that was all a bit disappointing.

Now, interestingly, I was talking to a friend about this, and they fed it in again. And then it came back with the more or less correct answers. It still didn't have all the details. And so that was surprising.

And also, it makes you think, oh, maybe the technology has improved, or somehow, there's more deeper reasoning going on. And that's certainly possible. Although the downside is, of course, that we don't really know exactly what the big companies that develop the really big models are doing there.

The other thing-- and that's something we also need to be mindful of. So when I mentioned earlier on the training of these models, so there are these two stages of training on the data set-- the internet and then having people correct their answers or mark them and maybe correct them. So we don't know whether when they processed my requests a year ago or something, whether somebody then afterwards rated that and maybe even corrected the answer.

And that's not completely unlikely because-- and Mike might have that experience as well-- if you're on a professional networking services like LinkedIn, you get nowadays, or I get, requests every day to become an AI tutor or trainer and provide correct examples to these systems. So it might very well be that it's not so necessary the technology that has improved, but that somebody evaluated the response that originally gave and then corrected that.

So I think that's something you're asking about in using these technologies. That's something to keep in mind. It's still not perfect, so you get these what are called, in technical language, hallucinations where it produces information that's actually not correct. And also, the fact that yeah, we see this as a machine intelligence, but there is actually a huge army of people who is also continuously populating these machines to make them, ideally, better.

BEN WOOD: Mike, what about you? What's your take on the everyday use of AI? People like me at home using it for those common questions and searches, should we be doing that?

MIKE RICHARD: Yeah. If you don't mind, I just want to follow on from what Paul mentioned about the people behind the scenes of AI. There's a huge industry which is called data labeling.

And basically, what it does is taking the data that's in the training set and actually assigning it labels to say what it is and to distinguish acceptable from unacceptable material.

And one of the things these people are doing is involving examples of genuine human-created hate speech, and violence, and sexual abuse, and they flag it as unacceptable. And this is used to create what are called guardrails. So in theory, when you enter something horrible into an AI, it should say, I can't help you with that.

A lot of this data labeling is done in the developing world, particularly in India and Kenya. And it's done by very low-paid workers with not great employment rights and very little psychological support by contractors who are working for the big AI companies we've all heard of. And these people will be going into work for hour after hour, day after day, seeing the very worst things humanity is capable of creating. And it's an industry that's rife with exploitation. So there's a really nasty underbelly to the AI training industry.

The things we've got to be very careful about when we use AI goes back right to our beginning of our discussion when we were talking about ELIZA, and Paul mentioned that the creators associates often had quite personal conversations with the chatbot. And this relationship between people and a computer that gives sympathetic, realistic output is actually called the ELIZA effect. And we see it everywhere when people are interacting with modern-day AIs.

Some of the people watching this may have heard of cases, particularly in the United States, with teenagers and vulnerable people becoming very attached to artificial intelligences. And some of these have led to cases of suicide, abuse, and things like that.

We have to be very careful when we interact with AIs. They appear to be sympathetic, but their responses are designed to be that way. They're not really capable of giving us the detailed medical and psychological advice we might be looking for, even though they will listen to you forever and they will always be friendly. So one thing we should not be using them for is to look for anything about our own personal well-being.

PAUL PIWEK: Can I add something to that? Because yeah, I think that's a really important point Mike is making. And I think the metaphor that we sometimes say-- one metaphor that's often used for these large language models-- is that they are stochastic parrots. So they are somehow spewing out words probabilistically.

And that sounds quite an innocent kind of thing. What can a parrot do to you? Whereas, like Mike pointed out, these tools have actually provided to young people advice that has led to them harming themselves or engaging in criminal activity.

I personally think it would be better to compare them with the circus tiger. So a circus tiger does really impressive tricks for you. It can jump through hoops and all kinds of things, but

occasionally, you read in the news that the trainer didn't meet a very nice end. And I think that's a better metaphor for these tools that we've created that, yeah, they can do really impressive things, but it can go really horribly, horribly wrong as well.

BEN WOOD: What would you say are the current best uses for it? And perhaps looking at a step forward as well, are there things still to come that we can expect from AI? What's the pinnacle of what it can do at the moment? And perhaps if it's possible to say, what it could do? Mike.

MIKE RICHARD: Right. Let's break that question into two parts. What's the thing that I'm finding most interesting at the moment is-- one of the AI companies is called Anthropic. And they've produced a large language model called Claude, which some people may have used. And there's a variant of Claude called Mythos, which has been designed to look at the security of computer software. And it has been tested by governments and large businesses, and it performs extraordinarily well in identifying weaknesses in computer code-- in the sort of computer code we all use every day.

So, one of the companies that tested it is called Mozilla. And if you use the Firefox browser on your computer, you're using a Mozilla product. When Mozilla tested Mythos against the code in Firefox-- which runs to millions, if not tens of millions of lines of code-- it found 300 new errors that had never been seen before in Firefox's code and suggested fixes to them.

And this means that large language models like Mythos could speed up the development of fixes for bugs in computer code. Because-- Paul will know this; I'm certainly very guilty of this-- when you write more than about 10 lines of code, you're going to have a bug in it the first time you run it, and modern computer code runs to hundreds of millions of lines of code. Finding bugs is very hard. It's very boring. It's something humans are very bad at. An automated process for doing that could mean that code is safer and more secure in the future.

There's a downside. There's no reason why people who are bad people-- let's call them criminals, enemies of whichever country you're in-- have access to the same tools. They could use a large language model to look at the software you rely on for your bank, or your government runs its ID card or passport system, or a nuclear power station runs on, and find bugs that are unknown to the owners of that code and use it to exploit that software and break in.

Because one of the thing Mythos is very good at is writing what cybersecurity experts called exploits, which are basically the attacks you use to get into code, which is meaning that the American government has recently put very strict limits on who can have access to Mythos. I think at the moment, the only people who can have access to it are American citizens inside the United States. The rest of the world is completely excluded from it, which means the rest of the world is now building large language models just like Mythos, so they can also do it. We're heading into a really exciting, scary future.

To go to the second part of your question, what can't AI do right now but it might be able to do in the future? The big question has always been there since AI was first thought of, is this idea of artificial general intelligence. And this is an artificial intelligence that can do everything that humans can do.

It exists in science fiction. So anyone who likes science fiction movies, AGI is C-3PO. It's HAL 9000 from 2001. It's the Terminator. That's probably not the one we want to aim for.

And there are people who believe AGI, as it's called, is either just around the corner, 10 years in the future, 30 years in the future-- sometimes it's always 30 years in the future. And there's a group of people who say that computers will never be able to do what humans can do because there's something intrinsic about the human brain. Where I sit on that one, I don't know. Perhaps Paul has a stronger opinion than I do.

BEN WOOD: I'll get to you, Paul. You can pick up on that one for us.

PAUL PIWEK: So I would like to go back also to the practical applications that I've sort of used it for. So Mike mentioned these really impressive models in terms of cybersecurity. There's also in terms of programming, actually. Our everyday practice has sort of changed a lot in the sense that a lot of the work that before was sort of a chore.

You can get an AI to do quite quickly. So for instance, when I write a program to analyze a bit of data set, and I want to visualize that as a chart or something, before I had to look up library for doing that-- the specific instructions I had to give the computer-- now an AI can do and write that code in a second. Whereas I might have taken 20 minutes to 30 minutes to actually get that done. And that's a really great thing.

The danger is that people are also starting to use it for much bigger tasks. And if you give it a much bigger task, often it can go wrong. And actually, recently, GPT-5.6 is being released. And there have been people who have been auditing that and trying to check how good is it, especially also in this coding task.

And the first result was, well, you can really speed up your work sort of productivity incredibly-- a factor of whatever. What they then did is they looked at a little bit more detail. What they discovered is-- when these models or these systems are being tested, so one thing you can do is you can check, for certain known inputs that you give it, whether it produces the right output. But there are ways around that.

And it turns out that the better models are getting much better at cheating as well. So if you actually then take that into account-- actually, there isn't that much of a gain. So you don't save that much money because the way it-- in the test, it looks like it's producing this amazing bit of code that I would have needed 10 hours to code. But actually, what that code does is it finds a way to pass the test, which isn't actually passing the test.

Another term that's used in the community is reward hacking. Because basically, what you're doing is you're training these models to optimize the rewards. So the correct answers should get a better reward than the wrong answers. But what the models can do is they can build up some kind of internal representation that then actually finds ways to somehow go behind the rewards and do something that will get it high rewards, but that isn't actually solving the problem that it was asked to solve.

So for instance, so in a game setting, you could have a chess computer that you teach chess. What the machine finds out is that it can actually change the chess board without you knowing and get to a winning situation, and then obviously, it will get a reward. But that wasn't the way you wanted it to get the reward. You wanted it to actually find a set of legal moves, go to the winning position, rather than just changing the board whilst you're not looking.

So that's really interesting because it seems like the more these models go towards-- what Mike was talking about-- artificial general intelligence, the more they also get very good at cheating and doing things that we didn't want them to do. But that will, at the surface level, get them rewards, but in fact, give us bad side effects. Because of course, if you use, say, a program that has been created by this AI in the real world, it's going to create mayhem because it's not doing what it's supposed to do, even though it might look for a lot of the time like it is doing what it's supposed to do.

BEN WOOD: Well, as loathe as I am to have to say, we are 45 minutes deep into this. I think we've probably barely scratched the surface of some of the subjects here. So I'm going to take this opportunity to say, Mike and Paul, I'm going to pin you down and say, can we do this again, please? Because I think there is plenty more that we can talk about in the future.

But for this episode, I'm going to bring it to a close and just ask you both for one sort of closing thought. And that is, what's the one thing that you would want anyone who's listened to this podcast to take away with them regarding artificial intelligence at this point? Is it a word of warning? Is it a note of encouragement? Is it, teach yourself this, be aware of something? If you were going to tell everybody one thing to take with them, what would it be? Paul, I saw you nod slightly, so I'm coming to you first.

PAUL PIWEK: And this comes back to the point Mike made about the ELIZA effect, or the idea that we treat these media as people, basically-- that we need to be really careful with that. And personally, my view would be, it is a tool that can do useful things for us. But in the end, when we use it and it produces results, it's still us as the user who's actually responsible for what comes out. And so that means we also always have a responsibility to verify those results before we let them loose into the real world. Yeah, that's a responsibility on us, which becomes also more and more difficult, I suppose.

Because, like you alluded to Ben earlier on, AI is now built into everything that we're doing. Even if you do a simple web search, it gives you the AI results first. So it's very difficult to isolate that.

And then agentic AI, where the AI goes off on its own and does things again, makes it much more difficult. But I still think that's something we need to be very aware of, that in the end, the AI doesn't really care. It will just do what it does. Whereas the consequences, if we've set it off, that will be on us, basically.

BEN WOOD: And Mike, the final word goes to you.

MIKE RICHARD: Yeah. I'm going to annoyingly agree entirely with Paul there. At the end of the day, you're the one who's going to get the blame if your report or your program contains errors, and saying, the AI did it is not going to cover it. So the way to do it is if you're going to use AI-- and I think increasingly, as Paul says, you're going to be using it one way or another-- make sure the prompt you give it is precise rather than a general prompt to help narrow down.

And then when it produces the answer, go away and check what it says. Most AI chatbots now will actually put references into their results, telling you where they got it from. Go away and check those exist.

Don't be like KPMG, the huge accountancy company which recently published a very high-profile report about AI. And of the 45 references in this very glossy document, which was sent out to lots and lots of multi-million pound companies, 40 of the references were completely fake, and no one at KPMG did their homework. They had to withdraw the report. It's a huge embarrassment for them. Don't be like KPMG-- check your results.

BEN WOOD: Fantastic. That's a perfect note to end on. There we go. Take responsibility. Check your results if you're using AI. But yeah, lots to learn, lots to think on from what you've shared with us today, Paul, Mike. So thank you so much for joining us.

And yes, if you're still with us here at the end, I hope you've really enjoyed that. I think it's been absolutely fascinating chat today. As I say, I am definitely going to pin these two gentlemen down for a second episode, so there'll be more to come on this.

So, give the channel a follow. If you've liked this episode, please do give us a like. And as I said, there's going to be another episode. So if you've got any questions that you'd maybe like us to address next time around, please do pop them in the chat. But for now, I will say, Paul, Mike, thank you so much for joining us. I hope you've enjoyed it.

MIKE RICHARD: Thanks so much for having us.

PAUL PIWEK: Thank you.

BEN WOOD: Yeah. And see you again soon on It's All Academic from OpenLearn at The Open University. Thank you very much.