## **Andrew Ng Guest** 00:00

Do AI systems actually understand the world or are they sometimes called stochastic parrots? I think AI systems do understand the world and they're also kind of getting better at explaining what they understand. What they don't understand, In fact, you know this is the term stochastic parrots. This is just parroting words, mimicking intelligence. But there was actually one study that really influenced my thinking, called Othello GPT, in which a model was trained to predict the next move on the game of Othello. When we think about existential risk, I think there are existential risks to humanity. I think high on the list would be, you know, maybe the next pandemic fingers crossed. Or I think that global climate change is a risk to certainly a large fraction of humanity. What it has to do with the dinosaurs there was many tens of millions of years ago, so it's less of an original risk, probably not going to happen in our lifetimes, but who knows? Very unlikely to happen in my lifetime, I think. But to me, when you look at the actual things that could be an extinction risk to humanity, I think AI would be a positive submission.

## **Craig Smith Host** 01:08

Hi, I'm Craig Smith and this is my AI. In this episode, I'm going to talk to Andrew Ng, a well-known researcher and pioneer in machine learning. Andrew has taught courses at Stanford and trained many of the AI luminaries working today. For a time, he was the chief scientist at Baidu, China's AI giant, and he co-founded Coursera, the online education company. Currently, he has an ML Ops company called Landing AI, which has a suite of tools to make it easy for businesses to build, deploy and manage machine learning models. I wanted to talk to Andrew because he's on the optimist side of the so-called threat debate, arguing that AI does not pose an imminent threat to humanity and that there's plenty of time to work out safety protocols, guardrails and international agreements to prevent catastrophe. Now here is Andrew. To begin with, do you believe that AI poses an existential threat, even in the long term?

## **Andrew Ng Guest** 02:21

No, I don't. I'm trying to keep an open mind, but I'm having a really hard time seeing how AI creates it. I know Jeff Hinton and Yasha Benjo have both spoken about the extinction risks to humanity from AI. I spoke with both Yasha and Jeff and I have the deepest respect for them. I'm trying to follow the reasoning. I'm struggling to see how AI, which is a fantastic technology, is making society much better off. I struggle to see how that creates any meaningful extinction risks for humanity.

## **Craig Smith Host** 02:55

Yeah, and just for people that haven't been following the debate, but see the headlines there's no evidence, no scientific evidence, to support the notion that AI poses an existential threat. Right, it's theory, it's speculation.

## **Andrew Ng Guest** 03:19

Yeah. So I feel like the two arguments about AI potentially causing an existential risk to humanity. The two arguments I heard is : what if AI is so powerful not GPT-4 but GPT-25 or something that allows some really angry person or maybe someone to create a bio-weapon that then causes humanity to go extinct? I think the risk of that seems small because it's actually pretty difficult to wipe out humanity, no matter how angry one person may be. I do think this is a risk of meaningful harm. Someone could create a weapon, maybe, but extinction is a scale of risk that seems very unlikely, very implausible to me.

## 04:10

And the second argument that I've heard is sometimes called the paper clip argument, not to diminish it with that terminology. But what if someone mis-specifies an AI objective, like building a profitable company, but it figures out a way to make a lot of money, accidentally wiping out humanity? I think if an AI is smart enough to do that, it will probably be smart enough to understand our intentions. So many people work on AI safety, so AI is smart enough to accidentally wipe us out. We'll probably be smart enough to know if we said make a lot of paper clips, we did not mean make a lot of paper clips and along the way. It's fine to wipe out humanity, so that seems unlikely to me as well.

## **Craig Smith Host** 04:54

Yeah. Is there a way that lay people looking at this debate can differentiate between real risks and speculative fears?

## **Andrew Ng Guest** 05:06

So I don't know. It's a good question, it's challenging. So AI, just to acknowledge it, does have short-term harms. Today They've been documented cases of bias, unfairness. You know we've seen that polyprogram self-driving cars systems can lead to car accidents, these kinetic events that lead to human death right now. So I feel like there are those harms that AI researchers are working to address. One nice thing about large language models like chat, gpd, bot, various models is I do see many researchers working to make them safer. So they're getting safer every month right now.

## 05:42

So if you are good about that, you know sometimes think about the aviation industry. With the rise of aviation, aeroplanes were really dangerous. They crashed, they killed people, completely tragic. No excuse for really aeroplane crashes. Having said that, humanity, we learn from aeroplane crashes and over time aircraft got much safer.

## 06:05

Today, I think about some aspects of AI, because people say you can never control a large language model and you know what else you can't control. You can't control an aeroplane either. You can't perfectly control where an aeroplane points because winds will blow it around. And sadly, even today, even you know, in society today, there are occasional aeroplane crashes which are very tragic, at least a really catastrophic loss of human life.

## 06:30

Having said that, I think society seems much better off to me with aeroplanes, then they're not. And even though we can't perfectly control an aeroplane, we can control them pretty well. So when I got on an aeroplane just a few days ago I felt fine, you know, working on my laptop and not really worrying about the plane crash. And I think that we're on that trajectory as well, that large language models, even the last six months, have gotten much safer, much harder even for some of the militias to make them do something bad Although I think it is still possible for some militias and determined. But I think that over time the risk of either accidental or malicious harm seems to be going down rapidly and I don't think we'll ever fully control them. But we can't fully control aeroplanes either and I feel okay getting in one and trusting my life to one.

## **Craig Smith Host** 07:19

Yeah, A couple of questions. One this concern about AI centres really on intelligence, how intelligent AI systems are becoming, and there's the debate that, well, you won't know when AI reaches super intelligence, because it won't necessarily tell you. Can we objectively measure the understanding of AI systems? Do you think?

## **Andrew Ng Guest** 07:56

So this is the debate about do AI systems actually understand the world, or are they sometimes called stochastic powers? I think AI systems do understand the world and they're also kind of getting better at explaining what they understand, what they don't understand. In fact, this is the term stochastic power. This is just parroting words, mimicking intelligence, but there was actually one study that really influenced my thinking, called Othello GPT, in which a model was trained to predict the next move on the game of Othello the game of reverse Othello and the authors of this study, which was published in the iQuer conference, I think, demonstrated that the way to predict the next move in this game is to actually build a world model. And specifically, the authors found that if you probe the state of the neural network after being fed a set of moves, it seems to be learning an underlying representation of the status of the board. So it's not just mimicking surface level, parroting out what's the next move, it's actually figuring out what the game of Othello is. So they're from Melissa Moose. It figures out what's the board, status of the board and therefore what are the possible next moves. So to me that was a fascinating study. A small scale, it was just the game of Othello.

## 09:17

But that really convinced me that I think today's large language models are building a model of the world and to me I believe they do understand the world. And also, if a large language model says something and you say, no, that isn't right, or you added that maths wrong in step five, the fact they can say you know what? I apologise for my error. I did make a miscalculation in step three and now let me fix step three and this is my new maths result for some maths puzzle.

## 09:49

I think that analogies between machines and humans are always dangerous, but I feel like I don't know that that's very different from how a child learning to solve maths problems for the first time, they interpret it. So I feel like the language models are getting better at explaining to us how they're seeing the world and just like when my daughter explains to me how she's thinking about the maths problem, I believe her. I think she's telling me the truth. So I think when the large language model was telling us how it's using the world for the most part I know it doesn't always tell me the truth, but I think it says something close enough to the truth. It feels okay to believe it at least some of the time when it says that's how I was thinking about the problem.

## **Craig Smith Host** 10:37

Yeah. So on the threat, your conversation with Jeff sort of the takeaway was that the research community needs to come to some consensus about the level of the threat and how to counter the threat. Because right now, with the divergent opinions or the split in the community, the public's confused and certainly policymakers, regulators are confused. So how can the research community reach a consensus on the threat?

## **Andrew Ng Guest** 11:21

Yeah, more conversations. I feel like I'm continuing to have conversations with very different views, including those actively worried about extinction to those worried about other risks. Also, many people are just busy building valuable applications, but not just me, I think everyone in the community can engage in polite, constructive discussion. I think that we can, maybe slowly, probably you'll be slower than any of us won't gradually come to a narrow range of views. I remember six, maybe seven, eight years ago, with the rise of deep learning kind of the wave of deep learning was started, maybe 10 years ago, a few individuals, including most notably Elon Musk's Bill Gates, warned of some of the risks of AI, and I think there was a spurt of worry at that time that then candidly died out.

## 12:17

I think I heard much less about that after a while. I think Elon Musk said something like with AI we may be summoning the demon. I may be disclosing him, but I would respectfully submit that in the last decade we've summoned a lot more angels than demons with AI. I mean not to dismiss the harm. There have been harms but net-net AI's have been a massive conlude to society, and so I think that wave of worries frankly died out, this new wave of worries with buy-in from prominent scientists, with Jeff and Yasha talking about extinction risks. That was a surprise to me, but I think it'd be worthwhile to keep having conversations to see where we end up with this as well. I'm really trying to keep an open mind, even though I think I said I just don't get it. I don't see how AI could lead to human extinction.

## **Craig Smith Host** 13:07

Yeah, yeah, the lack of that consensus. Is that a danger in itself? That either regulators listen to the people who are pressing the existential threat narrative and over-regulate, or they listen to those who say, look, those threats are far, far in the future, if they exist at all and under-regulate.

## **Andrew Ng Guest** 13:40

Yeah. So one of the things Jeff Hinton and I said in that video that I posted on social media Twitter and LinkedIn is I'd like you to mix this in real life for Kendrick One of the reasons why climate scientists have been an effective force is because they are aligned. So there are economic interests to ignore climate science and let's have lots of and just keep on generating carbon emissions. But the fact that climate scientists are more or less unified on the science of climate change, that's made it much harder for economic interests to find it more convenient to just not be regulated, to lobby regulators to ignore climate science. So I think, while the AI community is splintered, it actually makes, unfortunately, much easier for anyone with an agenda to lobby regulators to whichever argument is more convenient for the business. So the AI community is unified, then I think, like climate scientists have, we collectively do a much better job of helping regulators unleash AI to create all the value it can, while also mitigating against realistic risks. And just to be clear when we think about existential risks, I think there are existential risks to humanity. I think high on the list would be maybe the next pandemic finger stress. Or I think that global climate change is a risk to certainly a large fraction of humanity, and what the asteroid did to the dinosaurs there was many tens of millions of years ago, so it's less urgent or risk. Probably not going to happen in our lifetimes but who knows? Very unlikely to happen in lifetimes, I think.

## 15:27

But to me we look at the actual things that could be an extinction risk to humanity, I think AI would be part of the solution. So AI, I don't think the world responded that well to the last pandemic, but I think AI enabled drug discovery and monitoring of disease conditions. That seems important to me as climate change becomes worse. I think AI enabled solutions to mitigation, to the smart grid, to accelerate the electrification of society, or Candelie. I often think about climate geoengineering. Do we need solutions like using high-altitude aerosol sprays to slow down cooling of the planet? Or, if not doing it now because it's a dangerous technology, certainly to advance the science of climate geoengineering with AI-enabled modelling of the climate so we better understand if this is even an option which is seriously considered. So if we look at the real extinction risk to humankind, at least the ones I could foresee more clearly, it seems like an important piece of the solution. So I would say if you want humanity to survive and thrive for the next thousand years, I would rather make AI go faster rather than slower.

## **Craig Smith Host** 16:41

I want to give a shout out to this episode's sponsor, masterworks, an art investing platform that buys contemporary masters outright works by Picasso and Bansky and Warhol and others then qualify it with the SEC to offer it as an investment. Net proceeds from the sale are distributed to its investors. Last year, most investors lost money, even with a diversified stock portfolio, and that's why some of the biggest money managers in the US have been looking outside the stock market for assets with low correlation to traditional investments, things like Fine Arc. The opportunity was exclusive to professional investors until now Masterworks makes it available to ordinary investors like me. Since the inception, they've sold over $45 million worth of artwork and so far, all of Masterworks' exits have delivered positive net returns for their investors.

## 17:48

Of course, historical returns are not a guarantee for future returns. I'm not a financial advisor, so do your own due diligence. There's always a risk of loss. Masterworks has over 750,000 users and their art offerings usually sell out within hours, which is why they've had to create a wait list. But I on AI listeners can skip the line and get priority access right now by signing up at www masterworks art. Backslash I on AI, e-y-e-o-n-a-i all run together. Take a look and see what you think. Yeah, yeah.

## 18:35

I agree. I mean that's an argument I've given around the dinner table that the real existential threat is, or the confirmed existential threat is, climate change and AI can do a lot to help us adapt or mitigate that.

## **Andrew Ng Guest** 18:55

In fact, a few years ago I think Yasha Benjio and I were both authors on this, but David Roelnik wrote a really nice survey paper on all the ways that, on many of the ways that AI can play a role in climate change. So that paper, which I still really like David Roelnik and others did all the work but I seem quite proud of having participated in that work to think through how AI? Will climate change.

## **Craig Smith Host** 19:26

Yeah, that was those foundational papers for climate change AI. I remember that there's a lot of concern on the threat side and not only the existential threat but immediate harm about open source AI systems and that allow people to build beyond the purview of regulators or corporate management and that sort of thing I was surprised to find recently I mean I know you were it was quite a while ago but you were at Baidu I was surprised to find that there's a very robust open source ecosystem growing up around large language models or large models in China. Do you think that open source is a danger? And, specifically on China, have you used any of the open source stuff they've put on GitHub?

## **Andrew Ng Guest** 20:41

Yeah, I think the open source community is a global one. There's a lot of great work in the United States, there's a lot of great work in Europe and there's also a lot of great work in China. So what I'm seeing is that open source work is a fantastic force that is giving access to a lot of people around the world to large language models. I remember when Stable Diffusion released their image generation model the trained model open source in late 2022, I think there were justifiably reasonable fears of wow, what if this is useful harm or this open source one, anyone can generate images now whether that goes really poorly and the S&D. That has to be acknowledged. There were some problematic use cases.

## 21:28

But I would say, compared to the harm from a handful of problematic use cases to the massive amounts of innovation where people incorporating the models into all sorts of creative video and image editing software, new UIs, I think the benefits vastly outweigh the harms, and I'm not dismissing the harms at all.

## 21:48

I think that the models do exhibit bias. You ask the general people of certain professions. They only generate pictures of one gender or narrow range of ethnicities. Those are real problems that we need to work to solve and there are certainly negative, harmful use cases of deep fake imagery. Having said that, what we've seen so far is that releasing open source seems to create much more benefit when weighed against the harms. So I'm nervous about proposals to require licensing of models, open source models that seems like an absolutely terrible idea the aside for innovation and lead to regulatory capture, where it's much easier for big tech companies to satisfy regulatory requirements, thus concentrating power in large tech companies and disadvantages academic research groups and smaller businesses that don't have time or resources that deal with overly onerous regulatory requirements that really are not, in my opinion, effective at really protecting people.

## **Craig Smith Host** 22:51

Yeah, and I don't want to get off on a dog lag on China, but immediately, immediately after GBT three came out, there was a lot of talk in the press about how China is going to be left behind because their speech laws, their laws, the controls speech, restrict political speech, will limit the, the efficacy of large language models, and I argued against that to a lot of people because our open AI or Microsoft or Google's large language models also have restrictions on speech. So I thought that was a spacious argument. Nonetheless, I was surprised to see open source thriving in China. Does that surprise you at all, just from what you know of China, gosh?

## **Andrew Ng Guest** 23:50

I don't know, I can't say I'm an expert on that, but it has been fantastic to see open source models thriving all around the world and it's been interesting how global the research community is. A few weeks ago I was in Canada for CVPR, where you are speaking about landing AI's visual prompting technology, but it was nice to see researchers from all around the world come together and just share ideas freely and collectively, work hard to advance the signs of AI, a computer vision, at that conference.

## **Craig Smith Host** 24:26

But really AI will probably yeah, actually, and I'm sorry I've left this till the end, but can you tell us about the visual prompting developments that you've been pursuing at landing AI?

## **Andrew Ng Guest** 24:42

Yeah, so I think the text prompting revolution, which I've already seen through ChaiGPD and BARD and so on, is coming to vision, and so Lightning AI has been developing visual prompting technology to make it easy for anyone to build and deploy computer vision applications. So maybe yeah, I'm going to flash my neuro-licence and get technical for a second. But the text prompting revolution was really enabled by text transformers. The team I started and once led Google Brain, published the text transformer paper in 2017. And since then there was a wave of innovation by many, many research groups on scaling up text transformers and that led to GPT2, gpt3, megatron so many models along the way, until we got InstructGPD and then ChaiGPD and BARD and Bing Chat and so on. Not many people know, but the vision transformer paper came out about three years later, in 2020.

## 25:47

And if you go to computer vision conferences, I think since 2020, there's been this wave of innovation and scaling up and exploration of how to get vision transformers to understand images and the underlying trends for both text and vision. The vision, coming a bit later, is that these are models trained on very large amounts of data, text or images. The training is increasingly on unlabeled data, so this lets you have access to a lot more data or unlabeled or technically self supervision. And what was already seen in text is, with pre-training on neural networks, on tons of unlabeled text data. You can then feed it a very simple text prompt and have it make inferences.

## 26:34

What we're starting to see in computer vision, which I think is like maybe a couple years, one and a half years, two years behind text, is when you create a very large vision transformer on a lot of images on the internet, then you can give it a simple visual prompt and then have it start to make inferences in just seconds. So landing AI will release our take on visual prompting where anyone can go and use it. Go, use it for free to label a, prompt it with a few labels, a few pixels, and have it then automatically figure out what to do after that. As a developer too, this is exciting because, just as we've seen with text prompting, applications that used to take me six months to build, now anyone can build in maybe a day. Applications used to take me six months to build in computer vision that with tools it's now getting to be possible to build in maybe a day or a few days. So the innovation is exciting, not just from my team landing AI, but from many groups working on computer vision.

## **Craig Smith Host** 27:40

Yeah, that's fascinating. And just on visual prompting large image models, if that's what they are, they're vectorizing an image by converting each line of pixels to a vector. Is that right, yeah?

## **Andrew Ng Guest** 28:04

patches the image rather than catches.

## **Craig Smith Host** 28:08

Yeah, but they tokenize that in the same way that for a sentence, you convert that to an embedding and then you're trying to predict the next token, right yeah?

## **Andrew Ng Guest** 28:25

and just like now. Yes, so I think in text-based transform the core task, sometimes called the pretext task, is to predict the next word or, technically, predict the next token. So, because text is that linear order, predicting the next word, the next token, you know, is a nice way to use unlabeled data. One of the things, one of the reasons why there's so much buzz and exploration in computer vision, is the process for taking an image and converting that to a sequence of tokens. There are a few different options for that. So, you know, do you convert to the patches? Do you hide or mask some patches along the way? But so different researchers are exploring different ways to do that, but there's multiple ways to see them through working and that scaling up of vision transformers we're already seeing very exciting results From LAMI, aio and then Metasam segment anything model was another exciting breakthrough I'm just seeing.

## 29:25

Actually, when I was at CVPR earlier this year in 2023, I felt that vision transformers have become a solid alternative to convolutional neural networks and the buzz and innovation. I remember a couple years before, chatting GPT in the NLP on the text processing community, everyone knew something was in the air about transformers. No one knew exactly what's going to happen, but everyone knew something was well, people knew something was up and today at CVPR, I feel like something's in the air. People know something is up, even though the exact applications, the exact ways this will get used, is still being worked out.

## **Craig Smith Host** 30:05

No. And your visual prompting the landing AI tool is specifically for, I mean, it's primary use cases for labelling images. Is that right? Oh no.

## **Andrew Ng Guest** 30:19

Visual prompting is for building computer vision applications. So, for example, we have users using it for cell counting. Quite a lot of light signs as users, but given a picture of cells in a peak you dish in seconds, you can build an application to detect the cells and then post-process it to count the number of cells. Or I'll now use it to handle some satellite imagery where I show a lot of geospatial error imagery applications as well. Easy one finding tree cover. You can now do that in seconds. And then, of course, a lot of landing AI's.

## 30:51

Users have been manufacturing industrial automation, and so for many of the manufacturing defect detection tasks, especially ones based on texture rather than shape, our visual prompting is letting people build and deploy systems in seconds. But I've been surprised also, even though I mentioned some of what we are seeing as our most common use cases industrial manufacturing, light sizes and geospatial error imagery I'm seeing a very long tail of applications as well, where many people, many of our users, are coming with all sorts of applications in computer vision that we've not imagined and now able to build and deploy them. For example, there was one group that was doing cosmology. Since I'm not an astronomer, I would never have thought of doing that with a computer vision visual prompting application. The sort of creativity once people have this bewildering, just like, I think, when people had access to chat GPT. The creativity of things people did with it was incredible. And we're definitely not yet there at computer vision. But I think collectively the field is getting closer and closer.

## **Craig Smith Host** 32:02

Yeah At landing AI. Where are you personally? Where are you on the spectrum between peer research and building applications? Landing AI focuses on applications. I mean, where are you in that spectrum?

## **Andrew Ng Guest** 32:23

Oh, I think we're a product company, but the product, the tool for building so landing AI provides software that makes computer vision easy. So the product landing lens makes that possible, but the product is backed by deep tech. We've been doing our own internal research on visual prompting for like a year and a half, maybe not quite two years, but so I think the way we approach the product is via deep tech. So even today, I spend a lot of my time worrying about how to improve visual prompting and improve the algorithms as well as on the product and how to make it easier to use.

## **Craig Smith Host** 33:01

That's it for this week's episode. I want to thank Andrew for his time. If you want to read a transcript of today's conversation, you can find one on our website, ionai, that's EYE-ONAI. In the meantime, remember the singularity may not be near, but AI is about to change your world. So pay attention. No-transcript.