CAUGHT IN A WEB

PODCAST TRANSCRIPT

UNSW Centre for Ideas: Welcome to the UNSW Centre for Ideas podcast – a place to hear ideas from the world's leading thinkers and UNSW Sydney's brightest minds. The talk you are about to hear, *Caught in a Web*, features *New York Times* tech columnist Kevin Roose, and UNSW Sydney's Toby Walsh, Scientia Professor of AI, and was recorded live. We hope you enjoy the conversation.

Kevin Roose: A few years ago, I was radicalised on the internet. Now, I know what you're thinking after that introduction. I'm a journalist, I cover technology for *The New York Times*, I'm probably going to tell you about how young men are being turned into Neo Nazis by their YouTube recommendations or how your grandma is in grave danger of becoming an anti-vaxxer from too many Facebook memes. But my radicalisation story is quite a bit stranger and frankly, much more embarrassing than that, because I was radicalised by Spotify into being a Jimmy Buffett fan.

It started a few years ago, during the first year of the pandemic. I was bored, I was anxious, I was burned out. One day, I was in my home office trying to get a piece done. I was trying to get some music that would help me finish this piece – and I just cranked up Spotify and put on my Spotify discover playlist. And then this song comes wafting out of the speakers, says "wasting away again in Margaritaville". And my first thought was, this is strange. I'm not really a 'Margaritaville' guy. I'm more of like a jazz, blues, hip hop guy. And for a minute I wondered if my account had been hacked or if some engineer at Spotify had decided to troll me in revenge for something I'd written. But I kept listening, played it again and thought to myself 'kinda catchy!'.

Pretty soon I had an idea. I went into my house, I dug some pina colada mix out of the cupboard. I put on a Hawaiian shirt in my closet, I went back to my office, closed the door, cranked up Jimmy Buffett, *voila*, instant vacation. I was transported from my house in

California to a swaying hammock in the Caribbean. Over the next few years, this little break became a regular and embarrassing part of my self-care routine. I started calling it ‘island time’ and my wife just mocks me mercilessly for it. But I don't care because it actually works. Whenever I'm feeling the weight of the world, I pull up Spotify put on my ‘aloha’ shirt and head out to my office for a few minutes at the beach.

Now, I know that this was probably not the radicalisation story you expected. My usual reporting is much darker and more about extremism. A few years ago, I hosted a podcast series, as Toby mentioned, called *Rabbit Hole*, which featured the story of a 26-year-old guy named Caleb Kane, who was radicalised into the far right through his YouTube recommendations. I also wrote about this woman, Valerie Gilbert, a Harvard-educated New Yorker who was sucked into QAnon after spending too much time on Facebook. I even wrote about PewDiePie, one of the biggest online celebrities in the world, who has more than 100 million YouTube subscribers, and whose name was mentioned by the shooter during the 2019 Christchurch massacre. I still believe those stories are important. But this is the Festival of Dangerous Ideas. And frankly, the idea that the internet is turning some of us into extremists doesn't seem very dangerous anymore. It's just common sense.

We all have seen people who disappeared down online rabbit holes and emerge months, or years, later as completely different people. So, I started today by telling you about my embarrassing Jimmy Buffett addiction because I want to propose a slightly more dangerous idea – which is that it's not actually the extremists that I'm most worried about. We are all different now, as a result of our encounters with the internet. And I literally do mean all of us – me, you the people sitting on your right or left – no matter what your beliefs are, which politicians you vote for, or which apps are on your phone, none of us is immune to the forces that are being exerted on us by the internet. And actually, if I could make this take a few degrees hotter – this is actually a subject that I think that conspiracy theorists that I've met and reported on grasp a little bit better than some of us, frankly, because at least they understand that they're being manipulated. Now they think it's like Hillary Clinton or Bill Gates or the blood-drinking Satanists at *The New York Times* who are doing the manipulating, but at least they know that something on the internet is not right.

The machines that influence our behaviour go by many names. There are recommender systems and personalisation algorithms, 'for you' pages, reinforcement learning models… just to use a few. They are designed by some of the smartest people in the world – PhDs in computer science, experts in fields like persuasive design and choice architecture. Many of them are powered by artificial intelligence. Some of them are invisible, some of them are visible. Some are benign, others are malignant. But what they all have in common is that they are sitting on the other side of our screens, attempting to change who we are and what we do.

They're there every time we pick up our phones, sitting on our shoulders and whispering into our ear, 'don't you want to buy a different brand of dog food? Shouldn't you vote for this other politician? Wouldn't it be fun to listen to Jimmy Buffett?'. And it's important to understand that these persuasive features aren't just minor sideshows. In many cases, the machines actually run the show. So, I pulled up a few, like, statistics about how deeply these algorithms are embedded into what we do on the internet every day. So, at YouTube, for example, 70% of all time spent on the platform – billions and billions of hours a year in total – are the direct result of algorithmic recommendations. Amazon has said that 30% of page views on its entire site are generated by recommendations, which at its scale translates to many billions of dollars a year in revenue. 'Spotify discover' playlists, which are those totally algorithmically generated ones that I mentioned at the beginning, account for more than half the streaming income of something like 8,000 artists, according to the latest figures. And TikTok, which is just one giant recommendation engine, is the fastest growing app in the world. US teens spend an average of 89 minutes a day on it.

The point I want to make today is that these things matter deeply. That in many ways we are what we pay attention to, and that understanding the forces that are trying to manipulate us online – how they work, how to engage wisely with them – is the key not only to maintaining our independence and our senses of selves, but actually to our survival as a species. In order for us to live free and fulfilled lives, we need to know where the machines stop, and our human selves begin. So, I want to start by telling you a few things that I've learned in the course of years of reporting on these machines and how they influence us.

The first is that we trust machines, even when we shouldn't. A few years ago, I had a bizarre experience. This is another embarrassing 'algorithm happens to Kevin' anecdote. I had been using one of those 'wardrobe in a box' services where they send you clothes. I just don't like shopping very much. So, I'd signed up and you put in your measurements, and you make some selections about what kind of clothes you like and then they send you a box of clothes every month. And I remember I was wearing this blue bomber jacket that they'd sent me, and I'd probably worn it a dozen times. And I was standing in front of the mirror one day, and I just had this sort of experience of being like, 'I hate this. I don't like this at all. This looks terrible on me, why am I wearing this?' And I started thinking more about this and the many ways that we let machines influence us even when our common sense would tell us something different. You could call this 'algorithmic deference'. Because the truth is that often when our own judgment and a machine's judgment collide, we let the machine win. We just assume that it's right, that it's smarter than us more sophisticated, that it knows something about us that we don't know, and that we should just shut up and go along with it.

The second thing I've learned is that machines don't just read our minds, they *lead* our minds. A few years ago, a team of researchers at the University of Minnesota came up with a series of experiments to test the influence that algorithms have on our choices. And the way that they set up this test, they recruited a whole group of college students, and they took a list of popular songs, and they gave each of those songs, ratings between one and five stars. And then they said, you know, these ratings are personalised to you. And then they asked the students to listen to little snippets of the songs and to say how much money they would be willing to pay for them. This was during the iTunes era when people actually paid for songs. And this is actually a great way to design an experiment, by the way, because it relies on something that economists call Revealed Preference theory – which is that if you want to know what someone values you don't ask them, you look at what they pay for. So, they asked the students 'how much would you be willing to pay for these songs?'.

And what the students didn't know is that behind the scenes, the researchers had manipulated the rankings. Some of them got more stars than they would have otherwise gotten. Some of them got fewer stars. And then they forced the students to listen to the songs all the way through and asked them again how much they'd be willing to pay. The

researchers were pretty sure that forcing students to actually listen to the entire songs would neutralise the impact of these star rankings. The students would form their own opinions, right, and ignore the stars. But they were wrong. What they found instead was that these recommendations actually overrode the students' own individual tastes. They offered significantly more money for higher-rated songs, even when those ratings were totally made up. The results were clear – when an algorithm tells us that we're going to like something, we trust the algorithm more than ourselves.

Now, the moral that I'm trying to leave with today is not that all algorithms are bad, because in a lot of ways, they're great. Like when we open Netflix to find something to watch, we don't want to scroll through an alphabetical list of all the programs on Netflix – we like that it presents us with a smaller set. If we put something like a smart thermostat in our homes, we want it to learn from us what temperature we like in which room and automatically adjust itself. These things are not innately bad. The problem arises when the machines get their own ideas – when they can start messing with our options and our defaults, drawing attention to not what they *think* we'll like, but what they *want* us to like.

There's a technology researcher named Christian Sandvig who calls this 'corrupt personalisation'. And it's a real problem because we do trust these machines and this corrupt personalisation betrays that trust. What if Netflix decides that it would actually rather you watch its own original programming than the shows made by other networks, and just says that those are the recommended shows for you? What if the smart thermostat company strikes a deal with the maker of wool blankets to imperceptibly turn down the heat by a couple of degrees at night, so the blanket company can sell more blankets? That might seem far-fetched, but this kind of corrupt personalisation happens every day.

Social media platforms shove viral videos into our feeds, not because they think we'll actually like them but because they think we'll stay engaged and they need the ad revenue. We know that Amazon steers people toward its house brands in search results because it wants you to buy from them – it makes more money than when you buy from a third-party. And because these algorithms pass themselves off as personalised, and because we trust that they are actually personalised to our desires, we are predisposed to trust them. That is a

big deal. And I've said this to audiences. And one time someone came up to me, he said, 'well, isn't this just no different than advertising before the internet? Like you saw, you know, a TV ad, you'd be predisposed to go out and buy the whatever the ad was for'. But this is actually different – the veneer of personalisation makes it much easier to trust what they're telling you. In the old days, you might see an ad for, like, a pair of pants on TV. And you might think 'those are for somebody, but they're not for me'. But now we've got these machines that are telling us over and over again, like 'no, no, you like these pants, these pants were made for you. Trust me on this because I know you'. These machines don't just read our minds, they lead our minds.

The third thing I've learned is that all of these machines are getting smarter. A few years ago, I wrote a book called *Future Proof* about AI and automation. And just a few weeks ago, I wrote a column in *The New York Times*, it was called '*We need to talk about how good AI is getting'* – sort of updating some of the ideas from my book. And this column was my attempt to draw attention to the many advances that have been made in the field of artificial intelligence just in the last year or two. I wrote about the improvements that have been coming to systems like GPT3, which is a text-generating, text-completion algorithm that was developed by a company called Open AI – and it's sort of like a super-powered version of the autocomplete feature on your phones. It can take… You can give it a prompt like, you know, 'rewrite *Romeo and Juliet* in the style of Edgar Allan Poe', and it can actually do it. I wrote about AlphaFold, which is an AI engine developed by the Google subsidiary *Deep Mind* that has learned to predict the structures of proteins – which is actually, like, something that has bedevilled molecular biologists for decades. Like being able to take an amino acid sequence and predict the 3D structure of a protein. They trained this algorithm, originally, to play games like Go, and then it was able to learn how to predict the structures of proteins – which was a major ground-breaking accomplishment in the field of drug discovery and biomedical research. The *Journal of Science* actually named AlphaFold its top breakthrough of the year last year.

And I wrote about these AI image generators that you may have seen. These really just are one of the newest things in the field, they just came about over the last six months or so. They've got names like 'DALL-E 2' and 'Midjourney', and you basically… The way they work is

you plug in some text into a prompt, like, 'I want to see a Soviet-style propaganda poster of Pikachu riding a skateboard'. And it will turn them into realistic imagery. I published this column, and immediately I got dozens of emails from AI researchers who basically said like, 'good story, but you don't know the half of it'. Because they were telling me about stuff that they're working on, and how all of these AI fundamental advances are being commercialised and combined with other technologies, like targeted advertising, to give companies a much, much more detailed idea of who we are, what we want, and how to modify our behaviour. So, just to give you one example, a researcher told me about this new technology called 'pedestrian re-identification' – which is basically a fancy way of saying that companies now know how to use AI to track you across multiple cameras, with multiple resolutions and data sources, as you walk down the street and combine that with other sources of data. So, if Susan is captured on a closed-circuit TV coming out of the train station, there are now AI technologies that can extract her image, link it with a different image from a different camera of Susan going into a coffee shop, and put that data together with data from Susan's phone and credit card company to show that actually what she was in the coffee shop, she bought a slice of banana bread. And so now in a totally unsupervised way they can say, 'okay, every time Susan gets on the train, show her an ad for banana bread' – it can connect those dots. And that's actually… that's not like a dumbed-down illustration of what this is being used for. I thought that this research would be coming out of academic labs. But it's not. It's coming out of commercial enterprises, businesses are using this.

I also heard from AI researchers, that I was not spending nearly enough time talking about what you might call the 'era of ubiquitous synthetic media'. So right now, when I write a story for *The New York Times*, that story goes online, it goes in the newspaper, and no matter who you are, or how you're reading it, it's the same text, it's the same images. But in the very near future, there are going to be new sites that use tools like GPT3 and DALL-E 2 to generate entirely new stories on demand at the point of click – that are going to be personalised to a reader in real time. So instead of, you know a million people reading the same news story about the war in Ukraine, you'll have an AI model that generates stories about the war in Ukraine for you, in real time, by looking at your browsing history, your IP address the information on your social media profiles, and proactively generating a story that is specifically tailored to your interests with a slightly different slant than the one your

neighbour right next to you is reading. This is, to put it mildly, like, a very radical shift. We spent many, many hundreds of years in a media environment where information was produced and distributed by humans. We then, in the last 10 or 15 years, moved to a world in which information was produced by humans and distributed by machines – the Facebook newsfeed, the Google search algorithm, etc. And now, we are moving into an era in which information and entertainment will be produced without human involvement at all, and distributed by computers.

Another way to put that is that we are fast moving away from the era of all the news that's fit-to-print and into the era of all the news that's dynamically generated and personalised to achieve business objectives. So, what do we do? How can we cling to our humanity, and our sanity, even as machines are trying to drown us in a sea of corruptly personalised recommendations, and synthetic media? Well, there is some good news here – which is that it's actually not that complicated to get ourselves out of this mess. We don't all have to go back to school and become programmers. We don't have to throw our devices in the ocean, renounce modernity and move to the mountains. Well, you can do that if you want, it sounds kind of nice sometimes. But I have just three things that I think we can all start doing today to prepare ourselves for this new world.

The first, I'm going to borrow from one of my favourite technology writers, Socrates, who said, 'know thyself' – maybe a slightly different context in which he meant it. But this is really where it all begins. Because to have a chance of surviving this wave of accelerating AI, we need to look inward. We need to understand ourselves at a deep level. Where we get into trouble with these sort of algorithmic deference instances is where we're not actually quite sure what we value. We're not sure what brand of dog food we want, what TV show we want to watch – our indecision and our lack of self-knowledge actually makes it much easier for us to be influenced. So, in order to survive, we need to construct of positive identity, a robust identity that can withstand an assault from machines. Because self-knowledge is an area where we actually do still have the advantage. Right now, some algorithms seem to understand us eerily well. I was watching a video the other day on TikTok. There was a woman who was saying, 'you know, for years, I would go on TikTok, and I was just very confused because I would see all these videos of really attractive lesbians, and I'm married

to a man and, you know, I just sort of brushed it off as like, maybe the algorithm just misfired. And, you know, I'll go cleanse my algorithm by looking at videos of puppies or whatever'. And then she said, 'you know, fast-forward a couple of years, I'm coming out. I've left my husband for a woman that I met at a bar. Holy smokes! TikTok knew I was gay before I did'.

But I'm sure we've all had the experience of running into an algorithm that doesn't know us at all. My favourite example is when you buy something from Amazon, like, you know, you buy a screwdriver, and it says, 'would you like a screwdriver?' No, I just bought it. A lot of these algorithms, they're, they're getting smarter, but they're still pretty dumb. They get it wrong, they don't pigeonhole us correctly, or they miss what we think is important about ourselves. But this is changing, machines are getting better. And so, our knowledge of self has to get better too. If the machines learn to understand us better than we understand ourselves, especially on things that are sort of core to who we are – like our desires or our emotional needs, then the authority will transfer to them. Knowledge is power, and when it comes to keeping algorithms in their place, self-knowledge is power. We also need to resist machine drift. Machine drift is a term I came up with a few years ago, and it basically describes the opposite of self-knowledge. It's, it's when we give over more and more of our decisions to machines, without really thinking about it, like we let them pick, you know, which music we listen to, you know, what food we eat, etc. But also like, where we go on vacation, how we think, how we vote, what we pay attention to – it's this kind of passive surrender. And it allows machines to influence us at the preconscious level. And this is really bad, like this is, this is the first step in losing our autonomy.

So, as a first step to resisting machine drift, I recommend doing what you could call preference mapping. So, I went through this exercise myself a few years ago – write down all of the choices that you make in a day, no matter how small, and try to figure out which of those choices are actually yours, and which are being shaped by machines. Do you buy those shoes because the website recommended them or because you actually liked them? Do you?

When you commute to work do you just go the way that Google tells you to go because it's the way Google tells you to go, or do you go the way you want to go, even if it takes a few minutes longer? You can also start putting a little more friction back in your daily routine. You can take the long way to work. I actually disabled the YouTube auto-play feature that plays another video when the one you're watching is done. I also disabled the thing on Gmail where it completes your emails for you, which takes me a little longer but it's, it's actually me communicating with people, not a robot. There are these moments where you can make things a tiny bit less convenient for yourself in hopes of spurring some self-reflection, some moments where you step back and say, 'Wait a minute, is that really what I want to say or do?' Another good way to resist machine drift is just to spend less time on your phone. A few years ago, I did a 30-day phone detox program with the help of a professional phone rehab specialist. It was really bad. It was an amazing experience. It was not easy, but it did help me in many ways – it improved my focus and productivity. I rediscovered hobbies, it got me back into reading, and it radically improved my marriage. If you're laughing nervously at that, just find me after the talk, I will introduce you to my phone rehab specialist.

The third thought that I want to share with you is that in order to survive into the future, we need to invest in humanity. What does that mean? It doesn't mean that we should just go like fund a bunch of philosophy scholarships, although that would be fine. What it means is that we constantly need to be working to improve our deeply human skills – the things machines aren't very good at yet, like empathy and moral courage and divergent thinking. We also need to start valuing those skills more highly in others. And here is where I'm actually very optimistic, because I think that in the coming years, as AI continues to accelerate, we're going to see a renewed kind of humanism emerge, and a growing reverence for the things that machines can't do. I think we'll collectively spend more time on apps and social networks that actually connect us to people rather than just addicting and enraging us for profit. I think we'll gain a greater appreciation for people like healthcare workers, teachers, therapists. Maybe we'll even have an artisanal journalism movement and I'll get to come back to Australia and do this talk again a few years instead of a robot standing up here and doing it for me.

By working on our deeply human skills, and investing in the deeply human skills of others, we not only make ourselves harder to replace, but we fortify the kind of society that can be collectively resilient in the face of technological change. The machines can only beat us if we let them. And honestly, I don't think we will. It's too much fun here in 'Margaritaville'. Thank you.

…

Kevin Roose: Toby, how do you feel about Jimmy Buffett?

Toby Walsh: Well, I know of course, I'm gonna get very personal here because I know you used to be a barbershop quartet. I prefer Jimmy Buffett to barbershop quartet. So yeah, maybe the algorithms have progressed you on.

Kevin Roose: That shirt is actually a little margarita… that you'd be right at home in 'Margaritaville'…

Toby Walsh: It is! I was thinking I could take this off and start sipping the margaritas.

Kevin Roose: Yeah.

Toby Walsh: But anyway, thank you very much. Thank you. Thank you for making us think… Protecting ourselves. I wonder, how much do we have to be protected from ourselves? So, whenever I go to Silicon Valley, I always think it's, it's a really strange Kool Aid they're drinking there, you know, the house philosophers. And around the house festival is Burning Man. It's a pretty strange place. And I do wonder, and some pretty strange people. I mean, it was this idea of charities, you know, effective altruism. Tell us something about the strangeness and whether we should be protected from that.

Kevin Roose: Yeah, there's, I mean, to generalise about Silicon Valley is hard. And there's, there's a lot of people with a lot of interesting and strange and troubling views there. And some good ones like I'm not an effective altruism stalwart supporter, but I do think they have

some good ideas about how to effectively allocate resources and philanthropy. I think that one problem that Silicon Valley has right now, is that the only things that matter are the things that can be measured. That's sort of a philosophy of many people in the tech industry – is that, you know, if it matters, you measure it. And machines actually aren't that good at measuring things like long-term satisfaction. So, like, you know, if you ask me what I want to eat, in any given moment, I might say, "well, I want to eat a healthy diet consisting mostly of fruits and vegetables, and some complex carbohydrates and some protein". But I'm also going to… If you put a plate of cookies in front of me, I'm probably going to eat the plate of cookies. And then so then the question is, well, which of those things are these tech platforms optimising for? Are they optimising for our, our aspirational selves, or our more impulsive selves? And right now, the thing that is possible to measure is our impulsive self. What do we click on? What do we read? What videos do we watch? What do we swipe on? And so, I think there's sort of a movement and I'm hopeful, because I do hear people in tech now talk about things like how do we measure long-term satisfaction and optimise toward that? So, I think this is starting to change, but it's a slow process, and the commercial incentives are always going to benefit the short-term.

Toby Walsh: If you're measuring likes and clicks – people always say to me, "Well, Toby, you've just need to put a different algorithm in there". And I think, I think any algorithm, if the only input is like, likes and clicks, it's going to end up probably in pretty much the place that we have ended up with polarising content and extreme content.

Kevin Roose: I don't know, I think you could design an algorithm that optimises toward long-term satisfaction. I mean you could show people when they log in to YouTube or TikTok, you could give them sort of a time capsule. You could say, you know, this video that you watched six months ago, are you better off or worse off for having watched that? And, you know, people with some self-awareness might say, "well, actually, it felt good in the moment to watch, you know, 14 videos of you know, people getting hit in the groin with various sports balls, but it actually didn't improve me as a person and I wish I hadn't done that. And so please don't show me any more of those". Like, I think you could actually design something that was more long term.

Toby Walsh: People who use social media – I mean, there's various studies that suggest that people who use social media are less happy, have less self-worth, especially teenage girls, as a result of using social media than those who don't. I mean, so is… Is there hope? Or is that inevitable?

Kevin Roose: I don't think it's inevitable and the reason I don't think it's inevitable is because the market is responding already. One of the biggest things that's happening in tech right now is that Instagram is declining, especially among young people. And it's being forced to, you know, turn itself into TikTok and it's adding features to compete with these newer apps called BeReal and things like that – which are much less about displaying perfection and filters and airbrushing and making people feel bad about themselves if they don't have this idyllic existence. And so, I think there's actually a generational shift happening where young people are watching the people who are slightly older than them go through this real ringer of self-doubt and self-loathing, and they're saying, like, "I don't want that for myself". And they're using different apps. They're voting with their, with their, you know, their swipes and their taps. So, that does give me hope.

Toby Walsh: So, TikTok gives you hope?

Kevin Roose: It does. I mean, TikTok has its problems.

Toby Walsh: It's highly addictive.

Kevin Roose: It's highly addictive. It's, you know… Owned by a Chinese company…

Toby Walsh: The data is owned by a Chinese company. Yeah… It's got close ties…

Kevin Roose: But, but I think it gets rid of some of the problems with Instagram. I think what we can hope for is that each generation of popular social media app will be slightly less poisonous to teenagers and to, frankly, to everyone – and BeReal, the app I mentioned, I just started using it. I'm way too old to be on it. But I think it's an interesting case study in how the market can respond to the excesses of the previous generation of social networks – you

can only post once a day, there's no filters, there's no public likes, it's not sort of a popularity contest in the same way. And it's very popular. It's one of the fastest growing apps in the world. So, I actually think I'm not optimistic about the old social networks changing to be less poisonous. But I am optimistic that new ones will come in to replace them and will solve some of these problems, or at least ameliorate them through better design.

Toby Walsh: So, there's no hope for Facebook?

Kevin Roose: Well, for many reasons, I think there's, there's I would not, you know, place bets on Facebook being around in 10 years. But I think that, certainly we need to fix the problem. We need, you know, we need people. We need people like Francis Haugen to show us a way to improve these legacy social media companies. But MySpace never fixed its problems. It just died. And something else came along to replace it.

Toby Walsh: I want to go back to the promise that social media had. So, at the start there was a significant promise. I remember the Arab Spring, the Trump election, where you know, social media was used in a very positive way to empower people to get disenfranchised.

Kevin Roose: The Trump election?

Toby Walsh: The Obama election.

Kevin Roose: I mean, the Trump election definitely did empower some people on Facebook. It's just… They were all in St. Petersburg.

Toby Walsh: But you know, as you say, right. So, those same tools that we use to, to encourage people to go out and participate, were then turned for very negative ends. And, in fact, a lot of the manipulation was persuading people not to vote as much as it was persuading them to vote for Trump. It was persuading, you know, black people 'don't bother to go to vote this time'. Should we be… You talked a lot about protecting ourselves, but should we be, should government and people be, stepping in and protecting us from ourselves?

Kevin Roose: I think certain regulations could make sense. I think things like, you know, a national privacy law in the US would, would go a long way. I think transparency is something that only governments can really enforce – forcing these platforms to just tell us what's going on in them and making their data more accessible to researchers and journalists and the public. But this world moves so fast that I actually am not sort of a regulation-first proponent. I think that that actually I've, I've changed my mind a little bit on this just by seeing, I mean, the US government is still talking about Facebook as if it's the most dominant company in the history of the world, right? And meanwhile, you can't beg someone under 20 to use Facebook. They won't do it. And they're, they're struggling mightily to compete with this newer generation of social media apps. They're pivoting to the metaverse because there's, they're worried they're being left behind. So, I just I think that, you know, the pace of transformation has gotten so quick, that it's going to be very tough for regulators to keep up.

Toby Walsh: But even if it's tough, right, I mean, here's a radical idea. The idea that we micro-target adverts, that we tell everyone different untruths. I don't get to see the untruths that you get to see. They're specialised for my likes, and they're specialised for Jimmy Buffett for you. Is that really helping our political discourse? Or should we just say, you know, let's, let's ban micro-targeting and political adverts full stop?

Kevin Roose: Sure. That's, that's one possible solution. It's a pretty blunt instrument. But I think, you know, I, I think you could, you could say, you know, you can't target segments below a certain size, you can't target 15 people with an ad, you can target 15,000 people.

Toby Walsh: How about this – over the age of 18, I could make a vote and they're in my electorate.

Kevin Roose: Sure. For… I think for political ads, regulation makes a lot of sense. We already regulate political ads and other mediums. But that's, that's a very small slice of the persuasion machine.

Toby Walsh: And truth – should we regulate for truth?

Kevin Roose: … Sure, if you want to give governments the power to regulate for truth? Go for it. But the DeSantis administration in the US is going to have some thoughts about what constitutes truth.

Toby Walsh: So, I'm going to ask one last question, and then we're going to go to your questions, right. So, we've made, I think it's fair to say, quite a lot of mistakes in how we've rolled out the internet. Are you going to come back in five-years time, to the Festival of Dangerous Ideas, and tell us all the mistakes we made for the metaverse? We repeated it just the same again.

Kevin Roose: I don't think the metaverse is going to be here five years from now, maybe 15.

Toby Walsh: Okay, 15 – choose your time zone.

Kevin Roose: But certainly…

Toby Walsh: That will be more seductive because it's visual…

Kevin Roose: Sure

Toby Walsh: and oral…

Kevin Roose: I'm not a big… I don't think the metaverse is going to look like what Mark Zuckerberg thinks it's going to look like. I don't think it's going to be us all, you know, wearing helmets with our legless VR avatars, wandering around high fiving each other. I do think that computing is going to get more immersive, it always does. And it'll get closer to our bodies. And maybe we'll have glasses that have heads up displays or something like that. And there will be a tonne of new safety and trust challenges with those. And there'll be new ways that they manipulate us. But you know, I think that I hope that the lessons that we've learned… We really did something that humanity had never done before, we really connected the entire

universe, the entire world – billions of people with direct access to tools to communicate and broadcast – that's sort of a one-time change. It's like a zero to one thing. And I think, I hope, that we'll learn some lessons from the first 15 or 20 years of that experiment that we can then apply to the next phase.

Toby Walsh: Great. So, we're gonna take questions now. Please keep them short, please make them a question.

Audience Member 1: There's a growing trend towards software output, towards multipurpose super apps. So, if you take the data that Uber has – it knows where you go, and what you eat. You take an e-commerce app that knows what you buy. And you take a financial services app, which knows how you spend. And you put all of them together, you have a massive data set. And so, there is a trend towards building these to try and capture as much of your spend as possible. As a humanist, if you were designing such an app, what would be the watch-outs for you, and how would you design it?

Kevin Roose: If I were designing a super app? I mean, I think that there are some benefits to, to consolidation like that. So, if you have a lot of… If you have one company that provides 20 functions, rather than 20 different companies, each providing a function, you can sort of benefit from economies of scale when it comes to things like removing bad actors from the platform, targeting, manipulation, misinformation, things like that across different apps. You can just have one team that manages all of those surfaces. But I, you know, I'm not sure how I would design a super app. I've never been asked that before, so…

Toby Walsh: You just have to look at China – there's WeChat.

Kevin Roose: Well, WeChat and, and, you know, the Chinese government is no slouch when it comes to censorship and an oversight of online behaviour and speech. I don't think that's how I would do it, if that's your question. But yeah, I'll think about that some more. Thank you.

Toby Walsh: Question from the side.

Audience Member 2: You said the biggest defence against all of this is 'know thyself'. Now, that's all very well for you and I, who have had at least a large portion of our lives spent before the internet. But what about today's children? You know, they're getting a smart phone, phone thrust into their hands at the age of four, and they have no chance to know themselves. They're going to be influenced by whatever they're seeing on their phone from the moment they can think. How do we fight against that?

Kevin Roose: I think it begins by how we educate children. I, I've written about the need for more programs like social emotional learning. If you actually look at what the technologists in Silicon Valley are teaching their kids – their kids are not going to coding bootcamp, their kids are going to Montessori school, they play with wooden blocks in the mud. And that's how they educate their kids, because they understand, you know, those are there, you can teach vocational skills, like programming later, but in the early childhood years is when you really have to learn things like, like sharing, like, how to figure out how someone else is feeling. Empathy, collaboration, skills like that, that serve you well through your whole life. So, I think I think we really need to double down on those kinds of skills, especially in the early childhood years.

Toby Walsh: Very old school. Okay, next question.

Audience Member 3: How is the rise of porn – access to it and the algorithms within it – affecting and shaping sexuality in society?

Kevin Roose: Amazing question. I, I am obsessed with porn… in a journalistic sense. I think it's one of the most under-explored topics in modern society, because you're right, it is shaping attitudes about sexuality. It's something that we don't talk about or write about all that much. But it is ubiquitous. And we can all name Mark Zuckerberg. Who can name the proprietor of, you know, the biggest porn site in the world? And yet, they probably get about the same amount of traffic. So, I think it's, it's something that we need to do more... I'm actually like… I probably should write something about it one of these days. But I think it's, it's fairly important, especially because it does seem like an area where there's a lot of

potential for real harm, not just, you know, videos of, you know, non-consensual – whatever horrible thing you can imagine showing up on a porn site. But just that is one of the biggest generational changes, and it's talking about, you know, things that kids are exposed to early. That is one of the biggest generational changes in the past 20 years, is that, you know, you used to have to, like, you know, pull the magazine out from, you know, under the mattress, and now you've got this device in your hand that gives you access to anything you can possibly imagine from a very young age. So, I think it's a really important question. I know it's not an answer, but it's I'm affirming that that is, like, a thing that we need to be thinking and talking more about.

Toby Walsh: And troubled about.

Kevin Roose: Yes.

Toby Walsh: Okay, one more question here.

Audience Member 4: Throughout your presentation, you refer to this issue of personalised algorithms as being caused by machines. And all of these machines are generated by companies who want profits. So, do you think that our focus here should be on the machines themselves? Or do you think that the companies themselves play a huge role in this? And how does that relationship work?

Kevin Roose: Yeah, machines, for now, are built and trained by humans. So, that's clearly where the buck stops. There may be a time when algorithms are building and training themselves. I mean, there are self-supervised algorithms that train themselves to do various things now, but humans are still in the driver's seat – they control the inputs, the data sets, etc. So, absolutely, this is a story about human accountability. And every time in my presentation that I use the word 'machine', there should be a little asterisk that says, 'machine that was created and overseen by humans for mostly commercial ends'.

Toby Walsh: Okay, the final question.

Audience Member 5: So, you talked about knowing thyself, and following on previous audience talking about that – I'm actually interested to know for adults, most of us are adults here. When we wanted to fight against the algorithm, you suggested that we should find out, kind of like, our 'why do we do what we do?'. But if we can't even figure out, why do we choose a certain brand of tea or restaurant, how do you propose us actually examining, you know, the preferences that we have and the history of it? Because quite frankly, a lot of us probably don't even remember what rabbit hole we went through, what kind of practices that has influenced us to arrive at where we are today. So, I'd love to hear how perhaps even if, you know, tech founders, how do they not become a slave to these algorithms as well as kind of other people in Silicon Valley or even yourself?

Kevin Roose: They, they meditate, and they take psychedelics from what I can tell. Sorry, they do! All of them. No, that's a flippant answer. But I think there is something to be observed there – which is that the people who run the biggest technology platforms in the world are not glued to their phones all day. You know, they have assistants to who are, but you know, they're… Jack Dorsey goes on 10-day meditation retreats. And I mean, they're all involved in mindfulness practice of some kind. I think that's, that can be useful. I think that the question you're asking is the right one, though, which is, how do we understand? I think that we do need to do some, some sort of what, what might be called, like attribution in the advertising world? Like, where did I get the idea that I liked this thing? I think that's actually like, not super hard if you just sit down and focus, and it might not be true for everything. A lot of you know, advertising, for example, like, you know, operates in sort of a cloud of our subconscious, where it's like, 'why do I drink Diet Coke? Is it because I saw an ad, you know, 10 years ago, probably, but I'm not, I'm never gonna be able to connect those dots'. I think it's actually less important for things like, what brand of soda you drink, it's more important for, like, certain beliefs that you hold, you know, why am I angry about issue X, Y, or Z? Why is that salient to me? Is that consistent with some deeply held value that I have? Or am I just seeing a lot of other people get angry about it every day when I go on Twitter? That kind of thing.

Toby Walsh: Kevin wasn't that in some sense, despite all the pain, one of the gifts of the pandemic? We stopped, and we realised that being out in nature, going for a walk, meeting our family, meeting our friends – those were the things that were valued to us.

Kevin Roose: Yeah, but we also spent way more time on our screens. So, I think there was a little bit of a bifurcation there. Yeah, but I think I think you're right; we did… It did force a re-evaluation of priorities and values. I think was probably helpful.

Toby Walsh: Right before we thank Kevin, he'll be signing copies of his book and possibly other books. And if you want to know what an AI researcher thinks about machines, behaving badly, I'll be signing copies of my book. So, thank you very much, Kevin.

Kevin Roose: Thank you.

UNSW Centre for Ideas: Thanks for listening. This talk was presented by the UNSW Centre for Ideas and Festival of Dangerous Ideas. For more information, visit centreforideas.com and don't forget to subscribe wherever you get your podcasts.