My interests are in how we can better inform decisions around gathering of knowledge, sorting of data, what we can see instinctively as humans that possibly can't be seen through formal risk capture and data processing approaches. I've always been interested in this idea that humans have some access to a dimension of perception that possibly machines don't.

The Cognition Test

When you look at the history of AI, the original reference point that we all need to be aware of is the Turing test. This is named after the famous British scientist Alan Turing, who famously cracked the German secret high command code in the Second World War, built what's widely regarded as the first functional computer. And he said, There will be a future in which machines will catch us up in intelligence.

And so the test that I propose is, I want you to imagine that you're an interviewer and you're interviewing a candidate and the candidate entity, whatever it is, whether it's a person or a machine, is sitting behind a screen, so you can't see that you get no visual cue as to whether it's a person or a machine answering your questions. And if the conversation flows freely to the extent that you can't tell if it's a person or a machine, or if you think it's a person, but it turns out to be a machine, then the machine has passed the human cognition test. So, it is as if it is a human being.

I'm fascinated having conversations with my new friend, ChatGPT-4, over the last couple of months that it gives a very, very good impression of being a human being. I get it to interview me and test some things that I'm writing and so on. Very often it will ask intelligent questions, except when it doesn't. It's like a newly arrived research student on your team whose maybe they're not a graduate yet, maybe they're a fresh out of school, undergraduate. They're looking at gathering information. They're enthusiastic to capture all the information and to stick post-it notes all over the wall of all this stuff that they know, but they have no great capacity to analyze and absolutely no capacity to synthesize, that is, to gather all the data collectively and infer and induce and draw thoughtful, empathetic human conclusions from it.

The best analogy I can think of, is building a house out of Lego bricks, little building bricks. Generative AI basically just keeps on producing little bits of conversation. You want to call them morphemes or whatever we talk about in linguistics, little units of meaning. But essentially, it's just attaching little units of meaning to other units of meaning. A little bit like a child building a Lego house, and it has no concept of what the shape of the house is going to be. Just when it's ended up with something house-shaped, it stops building. So that is the limit of its perception.

Processing Capabilities

Now to discuss petaflops. A petaflop is a unit of computation. So it's essentially how fast your brain in a resting state is running. The power of compute, that is the hardware and software that drives generative AI, last year broke through the 18-petaflop barrier. It was physically not attainable for computers to think that fast. Even the fastest computers, up to about five years ago, were nowhere near the associative processing speed of the human brain. However, it is estimated that the computer will break through that processing threshold, so the upper limit of the human brain's capacity by end of year 2024.

What we are experiencing, however, is these exponential kickups happening simultaneously in more than ten different fields, not just computing. The banking industry, for example, uses voice traders - people making contracts using microphones, voice boxes on their desks. A voice trader maybe working a ten hour day, typically in a currency or a derivatives market, you've got maybe 2000 voice traders in London, you've got another 2000 in Wall street, you've got another 4000 in Singapore. So each bank may well have, let's keep the number simple, 5000 voice traders who are active for a ten hour day each day. Now, I don't know if you've ever looked at when you've taken a voice memo on your phone. I think it's like a megabyte of data for every ten minutes of trading in terms of the data capture of your voice, which remember, is an unstructured data source, your voice is just a continuous spooling, noise making machine. It doesn't sort of neatly divide up into data quantities which a machine can immediately peel out and analyze. So you have to have complete capture of your voice. And imagine that multiplied by 5000 traders by 10 hours a day, by, let's say 250 days trading a year, by whatever that figure was, a megabyte per ten minutes. Just the sheer quantity of data is unfeasibly massive to manage. And yet AI, as of now, can take all of those terabytes worth of data, not just listen, but it can do. Now what has previously eluded the banking industry, which is to spot patterns of misconduct at the early stage of their development.

Example of using AI to benefit the banking industry

Code switching is something that's used in social science a lot to describe the different sorts of vocabulary that we use at work or at home or with maybe close family versus sort of wider public realm. Code switching for traders can take the form of, we had an example yesterday, traders who are working in Spanish switching into Italian, or traders who are trading in Italian switching into Portuguese. Now, these are languages which are in some ways close, in some ways not so close, but classically really difficult for an auditor to spot. If you were a Spanish speaking auditor, you wouldn't necessarily immediately pick up on something switching to Italian.

A voice trading analysis of these hours and hours of unstructured data, an auditor would classically listen to maybe a sample random selection of sort of five minutes from this day and then five minutes from another day and five minutes from another day clearly is not going to capture any sort of devious behavior. The chances of even knowing what you're listening to are not necessarily that great. This machine, this AI, can listen to an entire year's worth of voice trade tapes and it can say, right, this trader in Rio and this trader in Rome and this trader in Hong Kong are colluding because they switch languages. They maybe go to a WhatsApp group and then they do an off market trade, or they use some gray space to do the trade that I think is absolutely fantastic. And that gives us the prospect of getting on top of money laundering, off market trading, insider dealing, gray market exploitation, which in finance has eluded us, frankly, forever.

Lagging Regulation

Human optimism leads people to explore. You know, in the same way that the explorers in the Middle Ages got into boats and sailed off the edge of the known world to see if there was another continent to be discovered. It's an extraordinary thing that's human optimism that makes people push out the boundaries. Unfortunately, in the social culture and the risk-taking culture of Silicon Valley, we have a new direction for that.

The testimonies that some of these developers have provided to US Congress or any the think tanks that have been starting to look at this, are concerning. The developers themselves still seem to be a little too much in the driving seat, which means that regulation remains behind. Everyone sees the benefit things go well for a while until there is a catastrophe. And then the regulation tries to make sense of what just happened. It tries cognitively to kind of reappraise the catastrophe. Regulation, unfortunately, always addresses the last catastrophe and this is how we had the banking crisis in 2008. The regulators failed to predict the thing that was happening next.

Metacognition and Hallucination

Metacognition is the human skill of knowing what you don't know. It's knowing where to stop talking because you don't know what you're talking about. I have a research colleague at Yale who looks at this, how many people are locked in literally incarcerated in the US justice system on the say-so of so-called expert witnesses who were actually exceeding their brief, beyond the limit of their skill. Now generally AI is largely lacking in metacognition. I have managed to get Chat GPR-4 a few times to say I don't know, but it's very rare what it will normally do is supply what it thinks is the best fit. Answer to the question.

In AI, such “fibbing” we call hallucinating. But let's be clear hallucinating. It's kind of like greenwashing. It's just an elegant word that means lying. Can you honestly say you've had a conversation, and you know your client well enough that simply getting our friend ChatGPT to, as it were, kind of pluck a load of stuff off the wall and stick it into admittedly beautifully shaped, clearly worded, sequentially laid out document. Is that the same thing as actually considering your client's needs, or is it just getting a machine to, again, to take my previous analogy, peel off all those sticky labels that the researcher has put on the wall and sort of file them into an alphabetically correct order?

I don't want to rubbish generative AI because it's exceedingly clever, but the problem that we have as humans is because it passes that Turing test that I talked about, because it can sustain a conversation, because it can give the appearance of answering a question, we've got to be incredibly careful. It's not actually answering the question. Maybe think a bit harder about client service? Don't defer to the machine so utterly.

Tech Utopia

For many, many years, technologists have sold us on the idea that the technology, because it now exists, is going to make our lives better. The simple existence of the technology will make our lives better. Let's just take three simple examples.

- **Antibiotics**

We can now operate. People could go in a hospital, they can have open surgery on parts of their body, and we can kill the bacteria, which previously posed a serious hazard that a lot of people who went in for quite straightforward operations died on the table because they contracted infections which we couldn't counteract. So antibiotics fabulous invention, completely changed healthcare. And we had over a century where antibiotics worked flawlessly and the people who would have died from infections simply didn't. And a triumph for healthcare. Except, of course, bacteria don't work like that. Bacteria are constantly evolving and it's like an arms race. So we

invented what we thought was the knockout blow, but actually it wasn't. And we now have so called MSRAs. So medication resistant bacteria. If you give somebody an antibiotic when they go into the operating theater, there is now a chance that they might die of the same thing that killed your great grandfather before antibiotics were invented. So that's one.

- **Autonomous transport**

The motor car was a fantastic idea. It gave people the freedom to travel, to have different patterns of work, visit remote relatives, all of that wonderful freedom that the motor car industry advertised very heavily throughout the 20th century. Except now it turns out there's a problem with pollution and the planet and mass overconsumption of oil and fossil fuels. Also, if you happen to live in a busy city, as I do, if you try getting in your car and going anywhere, you are now moving at a slower speed again, than your great grandparents did in the 1900s when they went to work on a horse. So, is the motor car necessarily a technology that has improved the quality of our lives without other social penalties? Clearly not.

- **Air travel**

Fantastic. We can get all over the world. We can fly around the world in not much more than 24 hours. The airliner is a wonderful invention. It's like a bus that takes you in the sky in super quick time to another place. Now, leaving aside pollution for just a minute, and leaving aside the extremely effective risk control profile of the aviation industry, big success story, by the way. Planes used to crash or hit each other very frequently. Now, the chances of that risk event happening are infinitesimally small compared with where they were a generation ago. So, air transport, fantastically safe, and yet cast your mind back to 911. I'm sorry, this is an unpleasant kind of example, but it's important. This super safe, effective, humanity transforming technology was co-opted, and as we call it, malignly repurposed, so used by bad actors to transform what was previously a completely or next to a completely safe object with a huge social benefit, to turn it into a flying bomb capable of destroying high rise buildings.

Every technology that invented presents opportunities, but also presents malign opportunities. And here's the thing, those opportunities are not foreseen by the people who invented the technology. That brings me full circle back to AI.

Accident Theory

Talking of aviation, there's a very useful piece of risk science referred to as normal accident theory, invented by a professor called Charles Perot. It says, where you have a large, complex system, where there are multiple interactions and complex calculations and lots of kind of this complex system talking to this other complex system. What Perot says is, it is inevitable that at some point, there's going to be an interaction between these two complexity generators, which will create something which was wholly unforeseeable.

An example is the case of fly by wire which was the first time that pilots could control the ailerons, the machinery of the plane, without physically having to move things. They moved a joystick, which was then controlling a computer, which was then talking to another computer, which was then generating the movement in the tail fin or whatever. So the actual movement was not the pilot doing this thing directly, mechanically connected, but a computer signaling to

another computer, which was then reproducing the signal in a mechanical form at the other end. Straight after flyby Wire was introduced to aviation, there was a series of spectacular accidents, including one at the Paris Air show, where the flyby wire malfunctioned and drove the plane into the ground. Predictably, the airline industry's response was to say pilot error. But after the usual sort of recriminations and public investigations and so on, it turned out, yes, it was a machine effect. It was a complex communication error, which was from an unforeseen exchange between two computers.

Now, let's talk about Generative AI. Taking that model into consideration. Generative AI is producing a greater quantity of information than the entire human collective brain has produced in its entire history. We are already at a state less than, let's say, two years into this new generation of generative AI, where the quantity of stuff coming out of the machine is by a factor of two or three, and in future, by a factor of 1015, a million, greater than anything we can produce. So the consequences of this for making sense of the world, making sense of risk, are absolutely enormous.

The only event I can compare this to in human history is the advent of the printing press. So you think your way back to the early 15th century, if you wanted to acquire knowledge, you had to borrow or write or copy a book. Now, I don't know if you know anything about book production in the 14 hundreds, but essentially, it took a team of 50 monks two months to create a book. Then fast forward to the advent of the printing press with Mr. Gutenberg and his friends. They could create the same book in maybe a day or two. Fast forward to the 20th century, high speed printing, desktop publishing and all that. I can write a book and it can be on the shelf within two weeks. I wish I could. By the way, Generative AI can write an 80,000 word novel in five minutes if that's what you want. Does that mean that it's going to be a novel worth reading? Probably not.

Bias Psychology

One of the wonderful things about human beings is we are super evolved monkeys. Essentially, we do need to acknowledge this, our brains, at one level. And if you dig into the middle of the brain, the amygdala, we still have this wonderful set of response bias activities which tell us to behave in a certain way. Fight, flight, feed, drink, chase the hostile animals away, or whatever it is. All that stuff that our primitive brain tells us to do.

Unfortunately, it also leads us astray. It makes us respond in ways which are not rational, which are irrational. Let's call them emotionally needy, affective. So, substituting anger or emotion, substituting emotion for rational decision making. Wonderful thing about humanity that it can do that both, for good. But there's a dark side to it, of course, which is we make assumptions. We stereotype. In order to conserve mental energy. We pigeonhole things into kind of easy place. Oh, that looks like this. That's analogous to that. Let's park it over here. But it can be a false analogy, and so we make a wrong decision.

One of the classic tests that I saw recently with generative AI in picture form. So Dali was asked by some people researching gender stereotyping to produce a series of pictures, rendered pictures of various occupations. And it was asked to draw a photoreal picture of a nurse. Guess what? The nurse was female and young and white. It was asked to draw a picture of a terrorist. The terrorist was a swarthy skinned person with a beard and a turban. It was asked to draw a picture of a CEO. The CEO was a 40 something, balding white man. Now, this is the AI, left

entirely to its own devices. Just asked to say, render me a picture of what you as a collective, you as humanity, or what the stuff you've read.

Draw us a picture of the thing that you have in mind when I just give you this notional idea. What it went for was not just the stereotype, but an amped up, reinforced stereotype every single time. Now, if we're serious about cognitive diversity, tackling discrimination, opportunity social mobility, education, all that stuff. General AI, I'm sorry, people, is not going to help us. It's actually going to make it worse rather than better.

The industry is outstripping the controls

There are some fascinating stories from the early days, all we now think of as sort of mainstream technology. One of my favorite ones is the Great Western Railway which was the first express railway designed as part of a fast transport link from London to New York with a fast steamship. The two engineers who designed it, Brunner and Guage, used to take the locomotives and race them up and down the tracks on a Friday night, and they probably had a few beers as well because they had this wonderful new high-speed technology that the world had never seen before, and they were basically kids. They were both just having fun with the tech. And it was only when they ran over a workman and killed him that they thought about, you know, signals and brakes and guard rails. We are at that railway opening moment with a generative AI right now, somebody's going to get seriously hurt. There is a phrase that is used in technology that really alarms me. They talk about seeking forgiveness, not permission.

As somebody specializing in human sense making and the regulation of human risk taking around that, I will say with utter certainty that there is going to be an AI generated disaster, a catastrophe of some kind, within the next two calendar years. I'll tell you why I'm so confident of that. That the history of regulation, the history of risk controls and audits of all of these things is a history of catastrophe followed by regulatory catch up. And I say, as one of the advisors to the government program we have in the UK on the future governance of AI, what worries me with that program is that most of the representation in the think tank is from the developers and not from those like me who are concerned about cognitive and social harms. I'm absolutely convinced that the industry is so far ahead of the control process that a catastrophe is at this point, inevitable. Please make sure that you are not the organization that suffers it.

Cognitive Pushback

Humans are well adapted to survive in hostile environments, to make snap decisions based on too little information. And unfortunately, in modern society, those snap decisions can lead us astray. Or more importantly, people are capable of being led astray by people who understand how that circuitry works. The only example I found of a collective pushback against this is the nation of Sweden.

One of Sweden’s neighbouring countries is not noted for its democratic uprightness and indeed sponsors, so called, troll farms. The Swedish government's view of this is that the best thing to do is equip Swedish high schoolers with cognitive skills with, you know, rationality tools with constructive challenge and teach them to critically assess every piece of information.

Unfortunately, in the generation of social media, technology designers have designed social media platforms that have turned off these cognitive skills. Social media works by switching on your amygdala, your fight or flight, your affect circuits and switching off your critical appraisal circuits. For as long as you're in that heightened affect state, you're going to buy more stuff. You're going to be more receptive to the advertising. We've had a decade of social media defeating these higher cognitive functions.

Nudging

Another action to be aware of is described as nudging. Nudging is the potentially benign science of framing a decision in particular way to solicit a particular response. For example, people didn't save enough to invest in their pensions, so the British government changed the nudge to encourage people and actually force people to set money aside for their retirement. So that was arguably a benign nudge.

In other things, for example gym membership. It's much easier to join a gym than to resign your gym membership. In such cases, your as your consent is assumed and then your decision framing is nudged in a direction which you might not necessarily want it to go. AI, again, amplifies this sort of manipulative element of commerce.

Mistreatment

IOSCO gathers all of the financial regulators around the world where they share, among other things, behavioral research and effective levers for prosecution of misconduct. What they have discovered is non-financial misconduct which are patterns of behavior where people are casually mistreating each other. These are very effective early indicators, early predictors, even for more serious misconduct consequences. We are also seeing this with AI in that the proponents of it are running so far ahead of the consumer capacity to make sense of what's going on.

Over-optimism

This technology has the shock advantage of surprising people. We're selling people this sort of drug of optimism around the new technology and we're inevitably, as with abuse of a drug, we're going to be let down at some point when we realize that this is not entirely real or it doesn't come without consequences. I think, can we agree that the law and regulation consistently throughout human history has underestimated forces of nature and human nature and technology? It's overestimated our capacity to get ahead of problems and frame them and head things off before they go wrong. And we also consistently mis-predict how people will respond to the advent of new technologies.

Critical Thinking

1. Semi-intelligent machines can do the jobs that we find boring and do them better well, clearly, we're already there.
2. The second stage, which we're rapidly approaching in the next year or so, is where again a semi-intelligent machine can hold a fluent conversation with rational arguments and good analytical conclusions. We're not quite at the analytical conclusions, but we're getting very close to the good.
3. The third stage, which is the one everyone worries about, is the existential crisis, is when the machines are self-informing and can overtake us. For example, there have been some

slightly alarming training dialogues where people have had a conversation with a generative AI where the answer was to eliminate the human race because we are the main producers of all the anthropogenic pollution that's causing global climate change.

Going to go back to Sweden again, the Rolling Institute in Sweden has had great success in transforming the national education program such that every 16- and 17-year-old in Sweden learns critical thinking skills. A little bit of mild scepticism is good for you not to be living in this sort of social media induced state of constant approval. We have to manoeuvre a generation of young people to think critically.

Think for example on Social Engineering. As an analogy, you invest $1000 on an incredibly strong metal front door with three locks in it to keep a bad force out when what the hacker knows is all they must do is talk you into leaving the key in the locks. Far too many organizations are investing more and more money in a stronger front door, but not then training people not to make these elementary cognitive errors. It is not the technology. It's the human social engineering element that is the biggest risk by far and is also the biggest thing that generative AI is good at.

Conclusion

I trained as an auditor with a classic Big Four audit firm. One of the things that fascinated me about that work, that still fascinates me today as a behavioural researcher, is this ability to ask people challenging questions to put them outside their comfort? I think the role of the audit profession and the role of the governance are consonant here in questioning truths foraging for true quality information we need to make a robust decision. Critical thinking is the skill which the audit profession and a well-informed governance environment can bring to this. If this is done with sufficient energy, then I think we can rope in this monster called Generative AI and actually it will be a force for good. I'm not optimistic right now, but it is within our grasp if we make that effort.