HIC 10 - xRisk and gradual disempowerment. Should we be quite so worried, and if so, what should the UK do?

Meeting at Nesta on Thursday April 24th
Anonymised Meeting Note for the website

Speaker1 8:21
Here are a few introductory slides to get the conversation going. Let's start with Musk. The idea for this session came from a session with him at the Abu Dhabi World Governance Summit. For those of us who've been in public service innovation for a while, a lot of what he says on the DOGE stuff seems right. Best example? The Federal government really can only retire 13k people per month because of the physical constraint of getting paper records out of and into an IronMountain facility. Digitisation has been running since 2014. "How is it going?", he asks. "We're at "B"", is the answer, which he interprets to be a grade for the programme - "good". No. They have got to the letter "B" in 11 years... He is right. This is ludicrous waste of human capabilities. From there, he goes on to describe a future of humanoid robots, of the end of money, of the age of abundance... And then he veers into topics that are distinctly less comfortable. DEI, for example, and how it is being baked into AI models. He thinks that in a few years' time, if nothing is done, male leaders will be executed by DEI terrorists. DEI is simply a form of lying, he thinks, and then he goes into this stream of consciousness about a bimodal human future.

Perhaps as expected, it is all a mixture of sensible, interesting and crazy. I bring it all up in terms of introduction to AI harms because one question I had sitting there was: why is Musk taking 45 minutes out of his time to be here? Abu Dhabi announced the purchase of a Loop for itself. A few billions of sales for Musk. Of course you show up for that. The point is that there is a political economy to all of this. When you go to the Whitehouse, the anti-corruption is so fierce that all you are ever given is $4.99 Hershey chocolate bags, because they fall under the statutory limit on gifts of $5. But here you are with a huge gift to Abu Dhabi - a headline meeting with Musk - in exchange for a purchase worth a few billion. Who is policy being made for is a central question in this AI risk story.

We have had a discussion about a "paperclip scenario" of xRisk. But the article [LINK] many of us have read [one of the authors are here today] is about gradual human disempowerment and is much more closely linked to the sort of political economy exemplified in exaggerated form in Abu Dhabi. The paper's argument is that there are many formal and informal _humanising_ processes at work in the economy today. A government does not do this or that - say pass this labour law - because those making the decision are human. Or "if government does this, then there will be a human backlash to deal with. Maybe they will withdraw their labour". But with more and more machine choice, those constraints fall away, or become less salient. For example, when thinking about Nudge policies, I often used to debate the idea of voting recommendation, or, going further the setting of default votes, as a way to combat falling turnout. But in a way, the whole point of voting is for people to do the choice work. Imagine having machines do that kind of choice work. We get to dark place quickly. And the point in the

paper is that the more we do this, ever incrementally, the more the decisions might tilt in favour of the machine's interests and away from the human interest. I like to look at this through the lens of Acemoglu and Robinson's "Narrow Corridor" - progress has been the result of increasing the state's capacity for coordination and conflict resolution while resisting the tendency of high capacity states to abuse their power in their own interests by adding to the capacity of counterveiling, civil society powers. We can view this now in terms of machines versus humans, rather than States vs Society - where are the counterveiling powers to the machines that will keep them withing the narrow corridor?

So how about the optimistic angles on all this. The first is about focusing n labour-augmenting rather than labour-substituting uses of AI. Look at where labour does a bad job, rather than were AI can do a cheaper (good) job than is currently done by labour. A good example is in clinical uses. Yes, you can train AIs to be better at looking at scans than radiologists. But the differences are minuscule. How about something that the clinicians are bad at, like predicting who is in mortal danger imminently - obviously an important input into hospital prioiritisation and management. It turns out that you can train AI to do the latter, which humans tend to be bad at. That avoids the sort of problems highlighted in the paper we are discussing, of humans being side-lined. So maybe a policy of chasing labour-augmenting uses of the tech. another example is in sentencing policy. It turns out that one of the factors that seems to determine whether you get sent to jail or not in US courts is the shape of your face. You could cut prison population without significant impacts on crime by getting rid of the "full face bias" in judge sentencing. This can be seen from the result of AI modelling of who gets sent to jail and who doesn't. We don't want to substitute judges. We want AI models to help train judges and make them understand what it is they are doing.

Finally, let me end with this classic behavioural science slide about nudges to get you exercising. We don't like going to the gym. So let's make it easier - we put an escalator to get you into the gym. That's funny, obviously, but also has a profound truth that carries over to AI risk: system 2 thinking is hard work. "If the bat costs £1 more than the ball, and the total costs £1.10, what does the bat cost?" 90% of people let system 1 answer that one incorrectly as £1. That is because getting system 2 get to the right answer is hard work which the human brain avoids if the quick & intuitive answer seems like it has a good chance of being good enough. The Flynn effect, by which human IQ increased throughout the 20th Century by substantial amounts, has been going into reverse for 15 years. The more intelligence we put into the world, the less we need to carry on our shoulders. That is perhaps the most telling effect of the rise of machine power against human power.

So, as policy makers, what are the solutions that we can lean into?

Speaker2 23:49

Can I just emphasise that last point. Ilya Suskever (ex OpenAI CTO) points out [LINK] that the basic research hypothesis of the LLM movement is that anything that _any_ human System 1

can do will be done by a non-reasoning LLM. And we know how remrkable some human System 1s are, even in the kind of intuitive maths of the bat and ball problem. So what we are doing with LLMs is putting brilliant System 1s at our disposal everywhere. So does the basic Kahneman & Tversky model suggest that System2 will be less and less solicited? What does the model say about how the brain decides which model to employ?

Speaker1
I am meant to be writing a book called "Boost" at the moment, which is precisely about the sorts of things that can be done to invoke System2. One nice example is the "Control Shift" programme that was deployed in a very violent part of Chicago as part of the "becoming a man" problem. The cognitive intervention was, in crude terms, getting men to actually pause before resorting to a violent solution and to think: "is there a way to solve this problem non-violently". There are interventions that work.

Speaker3 1:27:14
We at Poetry Pharmacy Live are working on Ai applications that connect people to poems to enhance their wellbeing, mental health, etc. The world of therapy bots is highly unregulated, but there are some amazing possibilities here. Just as the novel, the poem ,the magazine, the newspaper plausibly shaped the modern mind, so these forms will shape us. And transform us. But the novel did a good job. Perhaps the optimistic take is that this technology, properly regulated, naturally, can be a human enhancer in that same sort of way.

Speaker2 29:28
A question for the LLM experts in the room - is it a good shorthand to think of the non-reasoning models as System1, and the push into reasoning models as being an attempt to create System 2 machines?

Speaker4
Yes. LLMs still make very "System1"-style mistakes. You ask them to count all numbers from 1 to 50 except multiples of 13 and they will do it, exclude 13, but then "forget" what they were meant to be doing and simply put 26 back in. So-called reasoning models do try to get cross-checking and planning into instruction-following, which is much more System2-like.

Speaker1
Yes, so to repeat, one of the questions is whether we are going to build these models to be augmenting and boosting. In the case of the judges, for example, do we want to train models to tell the judge to "pause and consider" at the right times, to prompt the human into awareness. That would be much better than going with what most AI training is at the moment, which is

improving on but reinforcing the default behaviour. And if we reinforce default behaviours, we run the risk of the "boiling frog" problem of gradual human disempowerment".

Speaker4 33:37
Let me expand on that. There are already these very large systems that aren't doing exactly what we'd like them to do. We already know that the economy, it does a lot for us, but it also sort of tends to go a bit hard on the marketing and states, if left to their own devices, have a tendency to sort of slip towards being a little autocratic. But we have these control mechanisms in place.

The way I see it, there's a two-fold problem. The first of them is that as AI becomes more sophisticated, there will be strong incentives to hand things over to AI and hand human influence over to AI, because you have an AI that's better at your job than you are, then it would be somewhat irresponsible to not have the AI do some portion job if you're an employer, and you could hire a human or you could pay For a subscription to begin with. It might just be because the AIs are cheaper, but eventually they might actually be better. So just at the first step, there'll be reasons for us to hand over influence to them. And then the second step is, as we lose influence, we lose these control mechanisms, and we maybe don't appreciate how deep the control mechanisms are. Why? Why can I get coffee? Because there's this enormous supply chain that is invisibly combining enormous amounts of human preferences to build this elaborate infrastructure that is giving me something because I have money, because I'm doing useful work that people are willing to pay me for.

If you look at things like extractive states, rentier states, when they don't need human citizens as much, they just have a tendency to slowly move towards being more autocratic. What makes states nice to people? Well, it's it's because the citizens produce tax revenue, and it's because the citizens are necessary for military dominance and so on. This is a very cynical story, but I think it's at least worth entertaining the possibility that it's mostly these things that keep states in check. And then you have to ask, what happens when we enter into this racing period where every single company is asking itself, how much do I want to hand over power to AIs? Every single state is asking, How much do I want to replace my military with drones? How much do I want to have AIs making policy decisions? And once you get into that race, there's just going to be a very strong competitive selection pressure that pushes to more AI adoption and more sidelining of humans.

And if this process goes on far enough, then eventually humans basically just become like a drag. They become a limitation. And any entities that try to prioritise humans are sacrificing some competitive advantage to do so in a very general sense. And this is, this is really what I worry about. I don't know quite how far it will go, but I think it's going to go progressively further in this direction. And even on a small scale, it will just leave humans with less control of the world and being less represented. And if it goes very far, we get it up in a situation where we have very little power, and there are these independent companies that are run entirely by AIs, that are simply harvesting money and spending it on some slightly inscrutable goal. And as that happens to states as well, I think that could be very messy.

That is the essence of the worry presented in the article I co-authored [LINK].

Speaker5
I have been quite taken by the AI Snake Oil blog [https://www.aisnakeoil.com/]. And in particular their recent paper 'AI as Normal Technology". There is a lot of hype, which of course, is part of the narrative and cultural strategy we have been talking about. But perhaps also that means that more ordinary regulation will work OK. Take autonomous trucks. Yes, you can get trucks to drive on the motorway. But truckers do much more than that. They talk to customers, they sort out logistical complexities, they apply judgement to complex situations involving business, events in the world, people. That means that people will stay in the loop, and so the gradual disempowerment hypothesis looks less likely.

Speaker6 37:10
So I did read the paper at the time. And so I just want to make two points.

The first is, I think it's extremely valuable, very strong paper. Remember, we're only going to get more powerful models, and we do not currently understand how the existing ones work. No one can tell you what they will or will not do. They can't tell you how they work. Companies go around claiming they do decent interpretability research, and they kind of vaguely understand it. It's not true. The people making this technology do not know how it works. So when you start handing over big stuff to the models to make these kind of decisions, it is like going down a one way road. We actually don't, technically, as humans, have the capability to understand how to even come back from that. So that's why I worry. And I think the economic incentives that you've set out are extremely strong. We're seeing them playing out, as we speak, in the race with the labs, and we're seeing models being thrown out without proper safety testing, because the economic incentives are to keep going. Now that's just the labs, when that gets put out to the rest of the economy, and everybody starts delegating, the problem will be magnified and accelerated.

(I don't put government in this category because I think government's too crap to be able to use AI properly at any speed. One of the battles I had [NB not to be attributed] in government was trying to get any use of AI.)

I'm going to make one other point. I'm afraid this is a very pessimistic point. So please forgive me, when I read your paper, I was kind of pleased, because I think it's the best case scenario, in the sense that it does not have an acute crisis event described - lots of people dead, humanity still muddles on in its own way, even if it's maybe not in charge. I think that there is a very, very critical period over the next couple of years where the capabilities of the models become very dangerous in terms of misuse risk - bad actors taking the models and doing things that you don't want them to do. I've worked with the National Security establishment inside the UK Government on, for example seeing if the models can help you build biological weapons, chemical weapons, etc that you wouldn't be able to do on your own. We're at a critical juncture,

probably over the next 18 months, where the offense/defense balance on cyber attacks is dangerous. Offence is always ahead of defence anyway, but we're at a critical juncture where the AI is going to really push that out of balance in a harmful way. Defensive capabilities are not even the races, sadly.

So there is a critical window for mis-use risk. But then there's also a critical window for AI-endogenous model-led risk. What happens when they're self replicating, they're exfiltrating themselves. They augment their own capabilities. Things just start to get very, very weird at that point. And of course, it is a little bit sci fi. But not so much. I was in Silicon Valley the other day I went for a walk with one of the top researchers in one of the labs, and they are literally talking about how they need to find another job because the AI is doing 40% of their research work already, and they believe at some point that their job will disappear entirely. When AI tools can replace humans currently values in the £m per year for their AI building skills, we get to a place where the pace of improvement will be huge. The exponential we are on is very worrying, and "gradual disempowerment" would be , against that backdrop, pretty OK.

Speaker2 42:01
Let us go back for now to the non-catastrophic case. There is a very powerful and attractive argument by David Runciman who talks about the corporation and the bureaucracy as being "in-human" super-intelligences that we have had to contend with at scale since the start of modernity, which he places more or less at the publication of Hobbes' Leviathan. So we have some knowledge and practice of how to do this. I would like to bring in Martin at this point, with your experience of being an economic regulator in many spheres. The modern theory of economic regulation is strikingly similar to the alignment problem, and also the challenge that Runciman describes of our battle to civilise the Leviathan. The idea is that the firm is more knowledgeable than the regulator, is more powerful in the sense that it has the capabilities to do things that humans want (like deliver power to homes) and is endlessly devious in its own interests. The knowledge asymmetry gives the firm power over the regulator, whom we can assume, at a first approximation, acts for "humanity", or at least that part of humanity that is in their remit. The regulator can play around with incentives (rewards) to the firm, but ultimately the firm will game this system to the maximum extent in the interest of its shareholders, management, etc.

Martin, with your wealth of experience as a regulator and as an economist, thinking about regulating these entities that are unbelievably powerful, opaque, whose course is difficult to change... Are you hopeful, given what we've been able to do in the public sector in controlling for profit companies, that in some sense, those same tools can be used to control these new inhuman intelligences?

Speaker7 44:45
Well, I think you're absolutely right that the whole problem in regulation is that if you sit in a regulator's office, then you don't know 2% of what's actually going on in the company. More particularly, you know 0% really about what could be going on if they were behaving in different ways, if they weren't being misled by faulty incentives.

So that's a that's a huge problem, which regulators have tried to solve by trying to get more competent at asking the right questions, by hiring their own staff who can sort of figure out quite a lot of stuff - do technical analysis of the same kind that might be done within the firm, particularly when you're discussing things like investment projects and things of that kind. And that's the way it works. It quite honestly doesn't work well, but it probably beats the alternative and some regulators do much better than others. There's also the rather grim problem of regulatory capture, where the regulators are essentially bought. There's a famous case for that recently where the the chair of a state regulatory authority has actually been sent to jail for 20 years because he took money from the local gas company. That's an extreme example. But there is a kind of regulatory capture that goes on just simply through the process of socialising with the people who are in the industry, and the revolving door. If you think about the digital platforms, one of the remarkable things is that it took something like 20 years before anybody tried to regulate them seriously. And by that time, all sorts of things had happened - many beneficial - but the detrimental things are just adding up and multiplying. The European Union has managed to start regulating pretty effectively. In other jurisdictions, it's not going very well, but if you compare, for example, the EU approach to digital platforms, where they didn't do much, but then what they did looks as if it might be pretty good. But if you look at what they're trying to do with the same mindset that I've described with AI, then obviously they are far more in the dark. And indeed, everybody's in the dark because of the problem of alignment that you've described in your note is a very hard one - you are dealing with an entity perhaps even better than the regulated firm at deception. Classic regulated firms are very good at deception. That's the instrument they use to get their way. With AI, it seems even harder to verify what's actually going on. And so when I look at what the AI Act says, it's encouraging that it's focusing on the risks, and it's trying to produce a kind of classification to risk so that if you're in Category A you're subject to much greater surveillance than not. But then the question is, do they actually have any capability to understand dealing with the profit maximizing humans, and their much more professionally deceptive AI systems.

How are you going to do that? And so, to be honest, I tend to veer towards pessimism about the feasibility of making much progress with this.

Speaker8 1:25:39
In regulation, I am a poacher turned gamekeeper, and I agree with Martin's position but I am more optimistic we can do something. I think you've got to abstract yourself from the technology and ask yourself, what are the outcomes that I want to achieve, not the technology that is being used to achieve them? And once you do that, and you can look back historically at things like the formation of the BBC, things that are embedded in the enterprise act that say there are public interest outcomes that I want to achieve and pursuit of that public interest overrides commercial incentives, overrides competition law, and gives the regulators the ability to pursue that public interest, overriding those other foundational parts of law, then you have some power to control. So here, I would say the challenge for us is to say, what is the positive vision of a consumer and citizen outcome that we are looking for, what is our hopeful vision of the future,

and then putting legal powers in place that override competition, override commercials, override other protections, and give the regulator the chance to achieve those outcomes. That is how we have used Leviathan against Leviathan.

Speaker9 48:53
Just a brief tempering voice to that. If we are able to get AI into government, then we might hope that regulators can benefit from the consultancy of AI models, which can be terrible liers, but can be very good at seeing through the lies as well. So that's one hope. And to broaden the question a little bit, the slightly optimistic perspective is that much of the malevolence we may expect from LLMs and domination can also be tempered. So some of the insights they bring can improve governance, improve ways of doing democracy.

There's an incipient movement in AI safety, also called the defensive/acceleration (def/acc, [https://forum.effectivealtruism.org/posts/tc7z9tA55i2ScZS7Z/194-defensive-acceleration-and-how-to-regulate-ai-when-you]). What we need to do now is start deploying systems that produce defensive solutions. Where's the AI at the table here helping us solve this problem? There are people taking a more optimistic route and saying these AI's have got to be on our side too. For example, recently a Blackpool hotel was repurposed into a space for AI researchers to work on these sorts of solutions [LINK].

Speaker10 50:30
I worry that if all the junior staff, the junior lawyers, junior consultants, junior civil servants who learned on the way up, through the conversations with the departments and everything else are those with the jobs most replaceable by AI. So slowly, the ability to train new humans to have the ability and judgement that can direct AIs is being lost precisely because of the early successes of AI. Is there not a further worrying disempowering dynamic in this?

Speaker11 1:22:41
Some of the subtle feedback loops we should consider: AI is already being used to make food more addictive, for example by micro-targeting messaging about ingredients that you are are sensitive to. More AI, more addictive food. Then the neuroscience behind this: if, as a child, you grow up using addictive products, your amygdala over develops, which is the part of the brain that does fear and stress; then, when in your teens your prefrontal cortex is growing (which is the part of the brain that develops control mechanisms), then the amygdala dominates. So addicted kids lead to adults who are less able to control the use of their products.

Speaker12 1:29:16
Putting AI into government ... well, we might be worried, but it is hardly as if the status quo is so brilliant... But I agree with the chill involved in allowing AI first to change my preferences and then to act on those preferences by choosing my government... Let's not go there!

Speaker2 51:48

Can I pick up something else Martin said - about the relative success of the EU in its platform regulation, perhaps not mirrored in its AI regulation, and raise a point about current geopolitics. One of the very nice things in the paper under discussion is that it takes a "decentered", structuralist perspective - it identifies the "mega systems" in society of economy, culture and government; it then looks at how capital, labour and now AI have their interests felt in those systems. Now, a few stories have struck me: Vance linking tech and platform policy in Europe to the US defence umbrella (you need US values to benefit from it, for example on free speech); the US apparently saying to India that a trade deal will require that US platforms are allowed into parts of the India Stack that they have been (quite consciously) excluded from; and the general cold water that has been poured on EU and UK attempts to deal with Silicon valley market power in the platforms - the consensus seems to be that it is America's job to deal with the power of American firms. In view of all of these, and taking the structuralist approach of Speaker4's paper, should we conclude that AI interests have already captured the White House, and that, DMA notwithstanding, we are a bit late?

Speaker13 54:43
I'm still stuck on the idea that Musk was worrying about AI being a liar...

I thought one of the most interesting documents I've read recently was openAI's submission to the US AI strategy. And the reason that I found it so interesting was that by about page four or five, I realized they were describing the East India Company. And now we understand. So what we understand is we are exchanging our culture, our technology, our systems, our rules of engagement, for a security umbrella from the US government. And I think that everybody in this room has to start a little bit with that. Lots of leaders are giving up gold for glass beads. We know the model, and yes, we are late.

Speaker14 56:43
I would just like to pick up on that briefly. I agree with the East India Company parallel. I think it's quite instructive: you are either the one or two people able to impose the rules, or you are the rule taker. Then what do the UK's priorities actually become? We all care passionately about AI safety; that's why we're here. But the people who will actually make a lot of the decisions will be a few CEOs and maybe one or two presidents or premieres. So what do we do? You know, do we try and build something else? Do we have some advantage?

Speaker15 57:37
Back to being, you know, gloomy again. I gave up doing AI type stuff in about 2022, because I was realising I'm having the same conversations again about a problem that we have seen, in many ways, many times before: powerful firms that addict people and society to the goods that make them rich. Tobacco companies are still killing 8m people a year. How is that? We need to think about stopping inescapable availability & to do that we need to stop the lobbying, the shaping of narratives, etc. Those tobacco companies are killing people now. And we've known about it a long time. And we let them continue. So what hope do we really have of standing up

to the AI companies? This is not something that our society does well. This is what I am describing in our addiction economy project.

Speaker2 59:13
To be fair, even if we are late, we are earlier with AI than we were with the Tobacco companies that got themselves entrenched before the tobacco risk was all over policy circles. Philip Morris and Procter and Gamble and Nestle were so well entrenched by the time that we woke up. Whereas with AI, the moats are not yet quite dry on the mortar.

Speaker16

Let's remember that openAI, on its original release of ChatGPT, said: "this is to shock humanity into what is coming. after the release, we can talk seriously about what to do". And that was the time of the moratorium letter. Yet now, Sam is saying: "there is no appetite for slowing down"(in a strangely self-exculpating use of the passive tense). So two points: what is the crisis that will actually put regulation back on the policy table, and when it comes, shouldn't we be ready with the proposals in a top-drawer?

Speaker17 1:30:11
A few thoughts - perhaps the crisis will come from finance, where huge funds (eg AQR) are saying that they are AI-driven. The managers are saying they don't understand how the AIs are achieving the returns, but who cares! Sounds very 2008-ish, no? Other sources of crisis - extremely unsafe images and videos being generated "for clicks" which are clearly teaching the models through reinforcement learning what they should be doing more of. Finally, should the UK not be doubling down on quantum research, something that might give us a place at the policy table?

Speaker2

One of the compelling things in Speaker4's paper is that it presents a future we would not choose that does not involve a crisis - we are gently boiled frogs...

But could I bring in Henry at this point - the AI2027 scenario [LINK] was fascinating in its description of the geopolitical context of US/China rivalry. It suggested that the promise of AGI and the asymmetric advantage it will confer is so great that unrestrained development has become a superpower imperative. So Henry - you were closely involved in preparing the Bletchley AI Safety Summit and the process that it has put in place, with meetings in Korea and France since then. Are you still hopeful for that sort of process, and will it survive geopolitical competition?

Speaker6 1:03:27

It's very hard. I spent a lot of time engaging with the Chinese in the run up to the Bletchley summit and also to the Seoul Summit, and a little bit in the run up to the Paris summit. I was out in Beijing for Christmas last year. My view was you had to have China at the table at Bletchley, otherwise you're not actually having the discussion. Because they are the second largest AI superpower in the world. Unfortunately, there were people in the UK Government who didn't agree with me, who didn't think China should be at the table on anything, certainly not on this technology. There were people in the White House who also thought that. There was a battle to get them there. But I think it was the right thing.

I've updated my views a little bit on China, given the DeepSeek and the open source stuff. But my previous theory would be: there is certainly a section of influential academics in the Chinese system who understand the catastrophic risks and are worried about them. These are people like Andrew Yao and the Singhua University gang.

The cynical take would be that when you're talking to those people, you're talking to the marketing department of the Chinese government. But I think these people had genuine AI experience and knowledge, and they were worried. And they were certainly making the case to the Chinese government that they needed to be worried. The Chinese have more regulations on AI than anyone in the world.

DeepSeek had to put their model into the local regulator, and it took them three months to approve the model and then they could release it. They actually have a licensing regime. Of course, it is basically only focused on whether the model will chat about Tienanmen Square or Winnie the Pooh.

One of the missed stories from the French summit was that the Chinese actually announced and their own AI Safety Institute, which I thought was a good thing, but the ship had already sailed on that in the sense that the Americans, and this was under Biden, not Trump, didn't want an international AI Safety Institute network. The Americans then came up with their own separate one. Then you have Singapore, you have Japan, you have some other countries involved. And there was a network, my view is the Chinese should be in that network. But the Americans did not agree with that. The only cross party thing anybody agrees on in America is hatred for China. So I am worried. And I think that international engagement stuff is in a very tricky place now. However, on the plus side, I think Trump loves deals. And if there's anybody who would actually make a deal with China, fundamentally on all of this, he might be it, and I think he might have a better chance of doing that than Harris would have done. However, I go back to the point before, which is that we don't understand the technology well enough to actually even be able to do a deal. At least during the Cold War, you had scientists all agreeing, having a shared understanding of what the science was behind nuclear weapons, so you could then have a deal. At the moment, I don't think we technically understand the technology to be able to really even have that kind of deal. I agree with the def/acc programme - the only way through on all of this is using the AI to help us get through all of these problems. I think it's the only way. We, the UK, spend 75 million a year on the AI safety ship. That's more money than

anyone, any government in the world spends on AI safety. But that is minuscule compared to the budgets going into research without regard to safety.

Speaker18 1:32:12
Let me just say, if I may, as an outside observer that one of the critical things as we look at humanitarian vulnerability increasingly has to do with AI; if you take a look at the international institutions that are responsible for humanitarian assistance have a need to get involved in this. I think we have to push this argument and debate at a very practical level at the United Nations for the everyday tasks of the UN.

Speaker2 1:08:23
Henry, you know the American people and institutions involved in this and you understand the nature of this competitive moment. It sometimes makes me think about the time between the first use of a nuclear bomb, and Russia getting that the nuclear bomb. There was that narrow window of unipolar dominance, and all sorts of people were asking: "What should it do with its power?" Very strikingly, Bertrand Russell threatening the Bolsheviks with nuclear annihilation. Once Russia gets the bomb, Russell becomes a founding member of CND. In your view, is the US being motivated by the sense of what that moment of dominance could give them? Is that how we should understand it? And if so, what is the role of UK policy in all this?

Speaker6 1:09:47
Well, we're not going to suddenly build a massive LLM. We should start building data centers, make all of the moves necessary to get cheap energy into this country. So you have some sort of leverage on the AI stack, because you haven't got the AI companies. You're not building the foundation models, so you've got to have leverage somewhere else in the stack. So therefore that's what I would do, and then I would carry on doing this role where you can try and bring the rest of the world together and be the bridge between China and us. Many people were upset about Starmer not signing the Paris documents, but I think it was actually exactly the right thing to do to try and buy good will with the Americans. And there was nothing on safety in the document anyway (and the French don't care about safety), so it didn't matter. But what it did do, and I think it has allowed him to play this role that he's doing on Ukraine, and we have to keep trying. But it's very, very difficult.

Speaker2 1:11:01
We may have overdone the pessimism of the intellect section of the meeting, and we have too little time left for the optimism of the will. Chris - can you talk to us about the potential for AI model evaluation, and of using AI models to measure what is happening in the economy, and so keep a handle of the risks of disempowerment?

Speaker19 1:12:15
At faculty, we do a lot of pre and post deployment, like safety testing. We pretend to be bad people doing bad things, and scaling that across a lot of deployed systems. We are learning to control risk in a very practical sense. On the economic impacts, Anthropic have released their

economic index [LINK] which does keep track of trends and things we have to worry about. But with open source and the proliferation of models, we will lose that data. And there is a race to the bottom - OpenAI published their updated safety framework to say "we will completely update our model requirements if someone releases a more dangerous model; we can come down to wherever the competition is". That does not leave me terribly hopeful. Will there be shorter term risks, not catastrophic risks, but things that go wrong, that make us stand back and take stock? It is not clear to me where the near-term feedback loops will go. Will companies start to see degradations in corporate outcomes when they have replaced 40% of their workforce with AIs? Or will it go the other way? We do not know that yet. And to repeat, tracking these things is getting harder as the space fragments from a small number of cloud-based offerings.

Speaker2 1:14:24
Can you say anything about the degree to which there is something different in this set of companies - openAI, Deepmind, Anthropic - in being so influenced at their inception by universalistic value - Effective Altruism and also, for Hassabis, the strong influence of his Christian upbringing? Does this change their behaviour or how they might be regulated or the degree to which they can be trusted? It is, after all, very different from the minimalistic libertarianism of "do no evil"....

Speaker7 1:16:00
Surely they're about as ethical as the digital platforms. And indeed, they are the digital platforms.

Speaker19 1:16:34
The tech majors we work with invest a huge amount into safety - they have to because they are global companies with brands that could be trashed by errors. And being global requires them to look at things from any perspectives.

Speaker20 1:17:26

I've read a paper I find much more chilling than yours, and that's that's the recent DeepMind paper about so called experiential AI, moving from analysis of data to streams of experiences from which the AI will learn without having to be limited by human input, the idea being that by allowing human feedback to play the role that it has played historically, we've taken away from AI the ability to develop its own reasoning, so it hasn't had the ability to develop non-human reasoning yet. And the DeepMind paper is incredibly positive about this, saying this is the next big thing. This is what we should all hope for. Then I get quite worried.

One way is through regulation and control - control of the purposes to which our data is put. When we're talking to stakeholders about why they should share their data, what would make them comfortable with sharing their data? The governance is absolutely key to that. Part of that

is the overall purpose for which the data will be shared, but also who's going to access it, who's going to use it, and for what purpose? And that's really the key thing. One way that we might be able to improve this is through data governance and control.

Speaker21 1:19:16
Listening to the conversation. I'm both British and Kenyan, and I have assets in Kenya, assets here, family in Kenya, family here. And I couldn't help but think about AI and its use for good with my Kenyan hat on, and how I see and many citizens in Kenya see it as an opportunity to replace an incumbent, existing system that's broken. You perhaps have more of a glass half full perspective on the opportunity, even though, with my British hat on helping the ferocious tech players, for example, even grow across Africa. Yes, there's a realization about how, in many ways, my Kenyan being, is outsourcing the ownership of the data that analyzes and confirms my ownership to a business in California to oversee it. Well, this is the digital vassalage point, which is frightening, but at the same time, when you've got broken systems, anyone that comes selling you gold, even fools gold, is attractive.

Speaker13
I think one thing positive is that we've got to stop talking about data, just as if it's all one thing, because there is human experience, there is there is work, there's behavior, there's movement in the world. I mean, there's so many different things. And I think that one of the big policy wins that the AI companies got is to reduce everything to data so that it is all even, and if it's all even, then we don't know what to protect. And I think I've learned that sort of double down. Learnt it in child harms area where not everything is equal. And then I've seen it again in this campaign about creative copyright. And I also just want to say that I work with a group of kids from Kenya, and one of the things that they say is very much there is a huge pressure to join, to be in, to be on, to see AI as the future. But they say they haven't done anything for us. They haven't done anything for our childhood and the adult world has outsourced all questions of safety at the level at which we're talking about it, but also our individual safety to us. How do you explain that to them?