Transcript of ‘AI, Data Centres, and The Great Energy Problem’

Season 3, Episode 3, Transforming Tomorrow

[Theme music]

Paul: Hello, and welcome to Transforming Tomorrow from the Pentland Centre for Sustainability in Business.

I'm Paul Turner.

Jan: And I'm Professor Jan Bebbington.

Conversations about AI are everywhere, but how often do they lead to discussions about data centres and bat caves, or Meta’s official company ninjas?

If that's what you're looking for on a podcast, you’re in for a treat with today's episode.

[Theme music]

Paul: Jan, where do you stand on your energy consumption when it comes to AI usage?

Jan: Well, I have learned that I can do a Google search and then put minus AI at the end of it, and I don't get the nonsense that can sometimes produced out of that.

So I hope in doing that, I may be not inducing an AI search process on top of my existing Google searches. Because I, and I know I'm a bit of a Luddite and I'm sure that I'm uninformed, although I hope in the next hour I'll become more informed, but I just don't want to use AI. I want to use the, one of the things I've got in my own innate intelligence.

Paul: You could just go to a library, that would've a far less of an impact, you know, every time you’re planning on googling, just go outside, get on your bike, ride down to the library, look it up in Encyclopaedia Britannica…

Jan: …I think there's an optimal point somewhere, and I know this means that I'm like a thousand years old, but I think I, for me, I've passed my optimal point

of, um, how much I want to engage with new stuff. Blah humbug [in a humbug voice] humbug, humbug. [laughs]

Paul: But, but like I say then, why are you not getting on your bike? There's even less of an impact, uh, uh, carbon wise, anything wise, if you just get on your bike and go to the library.

Jan: Well, it, that's harder to tell. We might find the answer to that as well…

Paul: …could it be because there is actually some kind of level of convenience and usefulness and practicability to using programmes that, AI or not, do have some kind of footprint.

Jan: Ah, for sure, for sure. But, but the footprint as I see it and the footprint when it happens somewhere in the world where we don't see it, that's the thing that I feel uncertain about.

And where it'd be nice to know more about what happens after you've pressed the, the return button on you're, you're…

Paul: …yeah. Have you avoided all of these trends then, you didn't make a, an AI avatar of yourself in a toy box?

Jan: I most certainly did not. I don't even know what that is, and I have no intention of finding out.

Paul: You are so lucky that you've never come across that. [Jan laughs] It was one of the most irritating trends. What I found most irritating about that trend is people created these supposed like toy figures of themselves using AI.

If you imagine, you're buying an action figure and it represents yourself and what your job is, and it's meant to look like you with certain little accessories. So, if you were to be a binman, you'd have a bin with it. If you were to be Superman, you'd come with a Superman cape. If you were to be a mermaid, it would come with a few shells and a crab.

But people created them and they didn't seem to understand that action figures tend to come with lower limbs. So many of these action figures that were created cut off above the knee, so they weren't full action figures. The, the, the AI generated images, which I think this is all down to AI, cut off above the knee. So there were people out there who just had lost all their legs.

This is showing the massive limitations of what AI can do, because you could draw a better action figure of yourself because you know it would have feet.

Jan: And, and here's the thing, and maybe this is also where we go, is that the, the use of technology for various purposes is undoubtedly, you know, amazing and fantastic and, and the ability to, you know, scan a lot of X-rays, say at speed and accurately to, you know, predict disease or whatever. Really, really great.

I think the thing that sometimes annoys me is that some of the things that we're using a lot of energy for and having to build lots of data centres for are actually, you know, maybe not that useful.

Paul: Oh, no, they're totally and utterly frivolous. I dare say that 90% of use of AI is people doing something because they think it would look fun, or they can't be bothered doing it properly themselves.

Jan: Well, I don't know how this, this podcast is going to go, good listeners…

Adrian: …I just, just don't think you need me at all. [Jan laughs]

I, I can sit and listen to you two debate this all day…

Jan: …yes, but I suspect…

Paul: …we do have a guest…

Adrian: …yeah, we do…

Paul: …who's not meant to be speaking yet…

Adrian: …oh, I'm sorry…

Paul: …we'll get to you, we'll get to you…

Adrian: …no, I'm not needed, I'm quite convinced…

Paul: …I, I come across it so often though, that people talk about the benefits of using AI for this, that and the other, and I just, I have so many objections to AI. If I use it, it's because I've got no other choice for what I'm using it for.

Jan: Shall we put the AI question aside a wee bit?

Paul: No, 'cause I want to rant, [Jan laughs] but if you're telling me this isn't just gonna be a 45 minute podcast of me ranting, then yes, let's put it out to the side.

Jan: But because, it is related to, but not solely related to, data centres…

Paul: …yes…

Jan: …so tell me, who might we have as a guest, Paul? Pull yourself together! [laughs]

Paul: Yeah, I'm ranting about AI because of the amount of data it uses and the data centres that, that produce said AI results and the potential sustainability impacts of said centres.

And we're gonna welcome back today Professor Adrian Friday, who's a professor of computing and sustainability. And you remember Jan, he got to make up the job title himself.

Which, uh, the more I think about it, the more I wonder if he actually is employed here at Lancaster University, but it seems he's still here, so no one's got rid of him yet.

Um, we know he is not a big fan of the Sustainable Development goals, but we'll welcome him back anyway to talk to us about data centres and everything we can think of around that topic.

Hello, Adrian.

Adrian: Hello. How lovely to be back. [Jan laughs]

I, I still don't think you need me. If you want someone who's not gonna rant, you might, you might have picked the wrong person.

Jan: [recovering from prolonged giggling] Well.

Paul: Adrian is unique among every single guest we have ever had on this podcast. Jan, do you know why? And if you do know why, I'll be very worried.

Jan: Oh, no. I don't know why.

Paul: He's the only one who's picture is currently attached. Magnetically to my fridge.

Adrian: I would never have guessed that. [Jan starts laughing again] And, and now I'm even more disturbed. [joking] Why, why have you got me back here under false pretences?

Paul: During the summer, uh, myself and Adrian bumped into each for a wedding.

Adrian: Oh, that's true. Yes, that's absolutely true.

Paul: We, we did. And at said wedding was a photo booth, not using AI, but using silly hats. And so there is a picture of myself, Adrian, and my good partner Sabine, wearing silly hats and sunglasses and stuff at from said wedding.

And that's attached to my fridge at home.

Jan: I have seen that picture, Adrian, you'll be pleased to know.

Adrian: I, I think you should use it as the little profile picture [trying to repress laughter] for this, to be quite honest.

Jan: Excellent.

Adrian: [recovering] Yeah. Right.

Paul: He is wearing a hat that I'm really disappointed he didn't turn up in today.

Adrian: So, so I'm pleased you started off with a, with a small rant, I have to say, because, um, AI exists at different scales and AI has a relationship with data centres, uh, and it can exist in the very small and you can have it embedded in devices, and it exists in the very large and actually your, your instinct is correct if you use it to generate big things like, uh, images or series of images that become movies. This is much, much more computationally expensive than using it to generate text, which is much more computationally expensive than not using it at all. Generally speaking.

Jan: I think we're gonna get a sense of proportion. This is very good. So…

Paul: …it doesn't sound good to me… rant, rant, rant…

Adrian: …just to stop the, the generating the Lego figures, right?

But, but, but also, you know, we have to be asking the question, what are we using it for? And is that energy that's proportional with that energy well spent for society, because there are other challenges we're facing.

Jan: Yeah.

Paul: If you want to see what you look like with a totally different body, like an action figure, just go to Blackpool Pier, stick your head through one of the holes that's on the, the promenade, and, uh, you, you know, you can even see what you look like as a king, mermaid, anything.

Adrian: I, I love the fact that you're probably the youngest person around this table who's the most grumpy about this technology, [laughing] I think that's awesome.

Jan: Yeah, he's an old man at heart, I think we would say, we'll say.

So let's start with the, with data centres then, which, as you said, are linked to AI, which is kind of where we, we got excited about in the outset.

But what is a data centre and what happens in them?

Adrian: Excellent. So a data centre, um, again, these exist at different scales, but essentially it, it's a room full of computers. You tend to want to keep the computers running at a, at a temperature that makes sense for them so they don't melt down. So you have some cooling. Um, there's various ways of doing that.

Um, you don't want 'em to run out of power, so you tend to have some sort of power backup system in there, and you have to link them together and link them to the rest of the world. So you need some networking and you need some connections.

The data centre itself, you know, as they get bigger, so you've got kind of more heat to get rid of, so you need to think about a more industrial sort of cooling. So then you need sort of heat exchangers with the air or with water, uh, and you need to think about its relationship to, to where it gets the power and, and how you do the, the cooling.

But yeah, think of a, think of a massive, often quite dark room. Quite hot, full of racks and racks of computers jammed in together. Lots of noise of fans and, and things like that.

Paul: Is it just like an old desktop PC to the nth degree that had its cooling fan built in to keep it cool, only that has been writ, not just large, but gargantuan?

Adrian: Yeah, so, so they, they were optimised for the density. So, so you, you get a cab, a cabinet, and, and you, you have these special sort of builds of PCs, if you like, that you, you layer them into the cabinet and then in that lots of boards with computers and you sort of, you want, you want to get as many in as you can, because obviously the more dense that is, um, the more you compute you can get into that facility.

Jan: So in some of these, um, these computer arrays then, they'll be doing like running the banking system behind the scenes…

Adrian: …absolutely. So they, they can run, uh, I mean generally they're running software of any kind and, and, uh, there's been a, a massive shift in how we provide services and, and this sort of.

Back in the day, there were a few big mainframes where computation gets done and then there was a sort of revolution where the computation went on your desktop, and then people realised that it, it you could start providing these sort of compute services for people and sell that as a service. And then actually if you put all the compute in one place and cool it all as one, you can get that ratio about how, how much infrastructure you have to run to keep it running optimally, um, in one place.

So there's a sort of nice argument about centralising this stuff again. So it's all sort of shifted back into this, the cloud. The cloud is just the data centres and the virtualisation of all these sort of bits of software.

Jan: So the, the cloud exists somewhere on the ground.

Adrian: The cloud, the cloud has physicality. The cloud, the cloud is a physical thing. It's buildings and wires and optic fibres…

Jan: …okay. But from my perspective, it's not a building next to me. I sort of, it goes here by some way.

Adrian: …yeah, it could, it could be kind of anywhere in the world. So, so we have data centres on campus here at the university…

Jan: …yeah...

Adrian: Uh, and some of the services we interact with are there. But our email goes off to some data centres run by Microsoft. Uh, and actually there are many of those replicated across the world and we're, we're probably dealing with the ones in Warrington or Dublin or, um, you know, sort of close enough, 'cause, 'cause physical distance makes a difference because that's how long it takes to communicate with it and get an answer back.

So generally you want to interact with things that are closer in network terms to get that sort of interactive performance. So you tend to put, you know, let's say you're watching YouTube videos. You, you wanna, you want to get your data from as close as possible. To, to kind of get the data quickly.

Paul: So how big are these data centres then?

Adrian: Ooh. What a, what a question. Uh, bigger and bigger. And also, you know, they exist at a range of scales. There are private ones, there are, there are public ones.

There are, uh, the very biggest ones are sort of football pitches and football pitches. So the largest ones at the moment, uh, are sort of, uh, a hundred megawatts, couple of gigawatts of power. So you're talking football field size. You know, multi football, field size, warehouses full of computers.

So, so they, they exist at pretty huge scales.

Paul: Are they all above ground? Are they underground as well?

Adrian: They're mostly above ground, I think. Um, I mean there are a lot, there are lots of interesting experiments.

So one, one of the challenges as you move around the world is you have different temperatures outside and you want to keep the thing cool, right? So just the basic physics of that means if you put them in a cool air in the north, you've got a better temperature differential.

So people are playing different experiments. So there was a famous one Microsoft did, where they, um, had a data centre in a container, and they put it in the sea. Uh, to, to use the sort of thermal mass of the sea to keep the thing cool. Uh, and I, and I, and I think it sort of worked, although it's a bit hard to service the data centre 'cause you've gotta put a diving suit on.

[Jan and Adrian laugh]

Paul: Take the submarine down there.

Adrian: Yeah. And, and generally it's just pragmatic stuff, right? I mean, if you bury the thing, you, you've got all those challenges of getting, they, they turn over equipment, right. A computer, if you run it very hard for a very long time, will, will eventually fail and you have to replace it with new equipment…

Jan: …I hadn't thought of it, but yeah…

Adrian: …and, and you've gotta get the kit in and get the kit out, right? So the more, the more you bury it inside a, a clever place, you know, the, if you imagine the Bond lair, you know, with the waterfall going over the top and in the side of the mountain, uh, then that's gonna be a pain, isn't it?

Uh, you, you can't drive your truck of kit up them. You know, thousands and thousands of processing elements in, in a big data centre. This is, you know, there's a bit of a logistics side to it. You've just gotta get the kit there and keep it running.

Jan: And can we…

Paul: …I'm now imagining having to take an army of ninjas with you just to, to change a light bulb…

[Jan laughs]

Adrian: …oh, awesome…

Jan: [laughing] …yes…

Adrian: …better and better. [laughs]

Jan: …yeah…

Adrian: …I wonder if Meta have their ninjas…

Jan: …dear listeners, can you tell it's the end of a recording day?

[Everyone laughs]

Adrian: This is your fourth today? Is that what you said..?

Jan: …yes…

Adrian: …there you go….

Jan: …there you go.

Paul: The leaders of some of the, um, tech companies around the world I fully imagine do have armies of ninjas.

Jan: Oh, that, that could well be…

Adrian: …you know, I, I would put nothing past them.

Paul: Yeah. Don't say anymore.

Adrian: [inaudible]

Jan: So, do we know how many data centres there are in the world?

Adrian: Uh, I, I am trying to find out the answer to that question. It's, it's surprisingly hard to find the answer to that question.

Uh, the European Union has got a, an energy efficiency directive where they're asking their data centres to declare how much energy they're using, 'cause this is. These sort of questions about where does your energy come from and where does your water come from, and things like this are sensitive questions. Uh, and also, you know, things like what is your sovereign capacity or do you, you know, we are now investing data centres in the UK, aren't we, um, for doing AI.

Um, there are these questions about how many there are, but, but actually there doesn't seem to be much of a database, the data that's released in very aggregate form probably to, to sort of protect the commercial sensitivity. And there are businesses who specialise in trying to find out the answer to that question and then selling you the data, you know, so they're often behind paywalls.

So actually I, I haven't yet found a public source for this. The IEA do have a figure for how much energy they think data centres are using globally. You know, the last estimate they had was a sort of 460 terawatt hours. Uh, but predicting could be as much as a thousand terawatt hours by, uh, 2026 actually, which is not too far away.

Paul: Can you give us a comparison to the scale of that? What is 460 terawatt hours? How much energy would that be for a country or something like that?

Adrian: That would be more than Italy. That would be more than many countries, but that, that's obviously data centres running…

Paul: …mm-hmm…

Adrian: …globally. But, uh, yeah, the foot, the footprint of that is, is considerable.

And, and one of the interesting things is how fast it's growing. 'Cause people are making these very major investments in more data centres. Uh, and it's not that people aren't aware that this requires a lot of energy, but the interesting, you know, you can build data centres faster than you can build out the energy grid to service it.

So there's this interesting question about how much of our data, how much of our energy would we like to put to this purpose, I think is a really interesting question.

Paul: Yeah. I think we're a bit past the stage where all your data is stored on a 1.44 megabyte floppy disc.

[Jan laughs]

Adrian: I found my, uh, my, my floppy disc the other day. [laughs] It reminded me of a, of a much more parsimonious era where all my files were… [laughs]

[Jan is still laughing]

Paul: …yes, all your files on a three inch little bit of plastic, yeah.

Adrian: Now I wonder if I can ever find a piece of equipment to read it, of course.

Jan: Yeah. Yeah. Absolutely.

So, so with this growth in this, this demand for data, it won't only be AI, it'll be for other things as well…

Adrian: …mm-hmm…

Jan: …do we have a sense of the extent to which there's sort of this rapid development and championing of AI as a, as a, well as a business sector and as an investible sector and all those sorts of things.

Is that accelerating this process in a particular way, and do we even know if that's the case?

Adrian: Uh, yeah, I think, I think we do know that energy demand globally is, is starting to rise again. And that seems to coincide with, with the rise of, of AI and it's, it is a particular type of AI, right.

So the, once you've got, you always need data for training your models. And if you've got a lot of data, you can, you can train bigger models. Oh, and, and the sort of computational load of that is, is very considerable. So ChatGPT-4, uh, 1.8 trillion model parameters. I'm sure this is out of date, but if you imagine you've got to sort of tune this model and train this model it, it takes them, um, they get 25,000 graphics cards, uh, train it for a hundred days at a cost of about a hundred million dollars…

Jan: …ooh heavens…

Adrian: …um, just in energy costs. So I don't, I don't know, actually, I don't know if that, this was from a talk that I, I watched. I don't know if that includes

the cost of buying all that, [short laugh] that infrastructure. I don't imagine so, I think they rent the GPUs and do it that way. That's 54 gigawatt hours to train that.

And of course you don't train it once, you train again and again. And everyone else is trying to train their own models. So, so that, that demand is, is very considerable. But actually the, the sort of a lot of focus is on, on the expense of that. The, the inference cost, when I ask a query, is actually quite small for one query. The question is, what's the scale of, of queries, the number of queries you do, and does that add up?

And, and when you look at say, uh, Google's energy footprint, you find that they're sort of 60% inference cost and 40% trading costs. So actually just even though those queries aren't super expensive, compared to some of the other things we do in life, it's just the sheer scale of, of running those models with million, millions, and billions of users. That adds up.

Jan: And this is where maybe, I'm gonna try not to rant…

Adrian: [laughs] …no, don't, don't stop on my account…

Jan: [laughs] …well…

Paul: …remember all the bad things he said about the SDGs. Go on, go on, go on…

Adrian: ...I, I can't believe you remember that.

Jan: Oh, we remember everything.

Paul: [at same time as Jan] Everything.

Adrian: Yeah. Encyclopaedic, encyclopaedic…

Jan: …um, but if it's suppl, AI searches are supplied as a matter of course, and you can't turn them off. Then, then actually induces its own, its own, 'cause I don't want that inference cost 'cause I'm not interested in what the inference is gonna tell me.

Um, so in that respect, that automation of AI that's being stuck on with everything that you seem to seem to, you know, access these days must in itself then actually perpetuate this energy use.

Adrian: Yeah, absolutely. And I, I think these, so, so when, when people ask me what, what can people do? It's when they make those kind of business

decisions to embed the thing, and then you are doing it for every search, for every person, times the scale of the number of users of that thing.

You know, you build Copilot into Windows 11, you build that into your search. I noticed the other day Meta have embedded AI into my WhatsApp, uh…

Jan: …which is the least. Which is, why would I need it? [laughs]

Adrian: Well, well, I don't, I don't have that many friends. [Jan and Adrian laugh]

I don't need AI to, to, to tell me what my friends have said, right? I don't get that many messages. I'm not very popular.

Um, but the point is that it's very hard to turn it off. And I, and they're all sort of seeking, you know, what's the new business model? How am I gonna use AI to, to do well here? Uh, so, you know, um, I mean, Microsoft's famous, uh, Bing AI that started going more and more over to the, to the right.

You know, I mean, they're all sort of trying to find a way to use this, but, but, uh, but I think we're losing the ability to turn it off.

Jan: Yeah.

Adrian: Uh, and that's all, you know, and people think we need it, so they're building more and more infrastructure for it.

But actually the interesting question about the WhatsApp thing is, does that mean that AI is now reading all my private messages to be able to offer some value to me? I dunno about you, I thought those were private messages, I don't particularly want that.

And the same with the big investment in the UK, you know, I, I do not want my health records, you know, going, going in, in a format that, you know, to some US company, et cetera, et cetera.

There's some big questions for us as a society about how we get the data to the AI, where the AI is, you know, what control we have over that. And I think that's very interesting.

Paul: It, it drives me insane. Just like with you, with Google, 'cause Google in itself is a form of early, very early evolution AI 'cause it was searching for you and finding the information you wanted.

Now it's providing you with these almost entire useless summaries at the tops of pages.

Adrian: I'm, I'm gonna just briefly defend AI for a second, right?

Paul: Don't, don't do it…

Adrian: …no, no…

Paul: ...don't do it, you'll get kicked out.

Adrian: …oh, okay. [Paul and Adrian laugh]

Uh, I was listening this morning on the radio they were saying about this sort of Welsh language model, right. I would never, I'm sorry, Welsh people, uh, I think we have one in the room. I, I haven't learned Welsh, so this technology would, would maybe help me translate some stuff.

But we have, we have to question, you know, AI doesn't think, there is no I in the AI, right? It's not intelligent. It's producing language, which is statistically plausibly, like the language we use. You have to. You have to question how, how right it is.

Um, but I, but I think there are many use cases as, as Jan said earlier, right? There are training, deep learning models on lots of images of cancers and using that to scan cancer, you know, there were the, the sort of announcements around new drug discoveries by, you know, entirely AI driven virtual lab. You know, there, there are definitely use cases for AI.

The problem I have with that is that is used as the lever to justify all the ridiculously bad uses oft AI, for micro targeting messages, you know, increases cybersecurity risk through phishing scams that are getting more and more convincing because of the power of AI. There's lots of ways in which AI is accelerating things.

But I, but I don't want to just, I'm always very negative. [laughs]

I don't want, I don't wanna just say AI has no use…

Jan: …no…

Adrian: …it's the blanket default embedding of AI in everything that I'd love to challenge.

Jan: Yeah.

Paul: Yeah. I, I've lost count the amount of times that on any social media platform you get asked, would you like AI to write this message for you? Or would you like AI to edit this message for you?

And, if you're doing that, you're no longer a person, you are just an automated bot almost, because AI is doing the talking for you. It's not your voice, it's AI's. But yeah, I, I do understand there are uses for AI that are probably really important, and I just, you never see them in everyday life.

Jan: And I suppose that's the thing with all sorts of new technology and many of the sustainability issues that we talk about. And so thinking about modern slavery, employment brokers are really great to match workers to jobs, but it can go wrong. And I suppose we're in that kind of thing.

So. If we focus on then what's at stake with data centres and sustainability. So, so you said earlier that data centres were growing faster than the capacity of the energy system to, to generate, um, energy sources for them and in particular renewable energy sources.

Why are data centres after renewable energy and, and are they hoovering up the, you know, or about to hoover up the renewable energy capacity in the countries in which they operate?

Adrian: Yeah, no, that, that's a great question. So, so the, um, so demand for energy is growing globally and renewables are probably the most cost-effective way of delivering energy, um, where you do need to decarbonise everything from transport to heating to, uh, the energy that feeds data centres.

But the problem with having, um, a growing demand for energy is twofold. You're creating more that you have to decarbonise. Uh, you're gonna have to build those solar panels and wind turbines and hydro plants, which, which has a footprint, uh, in energy material terms in, in itself.

And these are very rich companies. These, these are, these are trillion-dollar companies in many cases who have more money than many countries, who can afford to, to buy all the green energy in a particular region to the exclusion of maybe other things.

So there's this, this sort of example I like to use in one of my talks of a, a bread factory in Sweden, who, who couldn't buy the capacity they needed 'cause it had already been reserved for, for data centre use.

So if, if your grid is of a certain size, and let's take the UK as an example. A lot of the good wind power is in the North. Uh, very controversial topic, uh, especially for the Scots. And, and a lot of the use for it is in the, the South. [laughs] And you've gotta get the power from one end of the country to the other.

Where are you gonna put your data centres? You quite often put them near the population centres. Um, for, for reasons I was talking about, latency and so on, but various other reasons as well. Um, there's a bit of a water shortage in the South. We've had a hose pipe ban for a year that it's getting drier and drier, global warming.

So you've then got competing demands of, already there were water shortages due to the, the sort of growing population in the south of the country. Um, you, you've got sort of energy and water challenges, uh, to meet there and, and, I would, I would, do I want more AI for… I'm going, I'm gonna try and trigger Jan…

[Jan and Adrian laugh together]

Jan: …it's very easy to do…

Adrian: …more AI for advertising and generating synthetic Lego pictures, or would I like my drinking water? I mean this is, this is sort of where we're at, isn't it?

Paul: What about something like rare earth metals? Have they been used a lot here as well? I know that we're on an area that's not exactly your specialty, Adrian, but is that something as well they're hoovering up to use in the manufacture of everything that's in there?

Adrian: Yeah, so, so, uh, so I know, I know people who are studying this. I've, I've, I've not looked into it more than reading their papers. But the, um, there is growing demand for rare earth at a, something like 10% a year, it's at a rate exceeding the growth of other material resources. Um, so that also has an environmental impact because to get the, I mean, they're called rare earth for a reason, right?

You, you've gotta dig up a lot of earth and, and spend a lot of energy and materials processing those to get the real, you need really pure materials in things like chip fabrication. Uh, and one of the other things that it's driving is if, if you need to train AI, you want bigger and bigger and faster graphics cards or,

or, or AI accelerator cards. Um, so these, these are the cutting edge in terms of the, uh, sort of integration of processing units.

So, so you need big fabrication plants like, um, uh, TMC are building, well, everyone's trying to build fabrication plants for the latest generation of, of integrated circuits. So, so the more we push for AI, the more demand for computation there is. The more computation there is, the more you need these very specialist processes. You need fabrication plants for that.

Once you've got those things, you're gonna, you need to get your money back. So you're gonna run those things for decades, producing chips at that process level, but you're gonna keep pushing for the next generation.

And there, there are roadmaps for more processes. Um, and I suppose the thing that I hadn't realised until relatively recently is, as you try and squeeze the performance at the, you know, you get the transistors so close together, there are kind of quantum effects.

You need to, you know, they've got roadmaps looking at really clever architectures to kind of minimise that. But one of the other things they're doing is exploring all the other mixes of rare earth they can find to get those gains, to keep pushing that to the next level of integration.

So, so the demands for what was on a processor in the 1980s in terms of materials, it was a much simpler process with much less rare elements to it than, than the chips we're making, uh, today, which is just about anything that isn't radioactive.

Paul: Is there any kind of planned obsolescence when it comes to these data centres, uh, with the ones that are going offline, and what happens to everything that's there, all the infrastructure, all the materials, anything like that?

Adrian: Ooh, uh, so I, so I don't know. Uh, I, I think the. It's not really a question of, um, planned obsolescence other than, um, they know very well how long a computer will last in that environment when you work it pretty hard.

Uh, and they have regular cycles of updating the hardware because you are going to have to, um, there's probably is a question there, at what point you do that, whether the thing's exhausted at, you know, you probably don't wait for the failure, you probably just replace 'em earlier, uh, than that.

And, um, yeah, I dunno what you do with, uh, 25,000 A100 graphics cards…

Jan: …well, presumably circular economy solutions might come into play…

Paul: …you'd like to think so…

Jan: … to retake those materials back out and reuse 'em in some way.

Adrian: Yeah, so, so I think, so I imagine so.

So the high-performance computing community has quite a good pipeline where, you know, the top end passed down to the next generation and things get reused. I'd like to, I could do with knowing more about this. All I know is that the, the sort of the, the Global E-Waste initiative thinks that 22% of e-waste goes through official channels and the rest gets recycled illegally in really kind of, uh, not very good ways for people on the planet.

So, um. Yeah, I mean, I, if those things aren't completely broken, I think, I think they could, could be reused and, you know, that, that's the, they're are pretty high uh, end hardware that will be a big upgrade for some other parts of the industry…

Jan: …yeah…

Adrian: …so I imagine it's a pipeline that moves down.

Paul: I was thinking you can't necessarily give the energy back and you can't give the water back, but the least you could do would be to give back the materials that you've used that you're no longer using.

Jan: Well, now as soon as you say that, it pop, pops into…

Paul: …you, you found a way for 'em to give the energy back and the water back.

Jan: I wonder about giving the energy back. So, so we're using water to cool heat, but in other parts, well, in some parts of our system, we would love some heat going into a district plan in order to be able to warm people's houses.

So could data centres' excess heat end up in district heating plants for, for people? Or, or is that like, like mad, mad complicated?

Adrian: …no, no, no, well. So…

Paul: …this sounds like something else from a Bond film.

[Jan and Adrian laugh]

Jan: It sounds more socially just than a Bond film!

Adrian: Oh, you got my, my feeble brain now.

Uh, someone was, was telling me about a project they'd done on where people will accept heat from. Uh…

Jan: …oh really? Oh, that's interesting…

Adrian: …I'm gonna have to come back to that one. Uh, yes. So, so this is, this is actually common practice. Right. So, so the famous example was the, the, you know, the, the data centre that feeds the local swimming pool. You need to keep the swimming pool hot. Uh, you've got the waste heat. Don't vent it to the air, you know, heat up the swimming pool.

And, and we had a PhD student who was looking at, can use it to dry logs or power greenhouses and, you know, various sort of different virtuous, uh, uses of the waste heat, which, which there are, which there are some.

Uh, and in the Netherlands it's pretty common that you, you won't get planning for your data centre unless you feed into the district heat network and that, and that's more common because there's more, it's more common to build district heat networks and housing that feeds from district heat networks.

Whereas in the UK this is something we, we've, we've not done very much. We've done a little bit, uh, more recently and some, some of the more recent housing projects are, are, are doing district heat networks.

I, I mean, I would, I would generally say district heat networks are a good thing rather than having a boiler in every house. You, you know, you, you use waste heat from all sorts of things and, and help share that.

Um. I was chatting to someone from Belgium who was saying they were, they were really struggling sometimes to match up the business case and also, um, the, the use model. So, so they had, uh, they were trying to get waste heat from a, uh, a factory. In this case to, to people's homes. And the factory was producing heat 24/7 and the homes only really needed it during the winter and during the day. So, so there was a sort of mismatch that people wouldn't really wanna pay for the energy, uh, that way.

And, and, and the same with data centres to some extent, they're gonna run year-round…

Jan: …mm-hmm…

Adrian: …uh, and, and our use of heat isn't always year-round. So, um. But I, but I think generally, um, if you're gonna have waste heat, sure.

Uh, the other, the other interesting question is do, do you want to live next to a football pitch sized, you know, light bedecked data centre with some diesel backup [laughs] generators when you build your housing? So it's not very good heat. It's quite low temperature and that means it cools quite quickly. So you, you need to be quite close to where you need it. Uh, and, and that's a bit of a rub.

But then maybe this comes back to Paul's point that, you know, may, maybe you would have micro data centres around, you know, we're building a housing, there's building a housing estate down the road from the University right now. Could they have a micro data centre that's, that's feeding in and doing some of the computation? You know, probably their broadband provider would quite like a small place to put their video content or whatever…

Jan: …yeah...

Adrian: …you know, may, maybe we could be thinking more cleverly about how to marry the two things up.

Paul: Oh indeed. If you are a James Bond villain, it's the perfect place to house all of your lackeys. [Jan laughs] You don't care what quality heat they're getting, but you're managing to house them and get that all through the residues...

Adrian: …you're back, you're back to the ninjas again…

Paul: …yeah, exactly. It's all coming together. It's certainly all making lots of sense.

Jan: This episode was always gonna go down this route, I think. So, so a couple sort of, you know, quick fire questions in a way.

Adrian: Okay.

Jan: Is a chat GTP query use… do we know how much more energy that uses versus a conventional Google search?

Adrian: Uh, right. I'm glad, I'm glad you asked that because, uh, I looked it up. [Jan laughs]

Colleague of mine at Glasgow, uh, Wim Vanderbauwhede, apologies for pronunciation. I can never pronounce his surname. Sorry. Sorry, Wim.

So his analysis, um, was that a Google search is probably something like, uh, 0.0004 kilowatt hours. Uh, which just to give you a kind of ballpark, boiling a kettle is probably about 0.1 kilowatt hours, right?

So, so, but, but that's, that's probably, I, I think that's for just a, a Google search. And then you're talking 0.02 grams of, of, of carbon for that. And he thinks that, um, a, a sort of ChatGPT enabled search, uh, is roughly 60 times that.

So, so it's, so it's, I guess, you know, that's still less than your cup of tea, to put that in perspective, but you can do it again and again. People don't usually ask one query and, and it's not so much that one, is it? It is the fact that everybody is now doing that. That is, you know, and, and I, I saw stats for ChatGPT-3, that it was something like seven tons of CO2 a day or something for inferencing that, so I guess ChatGPT-4…

Jan: …yeah...

Adrian: …is more.

Paul: And again, I guess we need to stress this is based upon a text enquiry rather than the…

Adrian: …that's right…

Paul: …can you create me the silly superhero figure?

Adrian: Yeah. So, so, so that, that's another thing. So, so general models, I, and, and I, and my understanding is, you know, 'cause they can do a great many things, including generating images, it's not really one model, it's lots of other models that are contributing into that.

So if you ask a general model to do a thing, it's gonna do a lot, it's a big model trained on these billions and billions, trillions of parameters. Uh, it's going to do a lot more work than if you can find a specialist model that is just trained to do the job you want. It's going to be lower cost.

So, um, if you took the sort of, um, BLOOM multi-language model, that, that will be much more efficient to train an inference than a general model like ChatGPT.

Classic one is writing code. It's very good at writing template bits of code, like bits of websites and calling APIs and stuff like that. If you find one that's just trained on a sort of corpus of code and, and use that, instead of using ChatGPT, it's going to be, you know, quite a lot less expensive.

Jan: And we were talking about this in a project I've been working on, uh, looking at, um, biodiversity impacts of solar parks, for example, and talking about, um, building language models, but only training it on like the academic research that is focused on these particular areas, so much smaller data sets, and to answer quite particular questions.

And I found that very engaging because you could then deal with all the parameters in each one of the, the studies and look at something synthetic, which I know is something that, you know, exercises us about how do you join together knowledge in ways that'll be useful.

And I was really quite blown away by that 'cause I think, oh, now that's useful. But again, much smaller, not general purpose and not, not picking up crazy things as well as uncrazy things, but like a, a contained training set.

Adrian: Oh, I, so I think that's getting to the heart of a couple of really good things.

One, one is you, you can't really trust what comes out if you don't know what went in, in a way, and the biases that has and, you know, the truthfulness of, of, of what went in.

And, and what you're talking about is a, a model that you, you, you pretty much trust everything that went in. So you're very likely to get a much more authoritative answer out of that, that model.

And it's also going to be much smaller. So it's going to be many orders of magnitude smaller in terms of, uh, the, the sort of, um, training space if you like. So the amount of computation you need to train it and, and also inference it.

So, uh, I, I think that's part of being responsible, I think. Is, is that we, we should be trying to, I dunno, use it where it has genuine value and, and use the sort of most appropriate models and techniques.

Paul: So as tech users then, what can we do to reduce the impacts that we have, uh, when it comes to our computer use, and the use of data centres?

Adrian: Am, am I allowed to have a, a problem with that question?

Jan: You are indeed, 'cause you always…

Paul:, [loudly] …no, no I've got all these problems with AI, which I've been told to drop, and I'm being overdramatic.

[Jan laughs]

Adrian: So, so I, I suppose this is a bit like, people say to me, you know, you, you're refusing to fly, you know, why - the flight's going anyway, you know. It, it's a bit like that, isn't it? One, once you've got the infrastructure in place, uh, then, um, you, you can do your bit to, to remove the thing.

But what you really wanna be doing is, is removing the demand in the first place or not building infrastructure we don't need in the first place. And asking these questions really early on is, is this the right, moral thing to be doing when there's a climate crisis? Um, and we need to decarbonise energy systems. And, and you know, I, I think, I think we've got some big things facing us about the future of work, the future of education. You know, AI is very disruptive in, in good ways and bad ways.

Yeah, so, but I mean, there are things, as we've talked about, right, if, if you're generating movies with AI, that's horrendous. Uh, if you're generating images, that's, that's probably not great for the artist whose work is ripped off. And, uh, you know, don't do those things in moderation. The less often you do them, the lower impact that will have.

There's probably, uh, so where you get the energy from and where the computation happens globally does affect how much, uh, greenhouse gas emissions there are, right. So if your data centre is in the, where you've got a high level of renewables, like in the, in the Nordics, uh, then the energy is likely to be greener, so the amount of computation, um, uh, will have, have lower emissions.

I mean, it doesn't change the, probably the emissions from building those data centres, which is probably done where it's manufactured in largely the far East. So we can be a bit choosy about where we do things and that, and that's great for model training 'cause we can do model training, uh, where the energy is greener.

I mean, you don't have to train it here. I mean, there's an access to data question, data sovereignty issues around that, but. Um, and the other thing we can do is, is you don't have to train the thing quickly, so we could put them near renewable, you know, wind farms and when you've got excess energy, you could do a bit of training and you know, you can help to, they're big demands of energy, but they're quite deferrable loads. So we could use that to help us with our, our energy transition story to some extent.

I would say it's not enough to generate to, to say why that, that's a justification for, for them. But, but we can be doing smarter things, uh, there. Yeah, and I mean, as you say, right, we, we can turn off our, if, if we, if we're trying to push back a bit on the demand, we need to sort of push to have options to take it out of some of the things we do every day.

You know, and you know, I, I put in the, uh, you are absolutely right. You can change your query string on your search engine and, and that will mean it's not running a, you know, somewhere else in the world it's not running, uh, those AI jobs, hopefully.

But, but I think the most important thing we can do is when, when we, when we are the ones making the decisions and pushing for the adoption of these technologies, you know, in, in our work and in our, our lives, you know, it's, it's, it's those decisions about what, what we're going to be doing and what we need, um, that are more critical than the, the ones, uh, on a day-to-day basis.

Jan: So, for this whole field, what's next for you in this, this area?

Adrian: Yeah. So, so, so I'm, I'm very interested in, in how do we, I mean, it's really hard, right? It's really hard to decide what are the right digital methods to adopt in, in almost anything.

So we've got a project on, on, uh, sort of responsible innovation in science and we're working with, uh, environmental scientists who are looking at questions about, do I run big process models for simulating what's going on, on the planet, or do I run AI to emulate the big process models?

And we're, we are looking at how, how can you work out what the footprint's going to be and which is a better choice for, for, you know, and what's that trade-off between the value of the answers you get out of the thing and the environmental damage of the computation that you're doing?

So this is a project with CEH that we've just got off the ground. So, so we're about a few months into that.

Jan: That's the Centre for Ecology and Hydrology.

Adrian: Uh, yes, that's correct.

Jan: Just, just for our listeners.

Adrian: Yeah, yeah...

Paul: …are you sure, Jan? You didn't sound sure.

Jan: Oh, well, 'cause I, because I'm a bit dyslexic. So those two words, you, so it could be CHE, it could be ECH, I mean, [laughs] it just sort of like, it swirls around in my mind…

Paul: …ECHR?

Jan: That's something different.

Paul: Yeah, yeah…

Adrian: …just because I know what CEH means, everyone must know…

[Everyone laughs]

Adrian: …yeah, no, you're absolutely right. So they, they do a lot of things like looking after, uh, what's the water quality in the UK and, and um, they're, they sort of look at, they've got a lot of data science going on looking at the water, uh, use and various other things.

Jan: Yeah. Well, that sounds very cool.

Paul: Well, I'm sure it's something we'll talk to you about in the future, but until then, Adrian, thank you very much for joining us, it's been great!

Adrian: It's a pleasure. Uh, I, I can't believe you'll have me back for a third time.

[Jan laughs]

Jan: We will, we'll wait and see. [Adrian laughs] I mean, he's promising. I'm promising nowt.

Paul: That's nice. He's meant to be your friend,

Jan: He is my friend. That's why I can say it.

[Everyone laughs]

Paul: Delightful. Thank you very much.

Adrian: Thank you very much.

[Theme music]

Paul: There's so, so much Jan to take into consideration whenever you're doing anything as simple as a Google search, isn't there?

Jan: [laughs] Much more than I ever, ever thought. But also the whole sort of, I think the, the life cycle of technological development, how it is, you know, how it might be used, therefore what kind of investment it, it, um, it attracts. And then what it induces in terms of materials and, and energy flows is just, just phenomenal.

Paul: Yeah, it's all the considerations you have to make beyond just the simple, oh, do I really want to know this? Or what, where can I find this? Yeah. Oh, it, it's crazy. Well, I still come back to the fact that people should stop making silly avatars of themselves.

Jan: Well, I'm sort of saved from that 'cause I, I didn't even know you could. And if, if I did know you could, I probably couldn't do it myself.

Paul: No, you'd get AI to do it for you, and that's the problem.

Jan: [laughing] That's the point. I also was very struck though, that there's different types of AI and different types of applications, and different sizes of applications, and what the AI might be trained on and what it might be, you know, what the querying might be doing.

So it seems to me that some of the, some of the things we might see in the newspapers that are full on or full against AI actually are missing all of that subtlety. And, and within that subtlety is where really interesting answers and decisions would rest.

Paul: And of course it isn't just AI, as we were saying, it's data generally and the demand for data, and then the various different things that you have to think about when it comes to data and the energy demands, the water demands.

I was particularly struck by the Swedish bread factory that wasn't able to buy its energy because the data centre down the road had bought up all the energy. So therefore, all the people in that particular part of Sweden weren't getting any bread that week.

Jan: Yes. I think they end up having, I, I read that story, I think they had to end up manufacturing further away, in which case you then have to ship the bread. They might have been otherwise. Yeah, so, so which comes first? You know, bread or bytes?

Paul: Yeah, that's it. People need to consider, do they want their daily bread or do they want their daily dose of silly avatars?

Jan: Yeah.

Paul: Yeah.

Jan: Madness.

Paul: It is.

Jan: But I was very struck, and I'm not saying this just because our, our producer is a, a Welsh language speaker, but, but maybe I am, just to keep on their right side, but also the extent to which, um, these kind of applications can really, you know, build capacity in a language, support people to learn it, all of those sorts of things as well.

Paul: Would it be the worst thing in the world if the Welsh language became extinct?

Producer: Yes!

Jan: Uh, dear listeners, you may have heard a very, a very resounding yes in the background and, um, and, and Mr Turner is going to pay for that comment later on.

He's gonna be edited in the most mean and horrible way from now until eternity.

Paul: [joking] The spellings alone, and the sounds that come out of a Welsh person's mouth.

But I know, I, I do take your point though, there are things that AI can do, and Adrian was stressing this, whether it be Welsh language or whether it be cancer diagnosis. I'll leave it to yourself to decide which of those might be the most important. Um, there, there are obvious practicable, really important impacts for it.

And there's reasons why we use AI, and there's reasons why we use data and therefore reasons why data banks and data centres are needed.

Jan: Yeah. And I was really struck, um, the example was, given that the Netherlands is, does quite a bit of integrated planning for the data centres and other needs, I don't perceive that that's the case across the built board.

So that's kind of impressive that a country would, you know, seek, seek a multi-functionality…

Paul: …mm-hmm...

Jan: …with what they're trying to do.

Paul: Strikes me, they're almost planning for say, the next generation, which is convenient, because next week we're gonna be talking about generational governance.

Jan: That was a smooth move. [laughs]

Paul: Thank you, thank you. I'm good.

Uh, yeah, we're gonna be bringing back Nick Barter, another return guest. Professor Nick Barter from Griffith University in Australia, talking to us about next generation, generational governance and how that tied in with his previous work on Future Normal. So that'll be great too.

Jan: I'm looking forward to it.

Paul: Until then, thank you very much for listening. I'm Paul Turner.

Jan: And I'm Professor Jan Bebbington.

[Theme music]