Human Led Artificial Intelligence
Marc Warner
27 November 2023

I'm certain everyone here today has seen Artificial Intelligence in the news consistently over the past year or two, perhaps even more this past week. You might be expecting this to be AI's moment, and then, just like crypto or the metaverse, in a year or two we'll move on to the next thing, and the peak hype around AI will be gone.

However, I'm here to tell you that this is simply the start of AI's journey. The implementation of existing AI techniques in fields such as medicine, education, warfare, and transport will lead to significant changes in how we solve these foundational tasks. And there will be many upsides. But why then are some people concerned? Allow me to illustrate with a little story about a coffee robot.

Picture me in Faculty's offices, where we don't yet have a coffee machine installed. Hence, I need to go down to the coffee shop to get a cup. So, one day, I say to the coffee robot, "Could you grab me a coffee as quickly as possible?". The robot dashes out, smashes through the door in its hurry to fetch my coffee, and then brings it back. I say, "Bloody hell! Thank you for the coffee, but next time, do you mind not smashing the door?"

Tomorrow arrives, and I ask the robot for a coffee, this time saying, "Could you fetch me a coffee as fast as you can, but please don't smash the door this time." So, this time the robot steps on the cat. "Dammit Robot", I say the next day, "Could you fetch me a coffee? Remember, don't break the door and don't step on the cat." The robot races downstairs to the cafe and snatches a coffee from someone who had just been served and races off.

The next day, the robot is sent off once again. I say "Alright, robot… get me a coffee, but please, don't damage the door, don't step on a cat, and don't steal from anyone". However, when the robot gets to the coffee shop there has been an accident - the kitchen is ablaze. The robot recklessly tries to brew a coffee amidst the flames and returns, charred, and melted but holding a coffee. I quickly adjust the rules again: "Alright, I want coffee, but don't step on the cat, don't break the door, don't steal from anyone, and if there are any physical impediments to getting one, it's not that important." On its next attempt, the robot notices that it's raining and returns without a coffee. By this point I'm ready to give up.

My point here is that even a seemingly simple task such as fetching a cup of coffee can prove to be complex when we consider how much common-sense human context surrounds a request. This context enables us to make subtle trade-offs about how important different elements are relative to the task, but it's exceedingly difficult to explicitly state this in instructions.

So, what happens if AI gets ever more powerful? What happens if we start asking AI to solve more and more important problems, like fix the climate, or eliminate all cancer?

We know from mythology that it is sometimes complicated to get what you wish for. King Midas wanted to be wealthy, but he actually wished for everything he touched to turn into gold. Very quickly, he realised the implications of a poorly specified objective function. He was unable to eat, drink and his family members turned to gold. It sounds silly. But this myth shows how hard it is to specify an objective function that contains all the notions of human common sense. It is currently an unsolved scientific problem.

And that worries sensible people at the cutting edge of this field. And if you don't believe me. Here is Sam

Altman, CEO of Open AI talking about benign misuse failure modes:

> ***Q***: *What would you like to see, and what would you not like to see out of AI in the future?*
>
> ***Sam***: *I think that the best case is so unbelievably good that it's hard to even imagine. I can sort of imagine what it's like when we make more progress in discovering new knowledge with these systems than Humanity has done so far; but in a year instead of seventy thousand. I can sort of imagine what it's like when we launch probes out to the whole universe and find out everything going on out there. I can sort of imagine what it's like when we have unbelievable abundance. When we have systems that can help us resolve deadlocks, improve all aspects of reality and let us all live our best lives. I think the good case is just so unbelievably good that you sound like a really crazy person to start talking about it.*
>
> *And the bad case, and I think this is important to say, is* ***lights out for all of us.*** *I'm more worried about an accidental misuse case in the short term. It's not like the AI wakes up and decides to be evil. I think all of the traditional AI safety thinkers reveal a lot more about themselves than they mean to when they talk about what they think the AGI is going to be like. But I can see the accidental misuse case clearly and that's super bad.*

…and now this is Dario Amodei, who's the CEO of Anthropic, another one of the three most famous Frontier labs:

> ***Q:*** *Paul Cristiano recently on a podcast said he thinks there's a 50% chance that the way he ends up passing away is something to do with AI. Do you think about percentage chance doom?*
>
> ***Dario:*** *Yeah. I think it's popular to give these percentage numbers and the truth is that I'm not sure it's easy to put a number to it. If you forced me to, it would fluctuate all the time.* ***But I've often said that my chance that something goes really quite catastrophically wrong, on the scale of human civilisation, might be somewhere between 10 and 25%.*** *When you put together the risk of something going wrong with the model itself with something going wrong with humans. Either people or organisations or nation states misusing the model, or it inducing conflict among them.*
>
> *What that means is that there's a 75 to 90% chance that this technology is developed, and everything goes fine, in fact I think if everything goes fine it'll go really, really great. I think if we can avoid the downsides then this stuff about curing cancer, extending the human lifespan, solving problems like mental illness. All this sounds utopian, but I don't think it's outside the scope of what the technology can do.*

These are big stakes; notice the sensible balance between upside and downside. Crucially, though, as we build these powerful AI, we will be required to encapsulate our values in software. As we develop these increasingly powerful systems, we're going to have to decide at a fairly profound level what we want our values to be. Of course, we can choose not to make any decisions, and AI will still uphold some notion of values, but it won't be in any way defined or chosen by us.

If we're going to encapsulate our values in software, it's important that the discussion involves everyone. Today, the purpose of my talk is to lay out the background and my understanding of this debate as clearly as possible, so that everyone can contribute as effectively.

My talk will be divided into four sections. 1) What is AI and how does it work? 2) How smart is AI? 3) What are our options for controlling AI? 4) So, what should we do?

1. What is AI and how does it work?

So, the first section is: What is AI and how does it work? Let's start with the definition of artificial intelligence. Quite simply, artificial intelligence is the field of research that aims to enable computers to perform tasks which we deem intelligent when performed by humans. This field started with Alan Turing. Turing was a critical thinker in the development of early computation, and one of the first individuals to start thinking, "What if you could build computers, what would happen if they could think? If machines could think, they might think more intelligently than we do, and then, where would we stand?"

Since Turing's time, AI has generally evolved along two streams. It's a simplistic view, but we can think of them as the "good old-fashioned AI" stream and the machine learning stream, with the latter gaining more traction over time. Now, good old-fashioned AI encompasses a plethora of techniques and strategies for making computers perform tasks. The key trait shared by these techniques is the input of human knowledge in one form or another. We provide them with expert rules or tell them things about the world, and they then perform computations based on that information.

Now, parallel to this, what about the other approach, machine learning? Machine learning has been around since AI's inception, but its popularity has soared in recent times primarily because its performance in tasks we care about has proven to be superior. Instead of manually specifying the details like human experts used to, machine learning involves showing the computer a bunch of data and having it learn the patterns directly from that data. This might sound a bit mysterious, and it certainly was for me when I first began learning about AI, but it's less mysterious than you might think.

Let's take a simple example. Here are some pictures of cats and dogs. Suppose we want to train a program to distinguish between the two. We might choose two attributes of cats and dogs, such as weight and tooth length, and we plot them on a graph. The data points scatter in a certain pattern, with dogs represented by squares and cats by circles. As humans, we can intuitively make a guess for a new point on the graph, but the computer needs a clear boundary between the two groups.

We ask the computer to draw a boundary that maximally separates the two groups, with all dogs on one side and all cats on the other. In order to do that, the computer has to guess a boundary to begin with. Let's suggest it starts with a line drawn here, then it looks and sees that one side is 100% dogs, but the other side is approximately 50/50 dogs and cats. That isn't a very good boundary.

The computer then adjusts its line and says, "OK, now I've got maybe 100% cats on one side and two-thirds dogs on the other. That's better." Then it shifts back in the first direction attempting to further improve. Can it do even better? Gradually, the computer refines its line until it finds a nice boundary in between both groups, where one side has all cats and the other, all dogs. So now, when a new point comes up and the computer needs to decide whether it's a dog or a cat, it can say "dog" because it's above the line. And this is the process of learning from data.

You might be thinking that this seems a bit too simple, especially if you've heard of neural networks. So, how do they differ from this relatively simple procedure?

In my example, I only used two categories: dog and cat, and only two dimensions they could vary on: weight and tooth length. But what if we consider that dogs and cats vary on colour, face shape, leg length, height, ear shape, and much more? We could measure all these variables and thus, have a highly multi-dimensional space. Although we humans can only visualise three dimensions, imagine if you could have four, five, six, ten, or even a hundred dimensions.

And we don't just have to limit ourselves to dogs and cats. We could have cats, giraffes, elephants, rhinos, and more. Therefore, there could be way more inputs and way more outputs. Neural networks allow us to take in many more inputs and draw much more complicated boundaries in our data.

While the boundary I used was a simple linear one, neural networks allow us to draw nonlinear, complex boundaries like this one. Using the millions, billions, and these days trillions, of parameters that you train on, you could draw intricate boundaries in very high dimensional spaces. This is all fascinating, but does it have any relevance to current technology?

I'm sure everyone here has heard about Chat GPT. The training methodology we just discussed is very similar to how that is trained. The first step involves gathering a bunch of data, primarily fetching lots and lots of sentences from the internet, such as "my dog has four legs." We then mask the last word of the sentence, so we have "a dog has four ___", with the last part blanked out.

We then ask Chat GPT to predict the blanked word - similar to guessing whether an input is a dog or cat, it now has to predict a likely word to fill the blank, based on existing boundaries. It might respond with, "A dog has four eyes." Obviously, that's wrong. So, we compare the guess against the actual data, and instruct Chat GPT to adjust its line of best fit to provide the correct answer - "legs."

Bear in mind, because there are so many words and sentences, the boundary is incredibly complex and the space is of high dimensions. But the process remains essentially the same: tweak that boundary in a slight way to increase the likelihood of guessing the right answer.

And we have to repeat this process across thousands and thousands and millions and millions of sentences to eventually make Chat GPT capable of making the right prediction. For example that cats have two eyes.

2. How smart is AI?

So, how smart can AI get? The first thing we need to do is define what we mean by 'smart'. Here's a capability spectrum ranging from low to high.

If I ask you to place a chess computer on this spectrum, you might argue that since chess computers can beat all humans, they rank quite high. But then, does that imply that I rank beneath the chess computer, because it can definitely beat me at chess? However, that same chess computer can't recognise a cat from a dog, nor can it order a beer from a bar. So, how should we think about these comparisons?

In truth, this single dimensional spectrum doesn't quite make sense. Intelligence isn't just one dimensional. There's at least one more important dimension to consider, which is the generality of the given capability.

A narrowly intelligent system can only optimise for what it wants within a very narrow domain, like a chessboard for instance. So, you can place a chess computer as narrow and capable. On the other hand, a general intelligence can optimise for its goals in a much broader spectrum of environments. Given this, I myself would be very general, fairly capable, with Chat GPT somewhat less general than me.

But the question is, is there a way to compare these varied intelligences? We can't carry out a thorough comparison because our understanding of intelligence doesn't provide a complete scientific theory yet. However, we can begin by asking some preliminary questions about the inputs to these different models.

The way we compare models based on their input computation is by looking at a floating-point operation, commonly known as a flop. Think of it as a very basic computation task such as adding one number to another. By counting the number of flops, we can compare the amount of compute used in different algorithmic training runs. That's what I'm showing here.

What you're seeing are some of the most important algorithms since the 1950s, and as you can see, in the last ten years, we've used a million times more flops than we did previously.

If we were to plot the compute, we're likely to use in the next year or two, we would have to shrink all of that from the previous graph into this small space. Then we can project the estimated compute for next year's versions of advanced models like GPT-5 onto this graph.

However, what we really want to do is compare that to a human brain. To do this effectively, we have to put this data on a logarithmic scale. Some of you may already be familiar with what a log scale is. Essentially, it's a transformation of the same data I just showed you, but we've changed the axis so that each tick isn't just an additional unit of the same size, but an order of magnitude larger as you go up the y-axis. Basically, what this does is it 'compresses' exponentially growing functions so that you can view them all on the same plot.

Where would a human brain sit on here? It's important to say that we don't have a perfect understanding of how a human brain works, therefore our ability to compare a conventional computer to a human brain is relatively shaky. What I'm about to show you here are thoughtful estimates, but they could be orders of magnitude out in reality. However, if we were to make a guess, it seems like the extent of an 18-year-old human brain would probably lie within these two bounds. As such, the amount of computational power that's used to process data over the first 18 years of life is roughly within these bounds that I'll illustrate. It appears that the most modern algorithms may have an approximately similar capacity.

But one might argue that if you actually took an 18-year-old in the state of nature and assessed how much they were able to learn, they would never be as capable as a modern human being due to our accumulated culture. We definitely have a huge advantage. Therefore, achieving the level of all human brains might be necessary before we acquire very powerful outcomes. This would fall within these two bars I'm about to show, and it's substantial - still significantly beyond what we've currently achieved, but we're starting to approach the lower end of it.

Some people might argue at this point that these types of mechanistic reductionist calculations aren't very useful because there is something intrinsically special about humans. I believe this perspective is misguided. Historically, thinking humans are special has been a very poor algorithm for understanding science.

Take astronomy, for instance. We used to believe that the Earth sat at the centre of the universe, as illustrated by pre-Copernican diagrams of the universe. However, we now know that we don't even exist at the centre of our solar system, let alone the centre of our galaxy or the centre of the universe. We exist in a very ordinary place in the vast expanse of the universe.

With evolution, we used to think that humans were placed here by God in a very special role. Nowadays, we recognize that this idea of specialness was wrong, and we understand that we evolved like other animals.

Then there's quantum mechanics. In the early 20th century, physicists earnestly debated whether consciousness was required to collapse wave function. Now we know consciousness is not required, any mass of roughly equivalent size can do exactly the same job, decohering a quantum system, as a human being.

Therefore, in physics, astronomy, and biology, we've found that this notion of human specialness misguided our intuition in such a way that it slowed the progress of science.

And I would argue that it's quite easy to fall into a similar trap with intelligence. If I showed you this plot again, depicting capability and generality, and asked you to locate Einstein on it, you would probably place him at the top right. However, I actually think it's more likely he would fall at a much more ordinary part of the spectrum. Just like Earth isn't the centre of the universe, Einstein won't be at this special point at the top right of this plot. Due to some technical reasons, I think he'll probably fall somewhere here. I won't delve into that now, but I'm happy to answer questions about it if people are interested.

So then, the important question is: what occupies the top right here? What could this be?

This is what people sometimes refer to as 'superintelligence'. So, an intelligence that is at least as general as ours but having much higher capability. There is a very natural tendency here to anthropomorphise superintelligence, to think of it as either a god or a devil. I believe both of those are poor mental models for this. It's much better to think of it as a sort of universal chess computer. In the same way a chess computer has some values that it's optimising for, a board that represents the world, and actions it can make with its pieces.

I propose that you should think of a superintelligence as an extremely powerful optimiser. Whether it has inputted or learned values, its board would be the wider world and the universe, and its actions would involve whatever elements it can control. This perspective prevents us from falling into simplistic anthropomorphic questions, such as whether a superintelligence would randomly wake up and become evil. Instead, we are much more aware of the fact that we can't even make a word processor that doesn't have bugs in the code. Can we build a super intelligent, world-optimising ‘chess computer ’that doesn't have flaws?

A crucial question, of course, is whether this superintelligence is even possible. As of now, no one can say for sure, since we've never built a superintelligence. But I will say, as far as I understand, there is no scientific reason why we should not be able to create one.

None of our scientific theories suggest that human brains are somehow the limit of information processing capability. In fact, in many ways, even our computers today surpass us in certain elements of recall or speed of processing.

3. What are our options for controlling AI?

Firstly, I'd like to address a question that sensibly comes to mind for everyone: why not just turn it off? If we were to create AI and something turns out to be dangerous, couldn't we just pull the plug on it? While this seems like a reasonable solution initially, a bit of reflection will lead you to understand that we'd have to make active choices if we want this option to be available.

Let's say we entrust this super powerful, slightly myopic optimising chess computer with a complex problem, for instance, reducing the amount of carbon in the atmosphere. It seems like a very laudable goal, and so we start it working.

This AI might notice that some people's livelihoods are deeply linked to generating carbon emissions. If it wants to succeed in its goal, it will have to ensure that these people, who are incentivised to prevent its operation, can't stop it.

So, what does it do? It might replicate itself across every computer on the internet so that it cannot be easily stopped. If we can conceive such a strategy in mere seconds, it underscores how challenging it would be to halt such a system. A genuine superintelligence, myopically focused on a specific mission, could easily strategise to stop us from disabling it. And it’s important to note that this behaviour comes very naturally from the notion of having a goal, not, as critics sometimes strawman, from some anthropomorphic competitive spirit.

Now, some scientists have very interesting ideas about how you might build a more permanent off switch into an algorithm. For example, programming the algorithm to maintain a constant uncertainty about human preferences. So, when it's about to execute a command, if it thinks there may be some uncertainty whether we actually want it to or not, it returns to ask us.

But it's not a given that we will automatically be able to unplug these tools. We have ideas about how one might build that into the algorithm, but the point I'm trying to make here is that it doesn't necessarily happen by default. We have to actively design these tools in a way that upholds this capability.

So, how should we interpret this map in a world of uncertainty? As I've said earlier, we genuinely don't know whether we can build superintelligence, although it seems very possible. However, we have to be honest with people outside of the field. We do know that there's a large green region in this space that is very familiar. We have used AI for decades in circumstances we completely understand and control. In contrast there is a red area where understanding and control eludes us.

Therefore, to me, it seems extremely important that we are very judicious about advancing as fast as is reasonably possible within the green zone. Many people don't realise that the UK is facing a fairly devastating prosperity crisis - since 2008, our GDP has essentially flatlined. We have the opportunity to bolster healthcare, improve security, and enhance renewable energy generation, all leveraging AI that is completely safe - the kind that has been used for decades.

As for the red zone, a very important question we need to grapple with is: should we venture into this area at all? Those advocating for advancements in this area argue that the wider world still faces many unresolved problems.

Global warming, disease, nuclear weapons, pandemics, etc. Potent AI could help us solve these issues in many ways. You could argue that humanity is currently bottlenecked on intelligence, and powerful AI could help alleviate that.

However, critics point out that this area is unknown and not obviously safe, and they're right. The dynamics are complicated. Whether it's companies or individuals, everyone is searching for different combinations of wealth, fame, and power, creating an incentive to race a bit more than they, or you might like.

One way we try to prevent these 'race to the bottom' situations is through regulation. However, it is vital to note that we often struggle to regulate even well-understood domains like housing and energy. The UK, for instance, grapples with energy prices that are five times more than those in the US, and housing costs that are several times higher. These regulatory choices we make inflict terrible self-harm.

Given these challenges, how can we expect to successfully regulate these new technologies, where even the most cutting-edge experts don't fully understand? Poor regulation could easily make the problem worse, not better.

This debate rages across Twitter: we should regulate this, we shouldn't do that. Now, I actually believe we can do better than the debate on Twitter, as surprising as that may sound: We can break this down along two fundamental dimensions: How long until we achieve superintelligence, and how easy will it be to make a superintelligence care about us?

Estimates of the time until we achieve superintelligence obviously vary. Some people think we have a long time, others believe it's a short time. Just for context, 10 years ago, the furthest estimates would have been hundreds of years, maybe a century or two, and the closest would have said a decade or two. Over the last 10 years, almost everyone's estimates have shortened.

So now, the long timeline proponents might predict a few decades, while those on the shorter end of the spectrum suggest two to five years.

The second crucial question is how easy will it be to get a superintelligence to care about us? This varies on a spectrum from easy to hard. Either *yes*, it'll be super easy, or *no,* it's going to be incredibly difficult. This gives us a two-by-two grid that looks something like this.

On the X-axis we have the timescale, and on the Y-axis we have alignment, or how easy it is to make AI care about us. We can then talk about two groups of people falling into these quadrants: the pessimists and the optimists.

The pessimists believe it's going to be really difficult to get a superintelligence to care about us and that its emergence is coming really fast. They advocate for strong regulation - putting a massive brake on the proliferation of technology, stopping most research, and creating a single international research centre to eliminate these race conditions.

The optimists say we should calm down. They believe it's going to be easy to address the problem when we get there and that we're so far away from it currently that we don't need to worry. They argue our current technologies are almost irrelevant to the ones that we will actually use to build AGI, or artificial general intelligence. There's not much point worrying about it now - let's reap the benefits and deal with the superintelligence consequences later. They want to encourage research and wide deployment of this technology.

Both sides can point to genuine downsides in the opposing viewpoint. If you halt proliferation, you're inherently stating that control over technology will be held by a select few. If you stop research, you limit potential applications. If you create an international research centre, you give various countries the chance to collaborate, but then potentially develop clandestine programs, because it will offer advantages to their own economies or national security.

Similarly, the 'green zone' policies also have their downsides. If it turns out that we are on the brink of creating superintelligences that are hard to align, encouraging more research increases the likelihood of creating superweapons. If you deploy this technology widely, you risk putting superweapons into the hands of everyone.

The truth is, we genuinely don't know where the world sits. Will we create superintelligence soon, or will it take us a long time? Will it actually be difficult to get it to care about our values, or will it be easy? We don't have the answers.

My suggestion would be that if everyone was a bit more aware of this framing of the debate, we could discuss plausible policies for each quadrant. Then, despite our uncertainty about where the world actually stands, we could evaluate which portfolio of policies offers us the best protection against downsides and the most opportunity for benefits. I’m going to present a version of that towards the end.

You might ask: What's our current approach? It's not ideal. Currently, we're mostly creating powerful but uncontrolled and poorly understood algorithms; we're on some kind of cusp between the red and green zones.

I can think of two approaches that would be significantly better. One is to adopt a more incremental approach: we'd build more general systems with safe and understandable components where we can control them, and gradually increase their power, ensuring that what we're doing is safe at each step. Perhaps making them open increases the safety further?

The second approach would be to target the red zone directly, but with a provably safer approach - meaning we have mathematical guarantees that what we're doing is safe.

One potential method for figuring out what to do is to look to other domains for inspiration.

We could learn lessons from many different fields - but there are clearly both lessons about the positives and the negatives for regulation. I'd like to take you through a few here. The first domain we can learn from is the creative industry. Here, we recognise the need to incentivise individuals to work on projects for the long term - this is why we have copyrights and protections for intellectual efforts. But various companies have managed to extend copyright, potentially at the cost of wider creativity.

The second field is the nuclear power industry. When you want to build a power station, you have to do a lot of work upfront to demonstrate both qualitatively and quantitatively that what you're proposing is safe. As we build more and more powerful algorithms, we might want to take lessons from the nuclear power industry on how to do this. However, critics will say that nuclear regulation has stifled safer nuclear power from getting adopted, meaning that more people die from the air pollution from coal than die from nuclear accidents.

From the realm of cybersecurity, we can consider red team testing. In cybersecurity, organisations wanting to protect their systems and fortify their defences will pay outsiders to try and infiltrate their systems. This way, they can identify vulnerabilities and rectify them. In much the same way with powerful AI, we are already starting to see organisations 'red teaming' their algorithms to comprehend their capabilities and enhance their safety.

From the pharmaceutical industry, we can take lessons from clinical trials which prevent potentially harmful drugs from reaching the market. Similarly, in the field of AI, it could be beneficial to put increasingly powerful algorithms through some type of testing process before we release them into the world. But again, critics would say that the bureaucracy of testing has created too big a barrier to getting medicines into the world, and people are dying as a result.

From the financial sector, we understand the concept of 'know your customer.' In the field of AI, we likely want to restrict the use of powerful algorithms, at least initially, to individuals we know and trust, making sure they won't use them maliciously. But, in the financial sector, it does seem that know your customer regulation has created enormous burdens on banks in ways that don’t correspondingly decrease fraud.

Finally, when things do go wrong, we'll want to learn from sectors like aerospace. They conduct very truth-seeking investigations - they have a learning culture. When there's a crash, they work hard to understand the most truthful version of what happened. That has enabled them to make enormous safety improvements. That would be a powerful thing for the AI domain too.

So, regulation comes with very real caveats in the actual implementation, and you have to decide which trade-offs you are comfortable with.

But are there technical approaches that are fundamentally safer? Yes, there are many. These include the Open Agency Architecture from Davidad at Aria, Inverse Reinforcement Learning from Stuart Russell at Berkeley, the Beneficial AI Roadmap from Yoshua Bengio at Mila, and the human-led AI that we work on at Faculty.

So, I'd like to take just a couple of minutes to explain how we think about this problem. We aim to build AI that is designed to be powerful and trusted. As we do so, we keep three tenets in mind.

Firstly, we want AI to be safe. This involves ensuring every algorithm has a governance mechanism controlling its actions.

We want it to be modular so that each component can be tested and understood separately.

And we want it to be human-centric. So, explainability is built into the structure and algorithms.

These are the foundations of the approach we’re taking at Faculty. This is already being utilised in the NHS to assist hospitals in making better decisions on patient flow within the hospital, aiming for maximum efficiency. Moreover, it is increasingly being adopted in the commercial sector to improve decision-making in areas such as supply chains and maximising machine uptime.

4. So, what should we do?

So, what should we actually do? I believe it's crucial to acknowledge that AI is an incredibly complex field, and I am greatly simplifying it in this brief lecture. Talk of ‘regulating AI ’is like talk of ‘regulating physics’, where you have radios on one end and nuclear weapons on the other. It wouldn't make sense to talk about regulating physics, because a nuanced approach is necessary.

With AI, if we’re in the green zone, we can proceed as quickly as reasonably possible, while in the red zone, caution is certainly warranted. But how should we achieve that caution? One group says that we should build openly and incrementally, and that will give us the best possible chance of making this successful. The other group says that we need to reduce race conditions, regulate, and solve the problem that way. If you think you know for certain, then you have more clarity than I do.

However, if we safely create the technology, we are still left with the question of what values should AI embody? And that is a question for all of us. It may seem like avoiding the issue, but the truth is that building transformative AI will require us to define what it truly means to be human. We will have to determine what we genuinely care about and how much effort we are willing to invest to uphold those values.

While this may seem like an ambiguous ending, it is because we are not at the end of this story. We are currently in the middle, and how it concludes will depend on the decisions made by people like those in this room today.