**Alex Kendall:** 0:00

The world model for me is a model that can understand the state of the world and predict how it's going to change, given an action you put into it. So, mathematically speaking, it's a function that takes your current state, your current action, and predicts the next state. Essentially a simulator. It's a model that can allow you to understand how the world will evolve, given different things you might want to do or interact with that world. The result of this is a world model that can simulate the future and, if you want to, you can take that state and decode it back to video so you can produce the video output of what's actually going to happen.

**Craig:** 0:36

But you might not want to.

**Alex Kendall:** 0:37

If you want to keep this real time and efficient, you can just stay in your embedding space and use it to drive your car.

**Craig:** 0:43

Hi, I wanted to jump in and give a shout out to our sponsor, netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,000, 25, 1. 36,000 because that's the number of businesses that have upgraded to NetSuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr. And more. 25 because NetSuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customized solution for all of your KPIs in one efficient system with one source of truth. Change risk, get reliable, forecast and improve margins Everything you need all in one place. As I said, I'm not the most organized person in the world, and there's real power to having all of the information in one place to make better decisions. This is an unprecedented offer by NetSuite to make that possible Right now. Download NetSuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free at netsuite.com. That's I on AI, e-y-e-o-n-a-i all run together. Go to netsuitecom I on AI to get your own KPI checklist. They support us, so let's support them. I'm Craig Smith and this is I on AI. This week, I spoke with Alex Kendall, CEO of Wave AI, to understand Wave's innovative approach to autonomous vehicles using a world model called Gaia 1. Alex explains the advantages of world models, which we've explored before on this podcast with Jan LeCun, and how they can be used in AI agents. The discussion offers a unique view on the progress, promise and obstacles in developing AI to act in the physical world. I hope you find the conversation as fascinating as I did.

**Alex Kendall:** 3:36

I'm an engineer at heart. I've loved building things ever since I was growing up and did so in the back garden, but got the chance to work on a bunch of robotics in my childhood in New Zealand. Whether it's building drones to chase some sheep around the field that I grew up in, or other projects I did at university, One way or another they ended up taking me to the University of Cambridge. I spent many years there doing a PhD in research fellowship in computer vision. I'm fortunate enough to publish some of the first work that applied deep learning to scene understanding algorithms like semantic segmentation, depth motion, and other forms of scene understanding. I think that work really inspired some of the ideas that I've had to be able to build machines that can make decisions for themselves. Interestingly enough, a lot of my PhD work stops short of understanding the future or doing future prediction, which is, I think, one of the big topics we've been able to address with world models and Gaia, which I'm looking forward to talking about.

**Craig:** 4:45

Yeah Well, that's fascinating. As I said, I just had Yann Lecun on the podcast talking about world models. He mentioned Gaia. Where do I start? I'm interested in world models as an alternative to large language models. Gaia won, your model. Yan says it's a little different from his JEPA architecture that he's using to research world models. Can you tell us a little bit about Gaia winning? Then I have a lot of questions. I'm interested in marrying this tech with robotics, because the big challenge in robotics, beyond the hardware, is building an AI brain that can plan and make decisions and that sort of thing. There's a lot of talk right now about LLMs being able to play that role, but I also think there are a lot of problems because of the hallucinations or the fact that large language models don't have a very concrete underlying model of the world. Why don't you talk about Gaia, how that came about, what it is both generally, and then how it's built and will go from there?

**Alex Kendall:** 6:35

Well, taking a step back, maybe some background. So I lead Wave, an autonomous driving company, and how? we've set off on a different path to build autonomous driving systems that have the onboard intelligence to drive different vehicles in new places, including places they haven't been to before, and understand the complexity and the long tail of situations that you see on our roads. And this is a, you know, taking an AI approach to autonomous driving is, you know, quite contrarian and different to how people usually look at this problem. When we started six years ago, in 2017, we set off to build an end-to-end neural net that could learn to drive. You know, take the data as input and output, emotion plan to control a vehicle. And this end-to-end AI approach I guess you know why a world model is interesting here. So, for me, it is a model that can understand the state of the world and predict how it's going to change, given an action that you know you put into it. So, mathematically speaking, it's a function that takes your current state, your current action, and predicts the next state. So it's essentially a simulator. It's a model that can allow you to understand how the world will evolve, given different things you might want to do or interact with that world. Why is this important for self-driving? Well, the first thing you might look at when you're building an end-to-end neural network. To drive a car is to build something that's autoregressive. Build something that creates a function that takes your input state and produces the motion plan that you should drive with. And that's kind of what people did in large language models. And the problem with self-driving is it's a safety-critical application. If you make the wrong decision, you're not just going to put out some hallucinated text, but you know it's life and death decisions of driving on our roads. So for that reason, it's really important that you are aware of the implication of your decision and you can understand the dynamics of the world. So that's really what motivated us to start off with world models as a concept, and in 2018, we actually published a blog with one of the first examples of doing this On an autonomous vehicle. We published a model-based reinforcement learning system where actually we're only on a quiet country road, but we learned to drive a car with a world model. So this system had never driven an actual car, it only learned in its imagination in a world model. But it used this model that had trained of the dynamics of the world to be able to learn to operate this car and drive it down a quiet country road and I guess over the last six years I can talk more about Gaia, but over the last six years we have scaled that approach to the point it is today where, with the latest and greatest in generative AI, we can now understand the full, diverse, rich, dynamic urban scenes that we operate in today, like Central London.

**Craig:** 9:22

In building the world model. I mean, the autonomous driving systems to date are helping me out here, but they're primarily reinforcement learning systems that are taking in data from various sensors and have trained on what the best approach is in any particular situation. Is that right, or what is the current state of autonomous driving systems?

**Alex Kendall:** 10:01

Well, we're having an AI conversation and, fundamentally, autonomous driving is a high problem. It's a problem of complex, high-dimensional decision-making, and so you'd assume that you're going to use a data-driven method like reinforcement learning to do it, but actually that's not the case. If you look at all of the large autonomous driving efforts out there today, outside of Wave, primarily the approach is a traditional robotics one. It's one of, yes, you use deep learning for perception, but once you have the state of the world, it's very much a hand-coded optimization approach to produce a motion plan that's aided by a set of infrastructure like HD map high-definition maps that tell the car where and how to behave. So it's actually not an AI approach, and what we've done is, I think, the first time an AI system has actually driven on the roads at this level of scale, so that's not how the industry has traditionally approached things. When you bring in an AI approach, of course, the challenges there are how do you understand what it's doing, how do you make sure it's safe and making the right decisions, and so that's brought up some of these challenges that led us down the road of world models.

**Craig:** 11:11

So well, that's interesting. I didn't know that about autonomous driving systems. So there they have all these sensors, that data is coming into a central decision maker, and you're saying that decision maker is a traditional control system and not a probabilistic AI system.

**Alex Kendall:** 11:34

Probably speaking, yes. A lot of the systems running around in San Francisco today, for example, are of that approach. Now, more and more machine learning has been used throughout over year on year in these systems. But it's not an end-to-end neural net, it's not a large transformer that decides the whole decision making, and that's the step that we've taken to replace that entire stack with one big neural network that learns how to drive end-to-end.

**Craig:** 12:03

And I've seen a lot written recently about using the reasoning power of large language models to play that role, to decide on actions, and then with some other piece of software, to translate that action, to execute on that plan. And can you talk about, well, first of all, for Gaia, the world model. So there's a world model that's building a state of the world in its I guess its weights, and then a reinforcement learning model that learns to act on those on that state of the world. Is that right? Maybe you can describe the architecture a little bit. Yeah.

**Alex Kendall:** 13:05

Let me jump into some of those details. We've got our research paper online that talks about them in great depth. But one of the interesting things for me is that if you look at the three big major trends that we've seen in large language models this year I mean at the start of the year it was all about scale. Everyone was talking about how many parameters, how much data, how much compute are these models trained on. In the middle of the year, it became about multimodality. We pushed scale to some degree and now it's about how do we understand across different modes, and a lot of image tech systems came out, for example. And then, more recently, it's about synthetic data. The benefits of synthetic data are clear: you can control the bias in your training data, you can ensure that the training data is equally sampled across the things you care about, or you can control the distribution of your training data, and you can often get information that is harder to understand from noisy real data alone. And those three trends that have really driven the state of the art and say large language models. The interesting thing is we've seen the exact same thing play out in robotics. So for us at Wave, we've been pushing the scale of our neural network that drives the car, and in the next year our roadmap is going to be pushing this in terms of parameters, data and compute by 100x further, two orders of magnitude further, and so the results, the emergent behavior we're seeing, come out as just remarkable. We believe the car to nudge its way through crowds of pedestrians, to do complicated, unprotected turns, to predict the behavior of other agents cutting in or moving around our vehicle. All of this kind of thing emerges at that level of scale. The second trend on multi-modality that's where I think it's really important to be able to learn to understand between different modes, because ultimately, if you're training a self-driving car just off the video data it has, it's going to be intelligent, but perhaps it will be more intelligent if it doesn't only have that video data but also text and other information sources it has when you and I learned to drive. I learned to drive when I was 16. And maybe my mom and dad probably had 20 or 30 hours in the car with me, maybe not that long. Maybe I learned something like five or 10 hours, I think. I was a fast learner but that kind of lengthened time to learn how to drive. But it wasn't just that that allowed me to drive. It was probably the 15 or 16 years of experience I'd had of learning how the world works, learning what objects are, how things might behave on the roads, and it was that observation that gave me the intelligence to drive, I think that's seeing the same as true in robotics. We can train our autonomous driving system now, not just on the video data of a driver but also internet video and text. We can literally feed it the PDF document of the highway road code that the government writes and give it that as further context to understand. And so I think multimodality is becoming really important to bring together different sources of information and improve the intelligence of your system. But then, secondly, with text specifically, I think the future of how we interact with robots is going to be through language. We are going to be talking to our robots, interacting with them. There's a reason why you and I have evolved languages. The way we communicate is because it's the most efficient way to get information across that we understand, and so, for that reason, maybe not most efficient, but most natural way, and so, for that reason, I think the accessibility of robotics will be greatly improved by us being able to just literally conversely, you should be able to be in your self-driving car and say take the next left, take the next right, drop me off here or I'm worried about this. Why are you doing that? And you should be able to build a sense of trust through it. And we've done exactly that at Wave. We've produced a system called Lingo, which is a first, a vision, language, action foundation model that combines those modalities of video, of action and robotics and language that allows us to talk to our autonomous vehicle and ask it why and what it's doing. And then, finally, the third trend, on synthetic data. This is where Gaia comes in. Gaia is our world model. Not only is it a system that allows AI to understand the implications of the decisions it's making, but also produces synthetic data. Generative AI is very good at recombining data in new ways, and so we have lots of experiences of foggy scenes on the car. We have lots of experiences of jaywalking scenarios, but we have very few foggy jaywalking scenarios. And Gaia can not only allow us to understand how the world's going to evolve, but it can create new examples we haven't seen before. And again, we can do that by connecting vision, language and action. We can prompt it and say, literally, give it a prompt and say I want an example of a jaywalking pedestrian in the fog. Or we can take a scene that exists in the real world and change it and ask it to recreate it with new features and things like that. I don't really get into the architecture of Gaia and I'm happy to, but I just thought those three trends have been really powerful for us in AI and exactly applicable to robotics as well.

**Craig:** 18:18

Yeah, so you can tell from my questioning. I'm a journalist, not a practitioner, so my knowledge is fairly surface. But I understand large language models to a degree. I understand the transformer algorithm and the large language models are predicting the next token in a sequence and because of the volume of training data it does a very credible job most of the time. I understand Jens' JEPA architecture, the Joint Embedding Predictive Architecture, to a degree. He says your model is something different, so can you kind of walk us through it at a very high level of how the model is trained and what it's doing? It's predicting the next state of the world, whether it's video or text or whatever. Is it doing that with a transformer algorithm? How is it doing it? I'm just sure the audience would also like to hear.

**Alex Kendall:** 19:42

Absolutely so. I mean Jan and I share a lot of common beliefs around these systems. We think that to go beyond the autoregressive nature of large language models, of just predicting the next word, and getting to systems that can understand and be safety critical, we need to have world models. We share the vision that these should be unsupervised. They should be able to be trained through self supervision, through whether it signals contrast of learning or building energy spaces or things like this. I think we share a lot in common there. I mean, jepa is a great architectural approach. I think there's a lot in common in these systems in terms of there's some representation space. You want to train it with unsupervised learning, and so for our approach, in particular with Gaia, what we first do is we tokenize up the different inputs, whether it's images, action or language, and essentially, well, today it's a large transformer, but you could use whatever your favorite flavor of neural network or, let's say, machine learning system that you want to use, but essentially you take those inputs and you learn this dynamics, this ability to take current state, current action and predict next state, and then the result of this is a world model that can simulate the future and, if you want to, you can take that state and decode it back to video so you can produce the video output of what's actually going to happen. But you might not want to. If you want to keep this real, time and efficient, you can just stay in your embedding space and use it to drive your car, and right now we're working on Gaia 2 and Gaia Drive, and these are systems that very much are going to see Gaia embedded in the vehicle, able to actually increase the intelligence, understanding and improve the safety criticality of our system in a production setting. So that's really the guts of the architecture Today it's a transformer that's able to predict future states.

**Craig:** 21:54

And again, forgive me, I'm sure you'll cringe at my repeating this back to you in sort of super layman's language but you're encoding the data coming in or tokenizing it and turning it into embeddings in. I guess you call it a feature space or embedding dimension, and you're making predictions based on that at that level. So there and then to see the video, you decode it into pixels but in that space you can predict the. Is it? How specific are those representations? When I said embedding space, I guess I meant representation space.

**Alex Kendall:** 22:59

Yeah, One of the other big challenges of self-driving, compared to large language models, is the amount of data you have as input is enormous. I mean, take our latest vehicle, for example. It's six or seven cameras. They have eight megapixels each, and then you care about multiple frames over time and not to mention, like if you consider an imaging radar as well or choose the sensors you want to use, just in cameras alone, the data eight megapixels, your RGB values there, so 24 million bytes alone from those. Multiply that by the six cameras, then you've got about 120 million bytes, and then multiply that by multiple time frames. I mean you're talking about gigabytes of data there alone. And that data what it does is it makes it you can't have an embedding space for your dealing with gigabytes of data. It's just too much to process, and so to make these models practical, you need to be able to take that extraordinary amount of input, data, all those videos, and compress them into an embedding space that you can reason about. So I guess the question is how many factors do you think you care about for a driving scene? You can list them out. You care about the positions of the cars in front of you, the pedestrians, the cyclists, the direction they're facing, the way they're going to move, the weather conditions, the traffic light, all these kinds of factors. Now, interestingly, if you go down the path of trying to list them out by hand, you end up in the AV 1.0, the classical robotics approach to autonomy, and that doesn't scale because it's very hard to enumerate all those factors and to reason about them a priori. So we can do it as a thought exercise, but I wouldn't advocate for that approach. But the point is that it's not hundreds and millions of numbers. There's probably a much smaller set of things that you care about there, and so we want to learn that you don't care about the pixels in the sky, except you want to know the weather, but you don't really care about all those things. There's a lot of redundant information in this signal. Compare that to large language models. You have sentences of text that are really high signal to noise ratio. Text is precise at saying this means that and impacts this right. It's a direct description of what you care about, compared to videos where most of the pixels in an image you don't care about are Clouds. Second story of a building you don't care about when you're driving. So the point is that you want to take this data and embed it in a very efficient space, and the way we do that is through end to end learning about what do we care about for driving, what actually is going to impact how the world's going to evolve, and that's what we look to learn. So we look to build a transformer and a self supervised learning approach that learns and embedding, that is really efficient, is as small and compressed as possible, but has the information that we need to understand the safety critical natures of the scenarios that we're driving through. So that's the primary task of that embedding model and that learning model of the scene representation.

**Craig:** 26:07

Yeah, and then, right now that you know I've seen some remarkable videos that you've done and maybe I can, you know, splice one into this podcast.

**Alex Kendall:** 26:23

But they're awesome when you drive the car. I go out most weeks and when you see it learn new behaviors week on week, like just yesterday I was when I went for a drive down a part of London and Notting Hill and we went through Portobello Road Market. It's this crazy area with loads of pedestrians all on the road. I've never seen our car sort of nudge its way through a crowd of pedestrians but it did so really safely, in a way that you know if you just stopped and waited for the pedestrians to clear, you'd be stuck there for an hour. So I have these kinds of new behaviors over time. There's tons of videos that we have online of this stuff, but it's pretty amazing seeing AI operate in the physical world like this.

**Craig:** 27:03

Yeah, the videos. So you're then decoding the representation space into pixels in a video. That's useful, I presume, for creating training data. But when you're actually driving in real time, how is that state of the world being translated into action? And that's where I think you said there's an RL engine or agent that's learning over time how to act on that. Can you talk about that part?

**Alex Kendall:** 27:52

Yeah, absolutely. I mean, one of the amazing things about Gaia properties is that you can create photorealistic and diverse scenes that are controllable by text. You can modify these environments and you can also create multimodal futures. I think the really powerful thing is when you only observe the past, you know you can't predict how the future is going to unfold in a driving scene, and the fact that Gaia can generate diverse, multimodal, plausible futures is a really important factor. But, yeah, in order to actually control the car. So this is what we call Gaia Drive. It's not just decoding to images, but also well, there's many ways that you can go about thinking about incorporating Gaia into a driving system. For example, you could generate future data and use that synthetic data to actually just train a system. You could use it to predict the future and use the information it learns about predicting the future to improve the driving representation. Or you could actually bring it into full on. You know, model based reinforcement learning or model predictive control or some kind of learning simulator. What that means is that, let's say you're driving, you're at a green light and you want to decide whether you drive through the intersection. What the system can do is it can run its world on a run Gaia for a few seconds ahead and see what might happen. Maybe it runs it a few times and sees how various different things happen and then it can make a decision based on how it thinks the future is going to play out. We do that in our brains and in our hippocampus we have mechanisms that are, you know, most famously referred to as sort of thinking fast and thinking slow. Thinking fast, the reactive decision making you don't really plan ahead, you just do, whereas thinking slowly you take a step back and use a reason for what might happen. Should I do this? And you sort of play chess. A few steps forward and we can do the same thing in robotics. Robotics has typically had, you know, two levels of control. There's a low level control that runs over 100 times a second that controls parts of a robot, and you have a high level control that operates typically around 10 times a second. That is the high level reasoning. But actually what these kinds of models that you do is maybe move one more level up, abstract and have a three tiered system. You can have a thinking slow thing which can involve a large language model to interact and to reason and to plan could involve a world model to actually understand the implications of these decisions. And that might happen at anything from one hertz, you know, one time a second to maybe even one time every 10 seconds. It's quite an infrequent high level. If you think about when you're driving, that sort of high level kind of topological task planning you do can happen at the highest level. Then that middle tier, you know you're designing the motion plan that you want to follow to ensure you don't hit things that you follow problems. That's more reactive and it runs it at that kind of 10 times a second and the 100 times a second is sort of the minute changes and breaks and steering to make sure that you actually achieve that plan that you've set up. And that three tier system I think is as I think world models can play a really great role in that top one a new level of higher level reasoning into robotics.

**Craig:** 31:17

And the advantages of that over agents built purely off of large language models is that they're there. They require less computation, I would imagine, but also their decisions or their predictions are grounded more firmly in reality, where, as a large language model, even if it's tokenizing pixels in a video, it's only predicting one token ahead as it goes along. Is that right?

**Alex Kendall:** 31:57

Yeah, look, I think large language models and world models can be complimentary, right. Like large language models give you the ability to understand well, give you a text interface. They really interact through language, but they give you the ability to learn a really incredible understanding of the world through internet scale text and world models. The advantages of world models is they give you the ability to understand the implication of your decision. Whether you are making a driving task decision, whether you are outputting sentences and text, you know. Whatever it may be, it allows you to understand, okay, what is the implication of that decision and the environment I'm operating in. Is that a good thing or not? And that understanding can help you do things in a safety critical environment. So I think world models become really important in an application like self driving, maybe not so much like an internet search problem, but when you have safety critical applications whether that's in medical correspondence, of language models or self driving for our application that's when world models give you the ability to do that. The other advantage I described, though, is not just at runtime, but also at training time. World models give you the ability to learn more efficiently. We do this in our brains as well as people. We daydream or nightdream. When we dream, we actually go through a process that solidifies our actions and lets us replay experiences to learn to do them better next time. Yeah, like if you're learning to play tennis and you hit a ball once you know you're not going to hit a ball with every single permutation of the angle that your racket might be, to learn how it might go. You might only hit it a few times, but from that you need to learn the general way of how to hit a ball, to get it in the right part of the court to be able to play tennis and what we do is from the ways that we hit the ball. We actually, you know, replay this in our internal world model many times when we daydream to be able to learn how to do that more effectively. And the same is true with machine learning. When you have a world model, you can get more out of your training data. You can replay, recombine, reconfigure and use that to understand your training data and learn a much more performant, effective or safe policy or decision making system from your training data because of the fact that you can learn and recombine those experiences in new ways. So world models are really powerful with training and inference or testing time.

**Craig:** 34:25

So and you're, you have this system, not only creating training data, but acting as an agent to drive a car. How many, how developed is that system? How far is it from commercialization, for example, how many cars do you have it in and how many road miles have you logged, and that sort of thing.

**Alex Kendall:** 34:55

Yeah, we've been. We've been spending the last six years building a different approach to autonomy, and we were at today as we've been able to demonstrate that it can do a lot of things that have been blocking the industry for many years. It can drive on the kind of equipment that's on modern vehicles today: single, GPU, computer, surround cameras, maybe a forward facing radar. It can drive in different places never been to before and it can drive different vehicle types. So we are very excited to be in the process of commercializing this technology now, to see it deployed across the world's most innovative fleets and vehicle manufacturers and to see this deployed in a way that can utilize value quickly and accelerate the growth from assisted autonomy through to full autonomy. And so we are. You know we are in that process right now. So far, we've been excited to partner with some of the UK's largest fleets like DPD, ASDA and Ocado Group and these are large fleets that do things like grocery delivery here in the UK and those partnerships have been wonderful. We've been delivering groceries throughout this year with our partners, as, for example, in London, at showing some of the value that this autonomy, this autonomy technology, can bring to society.

**Craig:** 36:16

Yeah, just on that. I imagine you still have a safety driver in the car. How is the regulation? I don't want to get lost in that regulation discussion, but how are you managing to operate? As you know, crews in California run into all kinds of trouble, but how is the regulatory environment allowing you to operate fleets, for example?

**Alex Kendall:** 36:49

Yeah, that's a really important question. So today we operate with safety operators, but we have a two-pronged approach here. The first is that we want to see the ability to build value and see deployment of the system still as a driver assistance system or with safety operators, as you say and there's extraordinary value that can be brought there, whether it's the helping, support, safety or improving the efficiency of operating vehicles. There's a big opportunity there that we can see today through driver assistance. But then, on the other hand, of course, ultimately we want to get to level four autonomous driving at scale and we've been really excited about the work that we've done with regulators around the world, but most specifically here in the UK. We've had a number of UK ministers for rides to show them the technology firsthand. We've sponsored the UK's parliamentary working group on autonomous driving technology and we've helped support bringing legislation to be considered to make this technology legal. And we've been offered, and are working on, a 1.9 million pound grant with the government to help understand and put forward a safety framework for AI systems. There's many more activities that we've been doing in the space, but we're deeply engaged with regulators and believe that empowering regulators to understand AI technology and to yeah, empowering them to understand it and to be thoughtful about how best to manage those risks is the best way forward. So we've really looked into our part on that front and have been excited about the results of the traction I should say.

**Craig:** 38:35

Yeah, can you? I have questions about the compute intensity of this and the amount of training data. But setting that aside for a minute, mercedes-benz has, I think, a level four autonomy in Germany, if I'm not mistaken, and the regulations there allow them to drive hands free on certain roads and that sort of thing. How is that system working and how does that compare to a guy of one? And yeah, answer that first and then on.

**Alex Kendall:** 39:29

So my understanding is that I haven't seen a level four system from Mercedes-Benz or other automotive manufacturers outside of the trials we've seen in China or in some parts of the US with some of the technology companies. I have seen a limited level three system, which is where the vehicle does take control and liability of the vehicle for certain scenarios and I believe Mercedes have a product that allows you to do this at low speeds on a highway in rush hour traffic. But in general, the automotive industry is very we get to see as an automotive industry. We get to see technology at scale which can do productionize, which can do driving any vehicle anywhere, whether it's not just highway at low speeds, but urban, suburban, different cities around the world, and that's very much what we're interested in solving. We want to build an AI driver, an artificial intelligence system that has the onboard intelligence to drive vehicles through all of the kinds of scenarios that we expect in our daily lives and the commutes and the travel that we do. We want to build this kind of technology to help assist people, to free up their time, to make it safer on the roads, to give them a more sustainable drive. All of these kinds of benefits. We want to ensure this technology can be brought out quickly and broadly. We think that this is a unique opportunity in space and something that of course the world. Yeah, I think it's a really important next evolution that we need to be able to deliver, to lift up, quite frankly, the safety and performance of all of the cars that we're putting on the road today.

**Craig:** 41:31

Yeah, and for training this system. I mean, one of the challenges, as I understand it, of other autonomous vehicle control systems is they depend very heavily on supervised learning and you have to label enormous amounts of data so that the system recognizes corner cases and there's this very long tail of those cases. And, for example, you can train a car to drive in California, but if you take it to Norway or something with a vastly different climate, it's going to run into trouble if the snow hasn't been labeled into the system and that sort of thing. How much? How do you train Gaia and can it generalize across environments in real time, or do you need to train it in advance for every kind of environment?

**Alex Kendall:** 42:57

The interesting thing is that when you are driving, you're generalizing all the time. You will never see the same thing twice on the road. Every time you go driving, the weather is going to be slightly different than you've ever seen before. You can have cars and other agents around you in different locations. So even if you're on the same road that you've driven you're commuted every day of your life that the specific things that you go through will be different in some way from what you've seen before. So I guess the first point to make is that autonomous driving is all about being able to generalize. We do it every time we operate these vehicles on the road. The question is how far can you generalize? It's one thing to generalize driving on the same road every day. It's another to drive on one road in the UK and then be able to drive in the US, where you have new things like four-way stop signs, right turn at red, other local driving cultures, and so we want to build a system that can effortlessly generalize to new environments to allow us to bring it to everyone around the world. And how we've done that is that our system is trained through unsupervised learning, so it doesn't need boxes drawn around objects or things labeled in the scene for us to learn how to understand the scene. It's all unsupervised. We watch driving data and learn from that how the world works, learn how to predict the future, train models like Gaia and these kinds of things. So it's all through unsupervised learning, and the key thing is that it becomes more efficient the more data you get. So we might need a certain amount of data to train in the UK, for us to train a system and for us to generalize a system to another environment like the US. We will need a fraction of that data because a lot of the experience is shared. Right, the same rules of physics apply in both countries. People tend to behave in a similar way, although there are some differences. You need a little bit more data. We found that going from a passenger vehicle to a van, for example, generalizing to a different vehicle, going from a Jaguar I-Pace passenger vehicle to a 3.5 tonne delivery van, that took us about 2% to 3% of the training data to be able to achieve similar level of performance on the van, and so generalization with N-Twin neural networks can be very efficient. It's like large language models learning to generalize from English to German, to French, to Mandarin, to other languages, and so those are some of the things that we think about when we look to scale our technology and generalize it to new driving scenarios.

**Craig:** 45:29

And what about the computer required to train the models and then to operate the models? And do the models operate through a connection to the cloud, or is there a GPU running in the vehicle? Yeah, so we don't train.

**Alex Kendall:** 45:57

The models live in the car. They are safety assured, validated, before they're deployed on the car. Rather, we take the experience we get driving all the time and we feed that back to the cloud in order to improve the new models that we'll be validating and deploying. So it's all trained at scale and we've been partnered for many years with our great friends Microsoft, and we work with Azure to be able to train at scale. Azure has been able to provide us with extraordinary compute power but, more importantly, the innovation that you need to be able to train these kinds of models. The tough thing with training video scale foundation models, like we do, is the data requirements. I talked about the size of the input set into these models. It's truly ginormous. And to train on this kind of data, you can't have it all sitting on the local cloud. It's tens of petabytes, maybe hundreds of petabytes, since you need to be able to stream it to your GPUs. This means you need a different set of infrastructure from when you're training a large language model. You can't have your data stored locally on the GPUs. You need to be able to stream it from storage that you can do fairly random access across, and that's quite a hard infrastructure challenge. So we've been thrilled to be able to do a lot of pioneering work in this space and we've been able to do a lot of work in the space with Microsoft to make it possible to train models at this scale.

**Craig:** 47:26

And is that, in order to get the video into a, do you vectorize it in order to tokenize it, or how does that work?

**Alex Kendall:** 47:48

Yeah, we take the video data or the experience for all kinds of data and, yep, it's tokenized, it's fed into the transformers at scale and that's how it's trained and it's not just video data, it's also the navigation prompt, it's the other robot sensors like GPS, wheel speeds, things like this. All these kinds of input data we use and feed into the AI.

**Craig:** 48:14

Yeah, I mean. The reason I'm asking about the computer requirements is that these large language models are amazing. They're people that are scaling them 10 times or more from what we've seen so far. But there's a limit in the availability of GPUs for the time being and for the foreseeable not the foreseeable future, but for at least a few years. And then, because of that constraint, there's a limit on the amount of inference. Anyone, a customer, can use the model for these rate limits. Does a world model face those same constraints?

**Alex Kendall:** 49:15

Sorry, Craig, can you clarify your question? Constraints around rate limits?

**Craig:** 49:20

Yeah, just on the availability of GPUs both for training and for inference, because it costs an enormous amount of money to train an LLM, but it also costs an enormous amount of money on the inference side and people are accessing these models through APIs and because of the costs and the compute constraints they can only process so many requests a minute or so many tokens a minute and that is really limiting the enterprise scale applications. So does that same kind of problem run through a world model? Or is a world model less fundamentally less compute intensive and so you can get around those problems?

**Alex Kendall:** 50:17

Well, the great thing about our space is that at inference time, most of the compute runs on the car and so, assuming you've got vehicle assets to deploy on that are there and fleets and are generating value, then you're in a great spot because you can run all the inferences you need on the vehicles themselves. So our challenge is really training time. Yes, there's some inference costs, but it's more manageable because it's a hybrid-edge cloud model and primarily you need to have most of it on the car because the car should be able to have all the intelligence it needs to be safe and make the kind of decisions it needs to operate in an environment on board the vehicle. So training time really matters. And there, yeah, I mean we are hungry for all of the compute and data storage that we can get to be able to power these models, and I feel fortunate to be writing Moore's Law and other year-on-year improvement curves that just bring down the cost and increase the availability of increasing orders of magnitudes of these systems. But no, we certainly do have an appetite to lift what we've got another 100x to next year in terms of scale of data and compute and parameters in this model. Training computers is a really, really important factor for us.

**Craig:** 51:40

Okay, then quickly, two questions. Is Gaia open source? And I've been writing a lot about why we don't see humanoid robots the way everyone wants to see, and one of the you know they're the hardware problems, but the other problem is having a reliable AI brain. Could this model be applied to other forms of embodied AI robots?

**Alex Kendall:** 52:17

That's an awesome question. I'm really bullish on humanoid robotics being a part of the future, the technology we're building. We want to see this kind of embodied AI system empower all kinds of robots, whether it's manufacturing domestic robots or self-driving cars. But I agree with you. I think self-driving cars will be the first big application of embodied AI at scale, because there is data, there are hardware platforms, there is a business case and so it can be built. Today, I think getting the data and the hardware platforms for humanoid robotics is is harder, but I hope that the scale of embodied AI we can build through self-driving can make that easier by taking that technology and adapting it to humanoid robotics in the future, and I would love for Wave to be part of that. Once we've got our self-driving embodied AI systems to scale. That's really the path we see. So I agree with you on that one, but I'm very bullish on in 10 years time, ai not just being chatbots and co-pilots but being all kinds of physical embodied AI systems in the world that we live in.

**Craig:** 53:25

Okay, and then the question on open source. Is Gaia open source?

**Alex Kendall:** 53:30

So, gaia, we have written an extensive research paper around it, and we're very much fans of openly engaging with the AI community and sharing ideas. We haven't open sourced the model itself for now, and that's something that we continue to iterate on and develop internally and to use to deploy on our fleets with our partners.

**Craig:** 53:52

Hi, I wanted to jump in and give a shout out to our sponsor, netsuite by Oracle. I'm a journalist and getting a single source of truth is nearly impossible. If you're a business owner, having a single source of truth is critical to running your operations. If this is you, you should know these three numbers 36,000, 25,1. 36,000, because that's the number of businesses that have upgraded to NetSuite by Oracle. Netsuite is the number one cloud financial system streamlining accounting, financial management, inventory, hr and more. 25, because NetSuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customized solution for all of your KPIs in one efficient system with one source of truth: Manage risk, get reliable forecast and improve margins Everything you need all in one place. As I said, I'm not the most organized person in the world and there's real power to having all of the information in one place to make better decisions. This is an unprecedented offer by NetSuite to make that possible Right now. Download NetSuite's popular KPI checklist, designed to give you consistently excellent performance, absolutely free, at netsuitecom. Slash I on AI. That's I on AI E-Y-E-O-N-A-I all run together. Go to netsuite.com, slash I on AI, to get your own KPI checklist. They support us, so let's support them. That's it for this episode. I want to thank Alex for his time. If you want to learn more about the conversation we had today, you can find a transcript on our website. I'm on AI . That's E-Y-E-O-N-A-I Remember. The singularity may not be near, but AI is changing your world, so pay attention.