The Future of Work podcast is a weekly show where Jacob has in-depth conversations with senior level executives, business leaders, and bestselling authors around the world on the future of work and the future in general. Topics cover everything from AI and automation to the gig economy to big data to the future of learning and everything in between. Each episode explores a new topic and features a special guest.

You can listen to past episodes at www.TheFutureOrganization.com/future-work-podcast/. To learn more about Jacob and the work he is doing please visit www.TheFutureOrganization.com. You can also subscribe to Jacob's YouTube channel, follow him on Twitter, or visit him on Facebook.

Jacob 00:01
Hey everyone, thanks for joining me for another episode of the future of work with Jacob Morgan. My guest today is Brian Christian. He's the author of several best selling books, including The Most Human Human, Algorithms to Live By, and his newest book, which is called The Alignment Problem: Machine Learning and Human Values. Brian, thank you for joining me.

Brian 01:40
Thank you.

Jacob 01:42
Very first question for is why did you write this book? What was the impetus for creating this?

Brian 01:50
So the story of this book goes back, well, really goes back to around 2014. And I had several experiences, where, you know, I had been in the AI community for some time, I'd written these earlier books that you mentioned. And I had been noticing the public conversation around AI really starting to take a turn. People, you know, when I would go to speak at a company or whatever people went from asking me, is AI going to take my job? To asking me is AI going to destroy the human race and life as we know it?

Brian 02:32
And so you know, you we started seeing people like Stephen Hawking, Elon Musk, taking this really kind of existentially grim view. And in fact, I was actually at a Silicon Valley book club, talking about my first book. And Elon Musk showed up and buttonholed everyone, including me on, you know, give me one good reason why we shouldn't be worried about AI, or one good idea about what we should do about it.

Jacob 02:59
Wait, he just randomly showed up,

Brian 03:01
No he, he was invited. He wasn't a normal member of the group, but he was invited. And I didn't actually expect him to turn up. But he did. So it's very, very kind of thrilling. And, yeah, he's sort of commandeered the conversation at one point, and basically forced everyone to explain why he should

not be worried about AI, or what we should actually do about, you know, our concerns around AI. And to me, this was an interesting moment, it was sort of a turning point in my own thinking, because I'd certainly been aware of this conversation of, you know, should we be freaked out? And I, I found it to be honest, a sort of a recreational topic, it was like something that, you know, we would talk about it over cocktails or whatever.

Brian 03:47
But it was that experience of being put on the spot forced me to realize that I didn't have any good reasons why we shouldn't be worried. And so that really started my cognitive dissonance of, Okay, I can't convince myself that I shouldn't be more worried than I am. So that maybe means I should be more worried than I am. So that started this long evolution for me of really starting to tune in a little bit more to the things people were saying, in that community. And you started seeing these concerns moving from kind of celebrity engineers and scientists like himself, Stephen Hawking into the actual computer science community itself.

Brian 04:29
So people like Stuart Russell, who is the co author of kind of the most well known AI textbook, started making some of the same arguments and then it was like, okay, you know, the fire alarms have effectively been pulled. And to roll the kind of roll the clock forward to 2016, which is when I started working on the book. By then you started to see two things. One was that these concerns that originally, kind of were most pronounced Outside of computer science, and then later echoed within computer science that had turned into like an actual scientific research agenda.

Brian 05:08
So by 2016, you know, open AI had been founded DeepMind had hired their safety team, the Center for Human compatible AI at Berkeley had been founded. And there was an actual research agenda underway for what's called technical AI safety. So where the rubber hit the road, we are actually doing things now there's this incoming generation of researchers, PhD students trying to do the work of making AI safe. At the same time, you had this increasing and sort of escalating series of public catastrophes. So you had Google Photos, captioning, this photo album of Jackie Elsie Nye, who's a web programmer and his friend, as gorillas. And they're both African American, you had the widespread use of algorithms and criminal justice, which was raising all these ethical alarms.

Brian 06:06
And so it started to me to feel like we were at a pretty significant crux in the history of AI. We have, on the one hand, this escalating series of public disasters, and on the other hand, this incredible growth of the actual scientific field around these questions of fairness, accountability, transparency, long term technical AI safety. So that was the story I wanted to tell. I thought this is one of the most high stakes, but also most fascinating things that I could see happening in computer science. And I wanted to try to tell that story through the perspectives of the researchers themselves. So that's really the project that I embarked on.

Jacob 06:52
Very cool. And yeah, I mean, of course, we've all seen enough science fiction, movies and read enough science fiction books to know that AI always wins and takes over the world and slaves, all humans. Usually what ends up happening, right,

Brian 07:06
Yeah, and I mean, in particular, you know, within the AI community. There are these thought experiments like the so called Paperclip Maximizer, where it's not so much that AI enslaves everyone out of some kind of misanthropic, you know, rage or whatever. It's like, we were just not quite precise in what we asked it to do. So it's sort of...

Jacob 07:30
Can you share the paperclip maximizer for people who are not familiar with with that story.

Brian 07:35
Yeah, I mean, this is kind of a modern version of the Sorcerer's Apprentice, you know, this idea of be careful what you wish for. So in the paperclip, Maximizer it's this paperclip company that builds this sort of all powerful AI. And they task it with maximizing the production of paperclips, because that's what their company does.

Brian 07:59
Unfortunately, you know, this is like the Midas touch, etc. Be careful what you ask for where the AI is so powerful that it turns the entire world into paperclips and, you know, destroys all living creatures and harvest the carbon from their bodies in order to make more paperclips. So that's the kind of thing that keeps computer scientists up at night. So this is where the title comes from. It's called the alignment problem.

Brian 08:25
So are the things that we really want these systems to be doing, actually reflected in the objectives that these systems have and the way that they behave? And this turns out to be an extremely deep and extremely interdisciplinary question. So it touches on not just computer science, but psychology, ethics, the law, cognitive science, etc, etc. So yeah.

Jacob 08:52
Yeah, lots of different things. Well, before we get into more detail into some of the things you talked about, in the book, I thought maybe we could look at a little bit of the history of AI because it's also very fascinating of even when this first appeared in, in history books, I remember, in like Greek mythology, there was the story of I think it was Telos, like this Greek mythical figure that was created, and it was an automaton. And so like, this concept of AI and creating something that's better and stronger and smarter than us. I feel like it's been around for 1000s of years. And then leading up to you know, there's of course, the deep blue and then we had the Terminator, like just so much stuff leading up to where we are today. Can you give us a brief kind of history about what some of those pivotal roles in moments were that that brought us to where we are now?

Brian 09:43
Sure, for me, the..a lot of what we're talking about when we're talking about AI in the 21st century, and in particular, the 2020s. We're really talking about a subfield of AI called machine learning. And the basic idea of machine learning is how to get computer systems to do things without explicitly programming them. So essentially just showing them examples and hoping that they'll kind of get the gist. And we'll kind of internalize that pattern and go generalize that behavior into the future. So of course, there's this fear of like, well, are they really learning what they think we want them to learn? Are there examples that we showed them really representative of what they're going to be asked to do later.

Brian 10:35
So for me that the story of machine learning really begins in the 1940s, the early 1940s, with this really poignant collaboration between Warren McCulloch and Walter Pitts in Chicago. So at the time, Warren McCulloch is this kind of 40 something mid career neurologist, he's just come from Yale, to the University of Chicago. And Walter Pitts is a 17-18 year old, homeless, math prodigy, who is just kind of hanging out around the University of Chicago campus. They strike up this kind of unlikely friendship, Warren McCulloch becomes his foster father essentially, moves him into his basement, and they stay up all hours talking about the brain.

Jacob 11:28
It was the AI winter, I think they called it right. And then like the AI Golden Age,

Brian 11:28
And so you have Walter Pitts, this logic prodigy, and Warren McCulloch, this neurologist, and together, they start to think about the activity of neurons in mathematical terms. And this is really to make a long story short, this is really the beginning of what we now know is neural networks. So it's one of the earliest ideas in computer science, but it took essentially until 2012, for the idea to really bear fruit. So I'm now compressing a lot of history of computer science. But basically, neural networks, this idea that we're gonna kind of directly mimic the structure of the brain comes about in the early 40s is then excitingly debuted in real machinery in the late 50s. But then, like, refuted mathematically in the late 60s as like, okay, this is actually a dead end is never gonna work, then revived in the 1980s.

Brian 12:14
Exactly, right. And yes, you have this revival in the 80s. But the problem in the 80s is that our computers were too slow. And we didn't have enough data. So we gave up again, but not for any fundamental reason, just because the time hadn't come yet. And so it was really in the year 2012. When you now had the ability to collect data at scale online, you know, if you needed if you had a model that required hundreds of 1000s of examples in order to actually learn, well, it was no longer a problem, because you could just go online and get hundreds of 1000s of examples of something. If it needed, you know, what in the 80s would have taking taken millions of years of compute? Well, by the 2010s, it was now possible in about two weeks.

Brian 13:28
So that really was the turning point was 2012, which launched what you could call the kind of the deep learning revolution, and everything that has come since so whether that's image recognition, you know,

automatic medical diagnosis, self driving cars, AlphaGo, alpha zero, et cetera, et cetera. All of that has followed directly from this kind of breakthrough in 2012. And I think it's hard to remember what things look like before then it's, it's very, very rare to see that much change in that period of time.

Jacob 14:10
Yeah, no, it's, it's pretty interesting evolution to see. Well, alright, so first, kind of high level question about the book and some of the things that you're exploring in there. When you think about AI in general, machine learning, what's your general perception of it? Good, bad. I know, it's super high level, but I want to start there and then we can narrow in on some some themes.

Brian 14:38
Yeah, I mean, it's it's extremely powerful and extremely dangerous for essentially the same reason right, which is it allows you to do things at scale that you don't know how to explicitly program. So you know, identifying recognizing objects and images is one classic example of, it's really hard to write code that would tell you how to identify, you know, a whisker on a cat's face or an ear or whatever, in a sufficiently general way, but you can just train the machine to do it. And so that's incredibly powerful. Because it now enables you to essentially have software doing things that you could not program software to do.

Brian 15:25
But it's really dangerous, because the very fact that it's doing things you don't know how to program indicates how difficult it is to debug. It's not operating in a normal, I mean for people who have a software engineering background, like there's this whole set of techniques that you can use to debug software with breakpoints, and, you know, integration tests, and all these sorts of things, that doesn't really apply to these cases. And so we need to basically reinvent a new safety, infrastructure for making sure that these systems learn the things that we think we're trying to teach them, and we think they're learning and to make sure that they're actually going to be safe when we deploy them in the real world.

Jacob 16:14
So are you are you saying that when we, I guess try to teach AI and these algorithms to do something that oftentimes when it makes a decision, or comes to a conclusion, we don't necessarily understand how I got to that point.

Brian 16:28
Yes. And often, it's sort of follows the path of least resistance. So an example that I give is, there was a PhD student named Willand Decker, who was trying to build a neural network that would detect whether there was an animal in a landscape picture, or whether it was just a empty landscape with no animals in it. And so he did the sort of normal deep learning things where you collect 1000s of examples of one category 1000s of examples of the other, and then you essentially tell your system, you know, figure it out. But what he discovered when he went back, he was able to use a set of techniques to figure out essentially, what part of the image was the most kind of significant or relevant to his model in order to determine whether there was an animal present, he assumed, of course, it would be the actual part of the image where the animal appears. But instead, what he found was the

system was looking exclusively, like, basically at the horizon, it was looking way into the background, and simply trying to detect whether the background was blurry or not.

Brian 17:41
Because his model learned that most landscape photographs of an animal put the animal in focus, which makes the background very blurry. But if you're just taking an empty landscape portrait, you set the focal length really far, and the background is very sharp. And so he thought he had built like a face detector, a fur detector and mouth, nose ear detector. But in fact, he had just built a blurred detector. So this is, I think, a very classic example of how, in some ways these models find kind of the path of least resistance. And often what they, if you will, what they think these categories mean, could be very different from what you think they mean. Right? Intuitively.

Jacob 18:30
Can you talk a little bit about how, and I know this, I'm trying to avoid making be, like, overly technical, but the idea of an algorithm and how it's created, because algorithm is basically a shortcut or a formula to get something done. Right?

Brian 18:46
Yeah.

Jacob 18:47
So how, how are algorithms created? I mean, I don't know if there's like a simple example we can think of so I want to do X, Y, Z. I mean, like, what's, what's the process of trying to put something like that together?

Brian 19:01
Yeah, I think so. It's gonna, needless to say, vary from one application to another. And I also, I, I'd want to draw a distinction between what I would call algorithms in kind of the traditional sense, which is, you know, explicit, an explicit series of steps that you write in code and sort of the old fashioned if x, then y, you know, then do this sub routine or call this method versus the, this new generation of tools, these kind of machine learning models. You can think of them as algorithms if you want to, but for me, it's more helpful to think of them as like, models or systems.

Brian 19:46
And so how are those models trained would be the technical term. The basic way that it works is you would say, Okay, I have some tasks that I want the system to do. And you're going to need to translate that into two things. One is a set of examples, which is called the training data. And the other is how you are going to mathematically define or, the technical term would be operationalize what you want. And so this is called the objective function. And so if you're building a system to classify whether something is a bird, or a plane, or a cat or a dog, that's going to involve somehow gathering hundreds of 1000s of images, somehow labeling them with what that image is.

Brian 20:36
So probably, you're going to go to Amazon Mechanical Turk, and pay people, you know, a few cents to tell you what's in that picture. And then you're going to feed this through neural network, and you have some objective function that defines mathematically the task that you want to perform. Now, typically, in image recognition, it's something called cross entropy loss, which basically says, you know, take a guess, between categories, one through 100. And if you're wrong, you get minus one point. And if you're right, you get, you know, zero points. And, you know, you want to minimize the amount of loss. So everything that happens in between is the result of the training procedure. And you know, the technical part of that we could get into Stochastic gradient descent and back propagation, and how, how do you actually adjust the parameters of the model for each example.

Brian 21:28
But I think, for most listeners, suffice it to say the training process is, you pull an example, at random, you run it through the network, and the network spits out what it thinks that image is, you compare that to the actual result. And then you fine tune all of the parameters inside your model, such that they slightly more closely approximate what the correct answer would have been, then you throw that image back in the hopper, and you pull a new one out at random, and you do the same thing over and over again. And then the miracle, if you will, that happens is that after you've done this enough times, not only is the model accurate for the examples that you've given it, but ideally, it's also accurate for examples that it's never seen. So that's, that's what's called the generalization. So that's, that's basically how I would describe how these work.

Jacob 22:22
Okay. And earlier you mentioned AlphaGo, and AlphaGo, zero, and I'm a big chess nerd, so very much involved with chess. And I know there's been a lot of implications in the chess world for this. And what I found fascinating between like AlphaGo, and AlphaGo, zero is I think, AlphaGo, where it learned to play the game of Go took that first approach, right, it looked at tons of examples from top players and analyze that. Whereas AlphaGo zero, if I recall, correctly, didn't look at any games. They just inputted the rules and said, this is how you play, go figure it out.

Brian 22:56
Yes.

Jacob 22:57
And the version of AlphaGo that learned to play on its own, crushed the other version of go that learn to play by looking at top players, I think it was like 100 to one or 100 to zero, which is like insane. And there's a lot of implications for this in chess as well, because even a lot of top grandmasters the way that they've learned to play chess, they're like, wait a minute, the computer, the algorithm is playing in a completely different way. And it's just changing the way that we think about chess and games like go. So it's, it's really fascinating how that stuff works. But there are implications also, for us, right? I mean, for society, how we live for business. Can you give a couple examples of where we're seeing these types of machine learning and AI, just in our everyday lives? Or in our organizations?

Brian 23:48
Yeah, absolutely. So yeah, first of all, you've drawn a really useful distinction, which is between what I was describing a moment ago, which is called supervised learning, where you just have a set of examples and you say, this was the right answer, you should do more like that. Versus a more sort of elaborate or advanced version of machine learning called reinforcement learning where you say, you're in an environment where you can take actions, and some of them will produce rewards, some of them will produce penalties and just do whatever it takes to get as many rewards as possible. And so both of these types of machine learning systems are just about everywhere you look in modern life. So I think this is, you know, to a degree that some of us don't quite appreciate.

Brian 24:41
For example, we're now seeing you know, in medical diagnostics, a lot of you know, you can feed a picture you take with a cell phone of like a lesion on your skin into one of these neural network models, and it will tell you with diagnostic accuracy comparable to a human physician You know, is this a malignant lesion or not? You know...

Jacob 25:04
That's terrifying.

Brian 25:07
But we're also seeing it in, of course, ad targeting, you know, advertisers are people have this intuition that advertisers are tracking your clicks, which is true, but it's that's sort of not even the half of it. In many social networks they're tracking how many milliseconds a particular image was on your screen. So even if you slightly hesitated your finger, as you were scrolling along, that's being recorded. And that's being used to train these giant models of what kind of stuff they want to show you in the future. So social media is basically just one giant, you know, machine learning system.

Jacob 25:51
Yeah, that's crazy. Well, so a lot of people listening to this, I'm sure there are some people listening and watching this who are more technical, and probably relate to a lot of the concepts and ideas that you were talking about. But I'd say a lot of people are not as technical. And they're probably wondering, Well, why do I need to care about AI and machine learning and algorithms? I'm not in that space. So why should non technical people, you know, just the average person like me care about this stuff?

Brian 26:19
I think this is really important question. I'm glad you asked. I think that this is increasingly becoming part of the the core curriculum for being a citizen. And in this current century, whether you consider yourself technical or not, is to have some working knowledge of how these systems operate, and importantly, when you can expect them to fail. So to give you one example, let's say you're a judge, and you've been sitting on the bench for 30 years, and you never learned about computer science or machine learning. Well, suddenly, you're finding in a number of states, including California, where I live, is there is an increasing mandatory use of machine learning statistical risk assessment models to assess things like parole, probation, pretrial detention, etc.

Brian 27:14
Suddenly, you need to have a sense of okay, what does it mean now, when someone I'm, I'm at an arraignment, hearing someone comes in, and I see that there is a risk assessment of 8 out of 10 likelihood of failure to appear at their court appointment. And it's up to me to decide whether to keep that person in jail, or one of them go back to their family waiting their trial, having some understanding of what what are these models really do? How can I tell whether someone might be an exception to this general pattern that that model has found?

Brian 27:51
You know, if you're driving a car that has autopilot, right, whether it's a Tesla, or you know, many cars increasingly have some version of autopilot in them. Sudden, suddenly there's a freak snowstorm, and the sun is at a really weird angle, and it's reflecting off the snow and so forth. Having some intuition of this is a very rare set of circumstances that is likely to have been underrepresented in the training data. Having that intuition might help save your life, right, that might be a clue that you need to disengage autopilot, because, you know, the road doesn't look like a road would normally look. And so you know, all bets are off.

Brian 28:35
So I think there's so many areas in life, I mean, if you're just trying to get a mortgage right now, there are machine learning models that are determining whether you're credit worthy or not. And so having some knowledge of how they work can help you kind of steer your own life. So at each of those scales, I mean, it's for anyone who uses social media, right, your your preferences are being studied, they're being modeled, you're in interacting with this kind of AI system. I think this is just the world we're in. And so, you know, that was one of the goals that I had with the book is that I think this has become, we're at a point where we just everyone needs to have a little bit of a crash course, in order to have a working knowledge of how the world around them is functioning, because this is just sort of it touches everything at this point.

Jacob 29:19
It's actually a little scary to think about because so much of how we live in work is determined by algorithms and technology, from what you're recommended to when you log into Spotify, to when you're looking to hire somebody and you have software that's determining if the person is a good fit, to I mean, everything. Even if you're looking to go out on a date, right, you're on Tinder or match.com You know, one of those sites and you just you don't know why and how things are being recommended to you just know that they're there. And I don't think a lot of people take a step back and say, Well, wait a minute. Why is this restaurant being recommended Why is this person a good fit? Why should I not hire that person? But it's, it's kind of creepy that it's almost like the code is determining how we live.

Brian 30:12
Very much

Jacob 30:14
It's kind of creepy actually, when you think about it

Brian 30:18
It is, and I think a lot of us have this increasing sense of, you know, the world just becoming a little bit inscrutable, a little bit like, we just don't really know, we're just sort of subject to the whims of whoever Tinder shows us this week, or whatever it might be. And I think it's, it's worth unpacking, you know, there are, there are a couple different layers at which these algorithmic decisions get made, you know, at one level, there is the business model of the company. So, you know, Spotify, for example, has to keep both the users and the bands happy. And so there's all sorts of things being done behind the scenes at Spotify in order to keep smaller record labels, you know, content with being part of Spotify. And so they're making sure they include a certain percentage of artists from certain types of labels and playlists, and there's all sorts of things that they're doing there, to basically try to keep all parties, you know, engaged, every, every services business model is going to be a little bit different. And that's going to affect how they're optimizing.

Brian 31:26
There's also times when companies just internally adopt some proxy metric that they think is going to be useful for growing the product, and they optimize against that. So a memorable example was for a longtime Tinder to use your example, what their engineering team was directed to optimize number of swipes per week. And that would be measured in terms of any new user interface change, did it make swipes per week go up or down? And, you know, anecdotally, I would hear from my friends, man, it feels like all I'm ever doing is just swiping endlessly, and I'm not really having conversations, I'm not really going on dates. And it's like, well, that is that's what's being optimized for, you know, just just because it was an easy thing to measure. And it seemed at first like it was a proxy for, you know, the overall experience. But eventually you realize that, you know, it's not right, it's sort of comes from what we really care about.

Jacob 32:30
One of my favorite examples night, I interviewed her a couple years ago, I think it was three or four years ago, Kathy O'Neill wrote a book called weapons of mass destruction. And I was asking her for one of her favorite examples of when something like this goes wrong. And she was telling me this story, I think it was a school district in in Chicago or New York, and they had some sort of algorithm, some sort of technology in place to evaluate whether certain teachers should be fired or not. And they were running this algorithm, and they determined that this, you know, this list of teachers should be fired.

Jacob 33:07
And one of these teachers was very confused. And she said, Wait a minute. My students love me, my teachers love me, you know, I don't understand, like, how did this piece of software determine that I shouldn't be like, oh, like, I want to see how this works. And it turns out that what happened is that there were a lot of students that came from another school. And from that other school, there was a very high incidence level of cheating. And so the cheating actually happened on behalf of the teachers, because the teachers, if their classes perform well, they get raises and bonuses. So when standardized tests were being taken, the teachers would go in there, and they would fill in the correct answers on behalf of their students, and the students didn't even know.

Jacob 33:21
And then all of a sudden, these students transferred into a new school, and they got into this new teachers program, and their scores plummeted. And the algorithm determined, well, wait a minute, they had high scores. Now they're in this new school, and now they have low scores, it must be the teachers fault. So this teacher should be fired. And it didn't take into consideration that there was this whole cheating scandal that was going on. And so I really love that story. Because I mean, I don't know how you can possibly account for something like that with, with technology. I mean, it doesn't seem like that's possible.

Brian 34:23
I mean, I think this is a perfect example of so many different things that can go wrong, right, like so for one thing, there's kind of a transparency issue, whether, you know, the teacher maybe has no insight into how this model actually works. You know, and we've seen cases in the criminal justice system where defendants have come to believe that there was simply a typo on the form that was submitted to the whatever risk assessment and even figuring out what data was actually fed into produce the score that was assigned to them. requires all sorts of legal maneuvering and things like that. So there's a huge transparency challenge.

Brian 35:07
And yeah, there's also this real question of how exactly do we define what we want these systems to do in numerical terms, like at a first approximation, rewarding teachers who make their incoming students grades go up. Sounds reasonable. The same way that, you know, Tinder, increasing swipes per week sounds reasonable. And this is really the heart of what computer scientists know as the alignment problem, which is, it's really easy to come up with a kind of working model for what you think you want the system to do. And almost inevitably, I mean, the same thing as with the paperclip, Maximizer, right? You, like clockwork, realize there was way, way more complexity in the system, then you accounted for the things that you wanted were way more nuanced than your like, simple, working KPI that you were trying to optimize. So that ends up I think being a cautionary tale that, unfortunately gets reproduced time and again, from, you know, across many, many different fields.

Jacob 36:18
Yeah, and I mean, it only takes one example for something to be proven wrong, right? I mean, it's like I remember in math class, what was it doing proofs? And you had to basically show the, you know, I hated that, because my worst subject ever, is doing proofs. And basically, if you find one thing that disproves the proof, then it's wrong. And it seems like it's very similar for this world as well. You might have this perfect idea, this algorithm, this this way of thinking about something, but if it gets disproven wrong, right, one person dies from a car accident. One mistake happens, and it's wrong. Like you can't like like, should we have that 100% expectation from these types of things?

Brian 37:07
Interesting question. I mean, there has been a lot of interest from the AI community in to like, aerospace safety. You know, the kinds of practices that we have, for, you know, making airplanes, the same process we have for like architectural safety or civil engineering. By default, machine learning has no such guardrails, right, by default. The way that it works is it has some mathematical definition of

what it's trying to do, and some set of examples. And it just pulls those examples at random and tries to turn the knobs to match the desired output. And it repeats this ad infinitum.

Brian 37:49
And so by default, in a way, it's almost the opposite of the proof example, where if there's kind of one, one piece of training data that doesn't quite fit the pattern. But there's hundreds of 1000s of training data points, the model may in effect, just decide to take the hit and get that question wrong. Rather than learn some super complicated pattern that it has to identify in order to be able to disentangle those things. And that can end up being a real challenge. It has obvious implications for things like disparate performance on minority groups, for example, you know, if you're building a credit score algorithm, and let's just say you don't have a lot of data about a particular minority group, well, because you don't have as much data about them, you are going to model them less accurately. But because you're modeling them less accurately, then you're not going to be able to, you know, give the loans to the most creditworthy people. And so then it's going to look to your model like those people, that particular group of people is less credit worthy on average than other groups. But that's not true. It's just a function of they were underrepresented in the training data.

Brian 39:08
So in some ways, I would say the machine learning needs something closer to what you're describing. Which is, you know, if there's some subset of these data points on which you're not performing well, don't just like, average that out with the larger group of people on which you're performing well and say, Okay, I'm, you know, I'm coming out ahead, that's fine. But rather, you know, take take an opportunity to kind of stop and figure out, you know, what, what's missing? Maybe we need to collect more samples from that population, or whatever it might be.

Jacob 39:41
I also wonder if we're going a little too overboard with with the stuff. So for example, we see this a lot in hiring right? I mean, I remember when you would apply for a job, way back in the day you would apply for a job and you would talk to a human and a human would call you and you'd have This human relationship and today, you apply for a job. And now you're like playing a game on your phone that looks at your soft skills, you're talking to a bot that looks at your, in other words, you're going through like all these things around technology, something is scanning your resume looking for keywords, and you go through all these steps. And if you're lucky, if you don't get flagged in any of these things, then maybe you get to talk to a human.

Jacob 40:28
And I remember I was talking to Nolan Bushnell, a couple years ago, he was a guest on his podcast, and he was the first boss of Steve Jobs. And he was telling me the story about how he hired Steve Jobs. And he's like, yeah, you know, I, I met the guy. I interviewed him for a couple hours. And I gave him the job. Like, then and there, I just shook his hand and he started working. It didn't take three months, he didn't play any games, he didn't get filtered through now. Like, it was just a human thing.

Jacob 40:58
So part of me wonders if we're removing the human aspect from a lot of our work, and a lot of our life that should just be human in hiring, interviewing people. I mean, shouldn't that just be a human thing? Why are we bringing technology into that?

Brian 41:15
I think this is such an interesting example. Because it's double edge. I think you can imagine someone critiquing the Steve Jobs anecdote and saying, well, this is a perfect example of privilege, right? Steve Jobs happen to already be in Silicon Valley, you happen to be friends with this guy, he shook hands, they had some coffee, and then he had a job. Well, what about all the people that can't afford to live in Northern California unless they already have a job? Or what about the people who, you know, wouldn't happen to be friends with this guy through his social network, etc, etc. And so I think there is a well intentioned push towards this more meritocratic society where we say, No, anyone in the world who wants to apply for this job, apply, and we're going to pick the best candidate out of everyone, rather than just going through the Friends of our existing employees.

Brian 42:05
I think there's something totally, you know, well intentioned about that. The problem you have on the backside is that you then end up with more job applicants, than you have the ability to actually filter. And so then you turn to something like this machine learning tool. And you know, there are many horror stories about machine learning being used in hiring. One of my favorites comes from Amazon. They built a tool in like 2016 2017 to,

Jacob 42:32
Oh, I remember that.

Brian 42:33
Yeah, it would, yeah, it would rank job applicants on a scale, I think it was from one to five stars. So they Ironically, the same way that users on Amazon ranking products. And one of the real problems with the model that they built was to make a long story short, it would look for sort of any linguistic patterns in common with kind of the hires that they had made in the past. So any any language on their CV or resume application. And just try to hire more people like that. The problem was that in the past, their engineering hires had been overwhelmingly male. And so upon closer inspection, they realized that their model was rejecting candidates or penalizing candidates who used the word women's in their resume, you know, I played on the women's soccer team, or went to a women's college or whatever.

Brian 43:29
They were able to delete this particular part of the model. But you know, on even closer inspection, was penalizing people for talking about, you know, playing field hockey, or sort of female, skewed hobbies, sports, ultimately ended up Amazon just ended up scrapping the model entirely, because it was in sort of inextricable, they couldn't get rid of all of these really problematic associations. And so I think that leaves us with a genuine conundrum, to your point. Where, what do we do, when, you know, we don't want to just go with the Friends of our existing employees, we can't filter, you know, hundreds of 1000s of applicants for given job. So I think there's, there's a genuinely complex societal question there, which

I don't have an answer for. But I can certainly highlight the horror stories from machine learning and kind of point to those as as a way to say this is not the silver bullet that we might think it is.

Jacob 44:32
Yeah, Because on the one hand, you want to be able to hire somebody quickly and just kind of go with intuition and what it's, you know, just meeting somebody and going for coffee. But on the other hand, like you said, you want to avoid any of these biases and make sure that you're hiring the best candidate and being fair. But then on the other hand, if you use algorithms and technology, it inherently has some biases in it, because it's created by human beings. So it's, it's tough. Yeah. I mean, why I want to say what's the solution? But what are you seeing organizations do? Like, has anybody figured this out? Is anybody doing a good job of, of using this stuff?

Brian 45:11
Good question. This particular story ends with Amazon dismantling not only the tool, but the team built the tool and just kind of starting over. So I don't, I don't know where they ended up, to be honest with you. I mean, I don't think they're eager to talk about it. I think being aware of this... the field of machine learning has, broadly speaking, had a pretty big wake up call in the last five years. And I think we're coming to understand that you can't just naively build some huge model and say, you know, here's a pile of examples, do your thing. That we need all sorts of tools in order to make sure that it's behaving the way we expect, and you know, there's an actual scientific progress that's being made there. But it's also kind of as much a social question, right, of who are the stakeholders? Who has a seat at the table? Who has insight into how the model is working? And is that actually kind of disclosed to the people being affected by it?

Brian 46:17
So it's a it's a huge, It's a huge question. But in some ways, I think this is, you know, without being too hyperbolic, I think this is essentially the defining question facing us for the next 10 years. Because this technology is coming in almost every single application that you could think of. And so there really is this question of, can we can our sort of wisdom and our judgment, keep pace with our enthusiastic adoption of these technologies?

Jacob 46:52
Well, before I asked you, where people can go to learn more about you and grab the book, any last parting words of wisdom for people who are watching or listening or any advice that that you have for people, maybe when it comes to becoming a little bit more aware of what's going on in this realm?

Brian 47:09
Yeah, I mean, I think for people who are interested in this area, in terms of making a career change, or if people are, you know, at the beginning of their careers, thinking about what to study thinking about what to get involved in. I see this, personally, as not only one of the most high impact things that you can work on. It's also one of the fastest growing areas in computer science, and really, in all of science. So part of what I just want to impart to people is like, this is the time, you know, if this if these sets of issues interest you, there's never been a better time to get involved. You know, universities are spinning up, courses, that touch on this area, you know, machine learning ethics, AI safety, you're

seeing a lot of resources coming along online, and organizations are hiring in this area in a hurry, because I think a lot of organizations realize that this is critical, whether you're, you know, you could be a credit card company, and you need to think about fairness and transparency and credit scoring, you could be a health insurance company thinking about how do you model patient needs?

Brian 48:22
There's truly no industry that I can even think of that's not part of this question. And so there's a huge appetite for people who can bring some expertise and insight to that. So if that sounds like something that's interesting to people, all I can say is, you know, get on it, you know, like, we need as much manpower as we can. I think this is a critical time.

Jacob 48:47
Very cool. Well, Brian, where can people go to learn more about you and grab your book, anything that you want to mention for people to check out?

Brian 48:55
Sure, yeah, that the book is called the alignment problem. And it's just recently out, in hardcover, it's on Audible, etc, etc. And if you want to know more about my work in general, I'm on Twitter at Brian Christian, and on the web at Brianchristian.org.

Jacob 49:13
Very cool. Brian, thank you so much for taking time out of your day to speak with me today.

Brian 49:17
It's been my pleasure. Thank you.

Jacob 49:19
And thanks, everyone for tuning in. And my guest, again, Brian, Christian, make sure to check out his book. It's called the alignment problem, and I promise it will be worth your time. See you next week. All right, we are all done. Let me push stop recording here.