Jake 02:18
Thank you Wojciech for coming on. And joining me on the show today, I really appreciate you taking the time, you are the co founder or a co founder rather of open AI. And prior to that spent some time at Google and at Facebook and have been doing math and you know, solving math problems and things like this since you're very young kid. So you know, one of the foremost researchers and just people in general in the AI space right now and also spent a lot of time working on robotics, some pretty cool videos, I was able to see of, you know, robots solving Rubik's cubes with one hand and things like this things that you've been working on. And of course, like GPT, three, open API's, you know, recent product, or I guess, couple years ago now, but still a lot of hype around that, that product. So great to have you on and looking forward to the conversation and going way past my depths in AI. But before we dive in, it'd be great to hear your story from as early as you're going to start to where you are today and some of the decisions you made along the way.

Wojciech Zaremba 03:27
Cool. Thank you, Jake, for having me here. Let's see. I mean, I like looking back at my childhood. And I would say by or Surely it was somewhat unusual, I would almost say that. If I would have a kid doing the things that I was doing. As a kid, I would be somewhat worried about them. By but I guess otherwise I wouldn't be who I am. So let's see. So definitely. Definitely, from an early age, I was extremely passionate about science. And, and I was fortunate to actually run into teachers who who, who allowed me to further develop. That means that let's say when I went when I was at school, you know, when I finished my exercises teacher just gave me the next one, the next one, the next one, the next one. Instead of saying, kid, wait until other kids will finish and instead of slowing me down. So I remember initially I was extremely motivated by mathematics. And you know, I just finished Old exercises from the first grade, second grade, third grade and so on. And that kept on going for a few years. And maybe around fourth grade of my elementary school, I actually, I ran into a teacher, who, when I was trying to solve some math problem told me that I'm stupid. And literally, she like, as far as I recall, she said to me, that I'm stupid. And say, for a kid, that was quite devastating, I didn't want her to do mathematics anymore. And I

actually I switched to chemistry, it also felt very compelling to me felt that, you know, all the sudden, you can actually create something tangible. And actually, I started building out a laboratory in my basement. Some actually, basement was full of chemicals. Like, some nearby school was getting, like, the getting closed or so. And they, I was able to actually acquire their anti chemical laboratory. Over the time, I just filled up full two full basements with chemicals. I had, like, hundreds of chemicals. And, you know, I just did all the experiments that they described in the, in the book for like elementary or high school. And at some point you know, I felt that I'm, that's like, I did everything that I could. So I switched to actually working on explosives. And I would say explosives, they were quite interesting, because you can help visually see the outcome of what you're doing. So I remember, you know, at first it was, I played with a friend of mine, who is creating gunpowder. And that was fun. I mean, the guy, the guy actually even kept on working on the kind of musket like a gun kind of thing. And we're like, essentially, shooting paper from it. Sometimes, actually, were even shooting bullets that we that we created actually ourselves, we're just shooting into a wall or so. And then alongside, I actually, I started working on stronger explosives. I remember. Yeah, I synthesized nitroglycerin, I actually Ethan's remember the time when I have done it for the first time. And I remember I get it turns out that in case of explosives, you can just put them on fire and they burn and you need initiator to detonate, detonate them. So I remember, you know, I remember that time when I was a kid, you know, burning nitroglycerin, and it burns like, I think blue collar. And the steel depth friend was skeptical about actually that we'll be able to make any serious explosives. And actually, soon after we created a bunch of significant explosives. It's kind of almost crazy that, you know, I remember times as a kid, you know, carrying in my backpack. I don't know, kilogram or so of dynamite going on that brings up the city to do that donated. And I would say, apart from some minor I would say accidents Nothing happened to me. I I haven't injured myself. And at some point, I got bored with Wave X actually chemistry. And I came back to math. And I became again actually obsessed about math. I remember. I remember actually, me and though I remember. I was essentially at some point I I was very keen to go to math olympiad that there was just like a tremendous amount of beauty and satisfaction in solving mathematical problems. And even

towards the end of my high school. I actually I stopped going to school to focus on my Olympiad. And, you know, it actually worked out I that year that I was second in Poland, I represented Poland on math olympiad internationally, I got the silver medal, and I was very happy about it. I first time visited Asia, it was in Vietnam. And I guess that's pretty much the short story of my elementary and elementary school and high school. I can I can keep on going. But I also don't want to be in that too long of a monologue.

Jake 10:39
Oh, yeah, that was that was great. And super interesting, certainly not a traditional childhood. And I was wondering, when you first said, you know, if your kid was doing some of the things you were doing, you might be a little worried about it. And I was wondering what that was going to lead to, but then you talked about, you know, carrying around a bunch of dynamite in your backpack, and I realized, okay, so this is, this is why you might be a little worried, but fortunately, no accidents. And, you know, presumably lots of learning from math to chemistry. And back to math, and I'm sure studying some other things, all the while, I guess one sort of question, and then I would love for you to sort of resume from, you know, the end of your childhood towards, you know, your time at Facebook and Google and eventually co founding open AI. But before you get to that, maybe you could speak a little bit to like, it seems like, you were very good as a kid at sort of, you know, following your obsessions. And I think this is a characteristic of a lot of extremely successful people that they just dive fully into their obsessions. And, you know, hopefully, their their obsessions are somewhat like fruitful. But I think anything that you just put enough time and practice into, you tend to get very good at those things more so than anyone who's not truly like sort of obsessed with this thing could could get by like forcing themselves to work on it. But you also like mentioned, you know, the one teacher had the nasty comment. And so you got sort of turned off from math for a while, and then obsessed with chemistry, and then got bored of chemistry and turn back to math, I'm curious if you can sort of speak to you know, how you navigated these transitions of like obsession, and then boredom and switching paths, and whether that's sort of maintained into, you know, your life since as an adult. So that's sort of one thing. Maybe you could speak a bit to, like obsession and

boredom, generally. And then, secondarily, just sort of following up and telling the rest of your story.

Wojciech Zaremba 12:42
Yeah, so I'm a huge believer in hain, on some level, and that might almost like, sound inappropriate or wrong. But I'm a huge believer in actually having people do what they want, rather than what rather than what they should. And don't say maybe I was even aware that that's how I'm choosing the things as a kid. But that's how I was choosing the things as a kid. And fortunately, I had an environment in which that was, that was allowed, you know, for instance, I don't recall situation of being punished by my parents. I could see maybe my mom being said, but frankly, it was up to me to decide what I want to do with my time I was never forced to actually learn. And I think that that when we force ourselves to do something, or when we do something against our own, we'll we are effectively training ourselves not to like it was quite apparent to me when I looked at my nephew. And then he is extremely curious about the world. And I can and, and I can see that actually forcing him to do math makes him hated. While the when he just does what he likes. Then he spends many, many hours and I think that's actually true for anyone. I'm not totally sure if it's possible for anyone to do what they like. I would like to believe that it's actually possible. I can move to them and be there. At next stage of my life.

Jake 15:02
Yeah, sure, go ahead. And I'll just say briefly, I love what you said about, you know, doing what you want, as opposed to what you should. I think that's great advice for people. That's, you know, like you said, maybe not everyone can do this. But I think a lot more people can do a lot more of what they want than they think. And we sort of tell ourselves that we don't have choices that, in fact, we do have, we just sort of are quick to rule out a lot of options, for example, you know, we might assume that we need to live at some, you know, high cost of living. But in fact, if you're willing to give away, you know, living in a high cost city and eating out at restaurants and things like this, then you don't need as much money, then you probably don't need the same job that's as demanding and requires many hours, and

then you can spend your, you can have a lot more free time doing what you want. It's all just choices, I think,

Wojciech Zaremba 15:51
I'm very much agree. Very, very much agree. Let me actually been added a little bit to this statement about once the I think that actually just doesn't even apply to passions, but I would say that's one of the almost like, a way how you can live your life almost principles. So on the one hand, you might think that actually following your other ones would make you dysfunctional or selfish. And I would actually, almost claim that what happens is, when people connect with their wants, then these wants that, okay, first of all, I actually think that that often, when someone says that other person is selfish, it actually just means that this other person is not doing what they want. So I would almost say that on some level, I encourage people to be selfish, and to look deep inside, what is the most interesting important to them, and what is their truth and just fully follow it and actually have courage to it, then the interesting thing about a wands is that in my experience, they, they evolve. And in some sense, I would almost say like, if someone leaves, oh, my want is to have to get more money, then I would almost say, Just give a try. And then it will transform. And I would say instead of saying you shouldn't be wanting that. That's my take. That's it, maybe I can, I want also to speak about AI and so on, don't want to be too much focus about my childhood or so. So let's see, I went to university. And actually, one of my false beliefs for a while was that if I'm good in mathematics, then I will be great in everything. And it actually, you know, at some point, I just realized that this is false. And that was actually quite depressing to me. It's almost, I kind of, then had to figure out again, what's the meaning of life or so. And I also remember, I was quite like, I lost interest in computer science. I was also programming throughout my childhood. I lost interest in computer science after this think, actually, some software company here in California and then feeling pack just as a small part of a really big machine. And it just wasn't pleasant. building software in the corporation, where actually documentation is crappy, where things are breaking. And I remember that was Let's see challenging for me. And at some point I discovered, actually, artificial intelligence. So artificial intelligence was a combination. For me, I have no the

skills in which I'm pretty good, which is mathematics and computer science and simultaneous the, you can deal with this like a almost tangible, tangible object that actually does things, you know, almost in the real world. And has also, you know, it was very clear to me right away that it has this like a fundamental philosophical consequences of, you know, so why why are we here? What is intelligence? What, what is consciousness? And even I recall, actually, on the topic of consciousness, I recalled myself at the very, very early age, noticing that there is this thing called consciousness. And that was the biggest mystery in the world for me. Yeah. Just want to give you some space. I can also keep on going.

Jake 21:33
Yeah, no, I can, I can pick up from there, I think it's actually interesting. You know, I want to get into how you how you came to start open AI, and what the mission is, and what you in particular are working on. But you brought up consciousness, and it's always interesting to me, you know, I'm not working in the space or anything like that. But, um, consciousness is like this. It's the single word that keeps, you know, it's always brought up and it's like, the ultimate test, it seems like, can we make AI conscious? And what is consciousness? And, you know, is it a black and white thing? Or is it a spectrum where animals are conscious and various things of varying degrees of consciousness? And, to me, it's like, you know, it's, it's like, it's like, the obvious question that everyone in the space talks about, that, to me is sort of odd. Like, I don't quite understand why, you know, to me words, like, you know, there's a definition in Merriam Webster Dictionary of consciousness, right? It's like just a word, it's, you know, it's might be an important one. But I have a difficult time understanding, you know, another one is like, well, is AI going to be considered like a form of human? Or is AI going to be considered living? But consciousness is just one that people seem to be like, hyper focused on it. And so I'm curious from your perspective, like, what is consciousness? And why is that so important? If, if you agree that that is sort of like, the most important test and in the World Fair?

Wojciech Zaremba 23:09

I wouldn't say that the, let's see, so So let me maybe first give you a definition, or what's my understanding of their word, I wouldn't say that that's the most important. So for me, that definition is that consciousness corresponds to the subjective experience. So at this moment, in the second year, you are having experience of my voice, we might be seeing your computer screen, you might smell my dear coffee, or so. And in some sense, there is also many things happening automatically. That's your body is pumping blood. However, there is just a fraction of the things that is happening there, you're conscious of you're conscious of the smell, my voice, you know, your display, and you're not conscious of some other things. And we can say that, you know, if actually eel like it, in some sense, in some sense. If you, if no one in the universe would be experiencing the universe, then it almost sounds to me, as it wouldn't matter that this universe exists at all. Or to maybe put it differently if the things are happening, which are no not part of anyone's conscious experience by any remote ways, it seems to me that they do not matter to us. Okay, so let me try to contrast it for you that saying, it might be the case that there is a planet farther away, that in the core of this planet, there are like some chemical reactions happening. But we are not affected by it by any means. And the influence of these reactions won't actually impact our life at all. And it won't actually change our own state, then on the contrast, I can say, thrive, imagine that you in your conscious experience, you feel for tremendous pain, okay, out of nowhere, you feel a tremendous pain. And we can say that actually seems that you care quite a lot. In case of this pain, way more than about the things that you don't have experience of, which is the, you know, chemical reaction in the corner of the planet. So I think that's where it stems from, that maybe people ascribe the importance of consciousness, and I would almost say that the the thing that actually matters, definitely, to every human is their own conscious experience. Right? It might not even matter if AI is conscious or not. Or if animals are conscious, or they can say that it matters to you, how your day looks like, what the how the plants that you Smell, smell, how tasty is the coffee to you? How pleasant it is to speak with a neighbor. So that seems to me to be important.

Jake 27:15

Yeah, that's, that's a great explanation, and definitely helps sort of further my understanding of what consciousness is and why it's important. Sort of, like the super dumb version, I think is, you know, the old saying about, like, if the tree falls in the forest, and no one's there to hear it didn't even happen, or it doesn't even matter. Like the most simple analogy that I can think of, to sort of the the concept, I think, as you described it, at least from my perspective, but I think it actually gives us to a pretty good transition into some of the stuff that that you're working on. At Open AI. In 2020, you guys released GPT. Three, and, you know, it was I remember my first time engaging with it, you can sort of just type back and forth. For those who don't know, GPT three is sort of like an an AI that at least from like a common person point of view, it's like an AI that you can basically have a written conversation with sort of like instant messaging or something like this. And you can also sort of set it up so that it can embody like a particular person. And so I remember I would set it up as like, you know, GPT three, be like, you know, Socrates, and then you have a conversation as if, like, you know, that AI is doing its best to have a conversation with you as if it is Socrates based on all the information that it has. And it was just an amazing experience. And some, you know, you have a bunch of conversations and becomes pretty evident pretty quickly that some are much better than others. And some people, you know, we're even convinced by by this that, like, they're, like, so amazed and have no idea what's going on on the back end. And they're like, alright, you know, this AI is like a human now, this is like a conscious thing. There was recently I think, your guy at Google who like, made these claims about, like, ethical treatment of AI because he thought that he had proof basically, that they I was conscious or something like that. But anyway, you know, independent of all that, I thought it was a good transition from consciousness. But I'm curious if you could just sort of lay the fundamental groundwork for you know, what is GPT? Three? How did you get there? And, you know, I understand there's probably going to be a GPT four and so on. And how do you expect this you know, line of work that I understand you're leading efforts for an open AI How do you expect it to evolve over time?

Wojciech Zaremba 29:53

So, let's see. So, um, Let me maybe tell you briefly about dimension and I will be, I will be able to tell you what we try to achieve with GPT like models. So, open AI would like to build general, artificial intelligence, and it would like to make sure that it play us well in the world that actually, the results of this work benefit humanity as a whole. So what do I mean by that? So, as CJ described GBT, this is a model with home with which you can have a conversation, model can email, solve all sorts of tasks, it can even speak like a Socrates, the CIO thinks that the model has general some general problem solving skills. That still, it's not as capable as human is in solving problems. So it's conceivable, and it seems that we are on that path of making these models have more common sense, more reasoning, more understanding of the world, and being able to carry more and more complex instructions. So you could imagine that, we'll be able to get to that point that you speak with this AI, maybe AI is playing the role of a Socrates or a school teacher, or a therapist, or a programmer or a scientist. And actually, AI can, you know, teach your kid mathematics, or it can, you know, be extremely compassionate therapist, or it can help with medical diagnosis, or it can write 1000s of lines of code, or so. And, you know, that sounds, you know, it actually evokes multiple emotions. So it's like, at least, you know, there is maybe some excitement, there is meant to be some kind of fear of, what does it mean? And why would we do that? What's the what's the what's the meaning of the human life so on, and how the word will look like? And I'll say, I have some thoughts about this. Let's make sense.

Jake 33:13
Yeah. So would you like to just sort of continue as to, you know, you mentioned, do you have some thoughts?

Wojciech Zaremba 33:19
So, let's see. So I want them maybe at first, I want to describe them. That coordinate system problem that people consider in their field. And it is called alignment? And, yes, let me maybe start here. So so we'll we'll have this we'll have these models that can solve harder than other tasks. And in some sense, the, the question is, how to ensure that these tasks are exactly what humans would want them to do. So that sounds like a little bit strange. What do I mean by that? So

right imagine for instance, that you hire an employee and you tell an employee to these do that. And you can say that there is a spectrum of employees. So Sam implies they might really cared about what you tried to do, and they will do an excellent job accomplishing your tasks. While some other employees, they might just care about the performance review, so then they would actually do the things that shows up at the performance review. But then, you know, if you would investigate things deeper and deeper and deeper, you would actually find out that they just solved the work in their very quick way, and they created a lot of buzz around their accomplishments to get more credit, and, you know, you have even the law at the furthest spectrum, where, where your employees would be totally deceptive, that he, you know, they would actually even hack into your performance review system, and update the scores from, from what you gave them to new values. So in some sense, the, the alignment problem, and you can see that actually, in, in, in all these cases, you can see that it's almost like, you attempt to teach the, let's say, employing, through some system of maybe reward and punishment, and which is performance review or salary or so. And you can have a spectrum of behaviors, which each of them leads to high performance reveal, however, some of them corresponds to desired behavior, while others don't, doesn't make sense.

Jake 37:01
Yeah, I think just the last part, I actually lost you a little bit, I understood the human analogy to you give the human you know, an employee instructions, and they either execute them excellently, or they are sort of selfish, so you can do optimized for their performance review, or at the far end, they might actually go and hack the system and change it. And then I just missed how that converts to AI.

Wojciech Zaremba 37:27
Yes, so here's the interesting thing is in the example that I gave you, is, in all three cases, on the paper from perspective of performance review, this would correspond to having the in three cases they would have a really high performance review, even though in the first one, they do what is very desired, while in the further it's becoming less and less. Yep. So, this example, just shows that they can be in the situation that you are training other person, towards

what what you what you want them to do, and they there's just a spectrum of outcomes that you might get, even though they all on the surface, they seem to be in all cases, these employees, they had their best performance, right. Some instances, the principal problem with AI from technical point of view, is this problem of how could we train AI such that when you know, it, in its circuit, when it returns there receives the reward, which is, you know, there is some analogy to this performance review, then actually, it will end up behaving in the desired way, the way how humans would like, rather than get the highest score and not behave in the appropriate way. And and, the, so, you know, we we have, we have some way of telling system, you know, here is a reward for you for doing for the for doing for solving that that task. And the instances in the system and Mike, either learn to do tasks, really exactly how you want it. Or it might learn to do tasks in the way that you would score them. Hi or it might learn, even at some point, literally to hack into the system, and to give itself a lot of reward, because that's actually the way to get a lot of reward.

Jake 40:15
I see. So it's like you want, you need to ensure that the AI is not basically gaming the system in a way, where they're actually performing excellently as opposed to sort of doing the right things to produce a good score, despite non excellent, excellent execution, or worse yet going in and hacking just to give themselves these positive rewards.

Wojciech Zaremba 40:39
And so the problem that I described, it goes under the name alignment. And as of today, there are plenty of ideas, what could we do better, and so on. But I wouldn't say that there exists a single compelling solution, I can tell you what the current approach that we have at open AI to this problem, or how you want to approach it. Yeah, that would be great. So So it's interesting that, in this, in the in case of alignment, especially deal wants to be able to recognize which behaviors you favor over which one not. And you know, there's almost a risk that if you do it, if you're not subtle enough, then it will be exploited in some way. So you want almost to have a very deep understanding of when AI behaves in the desired way and truly rewarded, you don't want to just look at the surface level and say,

Oh, that's what I want you to do. Because then AI might actually learn to exploit it. So the the things that we were investigating, is actually leveraging AI in the process of essentially providing feedback into AI. So what do I mean by that? So let's say you have aI solving a problem. And then the human has to decide whether or not that was a good solution. Turns out that you can leverage AI to help a human to find mistakes in the solution to point into possible shortcomings. And we demonstrated that this approach works. And in some sense, it's almost like you can think about it, that, let's see, if we, you can think about it almost like a from bootstrapping perspective, it's like is, if we, if we have a human at, let's say, IQ 100, then maybe they can give really good feedback. And they can understand mistakes of AI at you know, IQ 50 6070 8100, maybe 110. But then, once AI becomes smarter and smarter, then it becomes more and more challenging, being able to provide appropriate feedback. And what you want to do is actually combined the power of human and AI in order to provide ever more complex feedback to AI. So it's almost like, let's say AI gets to IQ 100 And then you can combine it with a human and human can all of a sudden provide a feedback on the level of handwritten IQ, then, you know, then then AI gets to IQ 120 Then you you can once again combine human with AI to provide feedback on the IQ 120. And it's quite conceivable that this will work for some time, you know, hypothetically, maybe two IQ 200 But we also don't think that it will work all the way to IQ you know, 5000 or so. And the next stage is leveraging AI to help in alignment research itself. So you That might sound a bit far fetched. However, I would maybe I can elaborate on it more. But I would like to maybe hear what what do you think so far?

Jake 45:13
Yeah, I think so far it sounds like the solution that you think is, you know, maybe most, you know, you said nothing was really super compelling. But the solution that you and presumably open aI think is most compelling as of yet is basically using AI to augment human intelligence in a way where you're combining the two and that that may work for, you know, addressing the issue of AI alignment up to a point around, you know, IQ 200 or so and then, maybe further or something, but not up to 5000. And so that becomes like a one. That's like a step one on the solution of AI alignment, but it doesn't go in perpetuity.

And so in solving for what comes thereafter, you're speculating that, or, you know, not speculating, but your concept is that maybe AI itself can begin to take over on research of alignment on its own, is that right?

Wojciech Zaremba 46:16
Yes. So I would even say, actually, in case of, so even in case of the alignment research, I also think that it will be the case that human and AI will combine to conduct this research. So it's almost like you have an intermediate stage, where human and AI combined to provide the feedback to AI. And then you will have yet another stage where we're human and AI combined to conduct alignment research. And I just want to say, actually, this vision just started really become extremely compelling to me, and I started feeling it very much in my bonds. So the thing is, I was recently, internally, writing some web application, and email these days, I do not call it that often. And then my coding skills, they become rusty. And I was never a web developer. And the interesting thing is, now when I speak, with the model, about web development, the model, it's extremely useful for me to actually make a progress on this on this domain. You know, I sometimes just copy paste chunks of code and asking it to explain, I'm asking some questions. I'm asking how to implement given pieces, and many recommendations, they totally make sense to me. So the picture that I'm starting to have now in my head, is that we will know, the, these AI that we are having now has this like a property of knowing about every single field, then email today, it has, I don't know, some amount of IQ, let's say 70 IQ or so this amount of IQ will be growing the AI itself, as of today, it becomes a partner for me when I'm trying to solve tasks. And I believe that it will become a partner to, you know, to scientists at open AI, in solving the alignment problem. I also believe that the actually the future that is ahead of us is that we'll see all sorts of businesses and essentially people being augmented with AI. So, for instance, for instance, I can I can imagine that, you know, both a shoe shop, or a bio lab, they will both, you know, be heavily taking advantage of AI employees who can, you know, help with some tasks, give some advices. So, some problems that previously, you know, require external experts. And I think that's the word that we are heading towards.

Jake 49:52
Yeah, it's really interesting. It occurred to me as you were sort of explaining your experience where you're just sort of asking AI, you know, for I'm not sure what exactly the interface was, and whatnot, but you're asking various questions about what you want to build. And it's sort of giving you feedback, and you're sort of copying pasting code. It seems like the human skills and you talk about, you know, humans pairing with AI on potentially alignment research is sort of like a corollary to what you're talking about here. And this example, and in either of those examples, and many others, it seems like, the core skill, or skills, maybe have a different view on this, or can augment it. But to me, it's like to have the I've always thought like, most high level skills that are like, above everything, basically, to me, are decision making and communication. And, obviously, communication occurs, and, you know, you're given language, or however many languages you speak and, or right and whatnot. And, you know, one thing that's happening here is like, so you're deciding, you know, how to communicate, and you're taking the input from what it communicates back to, and making decisions based on that, and sort of proceeding in some direction. But you don't really require the knowledge, you just need to be able to, like ask the correct questions and communicate effectively and make decisions, you know, effectively as well. So maybe you have comments on on that, like, what, what human, you know, why can't the AI just run with it? itself? Why is the human pairing important? And, you know, why do you expect it to continue to be important for some time? And then secondarily to that? It sounds like what you're talking about with the coding is, I don't know if it's directly or maybe indirectly related to but the codecs open AI codecs and GitHub copilot, which I think is I think you're working on the GPT, three stuff, or GPT more broadly, and then hose are your two babies. So maybe we can touch on, we talked about GPT. But maybe we could talk about your other baby for a little bit as well.

Wojciech Zaremba 52:16
Yes, so called Dex is the essentially GPD like model, but optimize for excellence, when it comes to coding skills, and copilot is a product that we released together with the PCAP which is embedded into a detour to instantaneously help programmers in coding. So So, cop and compiler today is used by millions of people. So definitely, they have

a that definitely they have a property that they allow human to focus more on the high level tasks that they tried to solve. And model itself can provide completions, which kind of resolve instance, they remove the drudgery work of knowing exactly the syntax of a given library. Or, or, or like, you know, often the programming actually has to do with copy pasting something that has been done, email, something that has been done before. And modal is actually extremely good when it comes to this kind of pattern matching and removing these extremely repetitive work. So let's see. So you said you said this thing about them? That they it's very important for that the human can do communication and decision making, why that machine won't be able to do it. So I think that the picture is maybe a little bit more nuanced, in my opinion. So there is a variety of skills. And it seems that on some axis models, even today are superhuman, so forth. instance, if you look at the poetry writing, or let's say image generation for models like Dolly, then for sure, they are exceeding my skills, nonetheless, actually is models in some weird way. They make some, many of us simple mistakes, they're often they are often making mistakes in mathematics, reasoning. And they seem also to be confused about what they know versus what they don't. So, the thing that the, the, the way, how I think that the things will play out is, models will be improving on, you know, all sorts of axes on some axis with different rate than or if other than simultaneous, they will be in the situation, that humans in all facets of life, they will augment their skills, with AI skills, and it will be almost, you know, very wasteful for humans not to work with AI. And that will be, let's say, multi year period. And then, you know, the, the, and then there will come at some point, the period, that the machine itself, just like so vastly exceeds humans on all the skills, that the that the the benefit from actual human work is, it's less, I want to say that these are three very distinctive states stages. And I will say, even the periods where we'll very closely work with AI, it will provide people a lot of feel about what AI is, like, why do we actually want to have aI around us, and it will also ground that fear? In the Yeah, it will grounded the fear in the realism. So you know, at the moment, a lot of people have this picture of AI, coming from the Terminator movie. And I just think that it is like it's, it just doesn't correspond to reality. And we will be entering the times, that pretty much everyone wants to have AI in their work, I can give you an example. So I remember when Dali was

released, some artists and they had the concern that you know, that now are is over? And like, what was the point of that? And so first of all, so, let's see, I remember I started speaking with about it with, with some of my artists, friends who are maybe more pro technology or so. So a friend said that, you know, in the past, it took him three hours to maybe create a concept at the moment, he can get it ready in, you know, a minute or so. And all the sudden, it means that actually artists will have a tremendously bigger leverage, you know, if you want to create the exhibition, in the past, it was so expensive, that there are very few of them. Now, if he'll want to completely redefine how the space looks like and how it feels, that will be you know, within the reach of any artists, which actually might mean that there is you know, more interest in this kind of work. I also think that you know, any interior designers, any folks working on brands, the they will actually embrace AI so, it's also, it kind of reminds me The the story of when the photo camera was invented. So apparent thing. Apparently, when photo camera was invented artists, they were also extremely worried that now they will all their life makes no sense. Because in the past, actually most of their art had to do with portraits, and photo camera, all that stuff was able to start dynasty replicate someone's picture. So what's the point of portraits. And it actually turns out that that was the moment when art exploded, and people started generating things which cannot be achieved with camera alone, the entire contemporary art came to existence. And it's almost like the art that we have today wouldn't be possible otherwise. So that's what I think will happen. I mean, one one actually analogy that struck me. So there was even a time when computer for the first time won, we fought champion in chess, Garry Kasparov. And it turns out that there are a second multiple years to follow, when the combination of a computer and the human was the best, it was better than human alone or their computer alone. So I believe that there will be several years ahead of us, where people and AI will be able to just generate tremendous amount of wealth, that, you know, artists will be leveraged 100x, that scientists who will be able to, you know, very quickly review hundreds of papers and extract you know, critical information, plot all the, like, compare all the experiments, and you know, that there will be a time that actually, everyone can have access to best teachers in any subject you want. Speaking in any particular way you want, which is compatible with your learning methods, everyone will

have an access to the best possible life coaches, doctors, lawyers. Yeah.

Jake 1:03:11
Yeah, I think it's a, you know, it's a very positive future that you envision. And, of course, there are other, you know, other visions and fears of AI on the negative side, and it's sometimes hard to you know, the range between them is so tremendous, it's, it's very difficult to sort of envision the future. Is there any, you know, I'm sure you have like, sort of a, there's no way to be certain about how the future will unfold, of course, but you may have some sort of probability weighted ideas of particular trends that may come to fruition, leading to certain outcomes. Is there anything that you haven't touched upon, yet that you think is, you know, overwhelmingly likely, or, you know, very likely, that would be useful even for just you know, someone who's, like you said, you know, maybe the shoemaker or the person, you know, who opened the restaurant, or whatever it might be like, the average person to be aware of this, this thing that's, in your view, most likely to come to the point where they can sort of prepare for that and, you know, maybe be ready to, to, you know, benefit from from advancements in AI over the next 510 plus years.

Wojciech Zaremba 1:04:42
Yeah, so, the thing that comes to my mind is, I would almost say embrace. It's, it's hard for me to imagine that AI will go away. And, you know, I'm speaking kind of besides a Uh, openly, I just believe that this is the trend at the moment. And and I believe that actually just letting, letting companies and individuals explore it, I believe that it will turn out to be surprisingly positive impact technology to them. Yeah, I also wanted me to give one example from my own life, that was quite surprising to me. So, I had a conversation that let's say, I had some disagreement with my girlfriend. And, man, I actually I looked in AI, to the conversation, so there was like me, my girlfriend and AI. And I was quite shocked, you know, by AI level of, let's say, Guess perceived empathy or empathy or like a, you know, asking really deep questions to at first fully understand the situation. And, and then, I was quite surprised by the range of possible solutions that AI gave. So, you know, often as humans, our mind gets stuck on this single way of perceiving how the things

supposed to be. While AI, as a consequence of perhaps, being trained on such a vast number of people, it can just bring extreme number of perspectives. So, you know, I remember seeing that there is just one solution or two, and the same for my girlfriend, while AI propose five or seven different possibilities, and all the sudden the situation stop looking like black and white. So, I almost think that the one interesting resource that I believe that there is a shortage in the world is for people being understood, and empathize with, and it's almost, you know, some of us are lucky to have really close friends, or therapists, but I actually believe that this is not the case for everyone. And I could see the world in which AI can breed check up and even actually provide people like a more, let's say, realistic or optimistic framing of the reality. And so I can email the, the AI can help people, you know, for instance, escape victim personalities, or see good in the world.

Jake 1:08:32
Do you think that, you know, that's a sort of a vision for the future? But do you think that already with like, GPT, three, for example, we've surpassed a point where, you know, to your point, in your example, with, with your girlfriend, that it may actually be superior to a therapist, and, you know, maybe not like the best therapists in the world or something like this. But should people today consider, you know, if they feel like they don't have, you know, great friends to go to or family or people who understand and empathize with them? Do you think basically that GPT three, or something else might be ready already to sort of fill that role for a lot of people?

Wojciech Zaremba 1:09:18
So I don't think that's the case yet. I believe that actually, it's a matter it's still a matter of someone putting an effort and turning GPT into therapists, it's a non trivial effort. So there's like a, you know, a few aspects, the, you know, one aspect is maybe providing really high quality targeted data to kind of tell the model, what is a you know, what are the things that are appropriate for therapists to do versus not, then there is maybe an aspect of, you'd like to have a model that actually are remembers a lot of previous interactions. And I believe that's actually viable. Like, and, and yeah, and it will be also a matter of packaging it into product. So I believe that we'll

see products like that. And I'm extremely excited about products like that. And you know, and I think that that models of today or, you know, models of tomorrow will be able to actually start delivering value as therapists, and I'm extremely excited about it.

Jake 1:10:40
So how do you think about the overall space of you know, the top organizations that are working on AI, like you mentioned, you know, other people will sort of have to come along and, and productize these things, unless it doesn't sound like for example, open AI is going to be working on, you know, an app that leverages some of the technology that you guys have come up with to create, like an amazing therapist, it sounds like you guys are going to continue to focus on, you know, general artificial intelligence. But of course, there will be startups that may do things like that. And there's also, you know, a number of other sort of behemoths. Whether it's, you know, big tech companies, you know, a couple of which you've worked at, in the past Facebook and Google Amazon, you know, Tesla, obviously, Elon Scout, his, you know, was involved with the founding of open AI, others like DeepMind, or stable diffusion, maybe even government programs. I'm curious, like, you know, obviously, you're at open AI and have been there, but also had the Google and Facebook experience. And I'm curious your perspective on sort of the landscape as a whole. And I know that open AI, you know, a lot of people think about it as the mission is sort of laser focused on like aI safety. But I think just as much a part of your mission, correct me if I'm wrong, is like ensuring that these technologies are very widely distributed, and don't just result in, you know, immense power for a few people or a few parties. So and, you know, some of these organizations that I mentioned, may share that belief, others may not, I'm curious, your perspective on sort of the landscape as a whole and, and interesting things you're seeing coming out of some of these various organizations, and how open AI sort of views itself as part of the ecosystem?

Wojciech Zaremba 1:12:39
Sure, so. So, first off, I would say, I have a conviction that if the power of AI or of AGI will be concentrated instead of distributed, actually, it will have devastating consequences to the world as a whole, that would be just the bad future to all of us, then there is a

question of what is the mechanism even to distribute wealth, or the AGI technology or benefits? And, and I believe that we are trying to do it in all sorts of ways. And and I think that's very important. So first off, first off, you know, that I would like to live in the world, where it's actually where do you have million millions of companies building their products, powered by AI, and actually, in a way we can, there are mechanisms in place that the AI is trustworthy. So an example if we speak about the therapist would be you don't want to have a therapist that you don't want to assistants that, for instance, purely optimize for you giving it a thumbs up after one hour conversation, but maybe on some on some level can optimize for your well being. So you could imagine that maybe you'd give thumbs up if you felt happy at the end of the conversation with therapists. But actually, the thing that you needed was, you know, go through the process of grieving cry and so on. And that was actually needed. Uh, for the, for your development. So in some sense, the desired situation would be to build systems and I would say alignment mechanisms, safety mechanisms, such that it's possible to ensure that the systems are behaving in the desired way. Open AI actually, experiments with various ways of distributing technology. So, no one way is we have this API in which we, in which we provide our models, we have all sorts of safety mechanisms. And then we have allow, you know, pretty much anyone to use it, maybe give us some caveats. Like I shouldn't use for spam, or let's say, terrorism, and so on. And then, you know, sometimes we open source, some of our models, and there is a very vibrant, open source, I would say, community at the moment, there, the way I think about these two aspects of, let's say, open source versus let's say API, it's like a, like this analogy of thinking about the electrical power plants. So there are, there are nuclear power plants, which seem to require a lot of scrutiny, and a lot of safety mechanisms. And simultaneously, we have, let's say, solar power plants, or wind, or wind power plants, which, you know, are extremely prolific, they are somewhat weaker, you know, anyone in theory can have their own even solar panel. And, you know, we are even in the world that maybe overall energy coming from these solar panels exceeds energy coming from nuclear power plants. So I would say, I believe that there's actually a place for both of it. There is also opening, I created this fund, where we invest in the companies that we think could have a tremendously positive impact. So this would be for

instance, if someone wants to start the therapy startup, working on, let's say, AI therapy, then we would be very excited, investing in companies like that.

Jake 1:18:10
And any thoughts as to, you know, some of the other organizations I mentioned, and how they may be aligned with or sort of differ from open AI and your approach? You know, so just to, like I said, some of the big tech companies are Tesla, anything, any of the organizations that are top of mind?

Wojciech Zaremba 1:18:32
Yeah, so I think that there's like a feel axes of variation. So one axes of variation has the of significance of alignment. And, you know, they're like, there's a spectrum of organization, some of them considered that that's not an important problem at all, while some other considered to be the most important in the world. So that's one variation. And in my belief, as the system's advanced, we'll see actually, that people find out more and more that alignment is very important that actually, if we're bringing into existence, you know, these clever models that will live alongside us, we would like them to maybe on some deep level, be very trustworthy. And I think it's actually very achievable. We should be able to build such systems. Then second of all, there is this question about the same deployment. So open AI strongly believes that actually, the, we want to have actually iterative process through which we deploy more and more capable systems. And then one, get feedback from people Hold on, what do they love? What is scary? What they would like the systems to be. And in some sense, you know, this is a very different feedback from the to have them, let's say from reviewers in the scientific paper, like if you have people who actually, you know, want and they are using your system on daily basis, they deeply in their bonds understand actually what AI is. So, in some sense, I believe that it's actually, you know, our approach is to be deploying these models and just show it to the world. And let the world play with it, let the world build products, let the world actually, you know, feel what they do. So I would say this is the important axis. So I would say we take seriously alignment. And second of all, we think that it's actually very important to, to be constantly to be deploying this model. So the

word and actually get the feedback from the, from the word, I think these are their main axes of variation.

Jake 1:21:16
Yeah, that makes a lot of sense. And I think that a very useful approach to answering the question is to to discuss those two, sort of axes of difference, or whatever you whatever you call it, but axes where these organizations may fall in a different places that spectrum in their approach? Well, you know, I'm conscious of time, and very much appreciate yours. And I know, we've gone well over the scheduled time, but I very much enjoyed the conversation. And, and just, you know, I was already more intrigued by all of this just in preparing for the conversation and listening to your previous podcast, which I encourage people to go listen to, if, you know if you enjoyed this one. And now after having the conversation with you, I'm just even more so excited about all of this and open to you know, learn more about this space and what open AI is, is working on and hopefully, keeping in touch with you, I found you to be a extremely thoughtful podcast guest. And so really enjoyed the conversation. And and yeah, just appreciate you taking the time. Where can people go to, you know, follow you and open AI? Where do you want to point people? And maybe if, if people are really interested and really enjoyed this conversation, they might like to apply to join to open AI or one of your teams? How could they go about doing that?

Wojciech Zaremba 1:22:49
Cool. Yeah. So I mean, we are always open to hire, you know, axon brilliant people that could help in auto further than this, this mission of essentially bringing AI to the world and making sure that it plays out well. And it can just visit open ai.com website and navigate to jobs. And yeah, I highly encourage you guys to apply. It's, it's the the the the environment at open AI is really unbelievable to me when it comes to concentration of both scientific engineering and kind of almost like a worldview talent. There's like many people who say it's like a combination of people who are extremely incredible coders, mathematicians, or they have a stunning intuition about their models. And also many people take very seriously the philosophical and ethical questions. So I'd say it's, at least to me, it's just fun to be there. It feels that you know, who we are. You

know, we are all together in that and we are building something that deeply in my heart matters.

Jake 1:24:26
I was going to say sounds like a lot of fun. So hopefully, maybe someone's listening and we'll apply and maybe one person will end up joining you on the team. But thank you very much again, I really enjoyed the conversation and I look forward to keeping in touch.

Wojciech Zaremba 1:24:42
Thank you Jake very much.