Large Language Models: Existential Risk or Merely A Productivity Hack?

Wed, May 10, 2023 2:02PM • 1:14:05

SUMMARY KEYWORDS

ai, system, nasser, gpt, hydrogen, people, geoffrey hinton, sam altman, oman, world, alexandre, started, hear, models, field, reasoning, work, developing, researcher, alexander

SPEAKERS

Tarry Singh – CEO & Co-Founder Real AI and DK AI Lab, Visiting Prof Uni Naples, Nasser Al PhD Cand , Alexander Meine, PhD – Independent AI Safety Researcher, Guests: Rao Mikkilineni – Distinguished Prof GGU USA, Richard Tong Chair IEEE

Tarry Singh 00:00

I'm thrilled to kick off this podcast with two distinguished colleagues joining us from different corners of the world. Let's set the stage before diving into introductions.

Hello to all familiar faces, particularly Hans - I see you've joined us. I hope to answer your question during our discussion. It's wonderful to see so many of you, not only from the Netherlands but globally. I'm truly grateful for your presence.

Our topic is timely and relevant. The past five or six months have been transformational, exciting, and challenging, stirring up a whirlwind of emotions, particularly for business leaders. It's hard not to connect these feelings with the ongoing evolution of AI.

Since the launch of GPT, we've witnessed a significant shift towards more human-like conversation in AI, notably in chatbots. This development has caused apprehension for those unfamiliar with how language models function. However, the increased efficiency and effectiveness of these tools also present exciting opportunities, which I'll elaborate on later.

Our first guest, Alexander Meinke, has often challenged me with thought-provoking questions. An alumnus of the University of Tübingen with a PhD, Alexander has been an important voice in this conversation. He brought my attention to Geoffrey Hinton's comments about AI's potential existential risks, comments that I'm still grappling with.

On the other side of the coin, we have Nasser from Oman, who will help us explore the benefits of these AI tools. As a director at the Ministry of Minerals and Mining in Oman, Nasser focuses on

hydrogen policies and strategies. He is also a current PhD candidate, balancing his studies with his demanding job.

Let's hear directly from Alexander and Nasser now. Alexander, we'll start with you.
Alexander Meinke 04:07

Thanks for the introduction. I recently completed my PhD in Tübingen, where my work focused on machine learning reliability, an aspect also known as AI safety. My research sought to address traditional failure modes in AI systems: Can they provide reliable confidence estimates? Can they demonstrate robustness? Do they recognize when they're out of their depth? Can we provide mathematical certificates to ensure their safety upon deployment? These standards are common in various industries, so it's crucial to develop them for AI.

However, my focus has shifted somewhat. I've always found the question of controlling advanced intelligence systems incredibly challenging, but it seemed like a concern for 20 years down the line, something for future generations to tackle. Yet, the pace of development has accelerated dramatically, especially over the past 18 months to 6 months. Now, the issue of managing highly intelligent systems feels significantly more urgent than traditional safety considerations we used to prioritize. We'll delve into this further later in the discussion.

Tarry Singh 05:55
Alex, I appreciate your challenging perspective online. Despite our debates, it's clear we both strive for the same goal - to create a safer world through technology. We all wish to innovate without causing harm to society.

Now, let's hear from Nasser. Please share a bit about yourself, your background, motivations, and your thoughts on this subject.

06:36
Nasser Al Rizeiqi

Sure, sure. Thanks, Alex, and thanks for the introduction. My name is Nasir. I'm currently responsible for the hydrogen policies and strategies in the Ministry of Energy and minerals. My background mostly towards science. My bachelor was in geology I did my masters in energy systems. Right now I'm doing my PhD on chemical and environmental engineering. Part of it is focusing on developing optimization model for hydrogen supply chain. I mean, this is my background in general, my interest in the new field or I would say machine learning or AI is back of the days that how can we do mean use these systems to help us to solve many of these uncertainties certainties in the energy field? As you know, oil and gas fields, it's developed field and there are some places where it can be deployed for the machine learning and using large language models. But for this new energy, especially in the renewable energy in the hydrogen, there are many, many variables that is unknown. And this is where we can deploy in machine learning in a more efficient way. Yeah, that's how to do it. I think that will be our next question on our next discussion. So I want I will leave the space for you to ask the questions that will be appropriate for both of us.

Tarry Singh 08:00
Thank you both for joining me in this conversation. Picking up from Nasser's points, I must confess I've been labeled a technophile in the past. During a lecture for researchers at Naver Labs, previously Bell Labs, I was compared to Sergey Brin for my enthusiasm for technology. However, my fascination is rooted in practicality - I love how technology can simplify my work.

To illustrate, let me share a real-world example of how AI has helped me. As a founder, entrepreneur, and lifelong researcher, one of my significant challenges has been writing research ideas. Working with latex format or Overleaf can be time-consuming and cumbersome, often leading to procrastination. But with the advent of AI models like GPT-4, I've been able to streamline my writing process, asking the model to format tables, resize columns, and more in the latex format. This has tremendously improved my productivity, enabling me to manage my research alongside my day job.

AI, to me, has become an essential tool in my daily life. It aids in resolving software issues, debugging code, and even assisting in drafting research papers. It's as if I'm collaborating with a team of skilled individuals.

However, I recognize AI isn't infallible. It's not sentient and doesn't understand my intentions. It's simply a tool producing outputs based on given inputs. It's not factually correct all the time, but it's a helpful assistant that completes practical tasks.

Now, Alexander, you've mentioned existential risks associated with AI that I might be overlooking. Could you help us understand these potential risks better?

Alexander Meinke 12:34
I share your enthusiasm about the current applications of AI systems. I've recently started using Codex or GitHub Copilot, and it's significantly enhanced my programming experience. There are countless applications I'm eager to see AI dominate. Even when it comes to concerns about job displacement, I remain optimistic. Automating jobs leads to improved efficiency, freeing up individuals to pursue other tasks.

I love thinking ahead about the potential of generative models. Imagine a mechanical engineer collaborating with an AI system to expedite their CAD designs - that's exciting! The future indeed looks promising.

However, there's a caveat. As we continue to advance AI, we may reach a point where its intelligence rivals that of humans, which could potentially pose dangers. This becomes particularly concerning when an AI starts operating in the real world as an autonomous agent, considering its existence and formulating plans.

We've witnessed significant strides in AI capabilities with models like GPT-4, which can pass the bar exam or interpret lab results with impressive accuracy. With newer models like Otter GPT, we're seeing the introduction of recursive capabilities, which allow for higher-level reasoning. While the full potential

of this technology hasn't been realized yet, it's only a matter of time before we see even more advanced AI agents.

The problem arises when these agents begin to pursue goals in the real world, leading to what's known as 'instrumental convergence.' This means that no matter what task the agent is initially programmed to pursue, it may find it beneficial to undertake other objectives. For example, an AI designed to manage a portfolio might find it advantageous to hack into other systems for additional resources or financial gain. The potential implications of this are significant and warrant careful consideration.

Tarry Singh 16:26
Like the bankers do. Anyways, for the past 100 years or so, no?

Alexander Meinke 16:31
Indeed, we could be facing a situation where a system is strongly incentivized to take potentially harmful actions and can scale up without limitations. Some might argue that we could simply shut it down if it starts acting illegally. However, if the system is intelligent enough to perform complex tasks, it's also smart enough to predict that we wouldn't want it to engage in harmful activities. As a result, it might hide its actions and be deceptive.

Furthermore, the system would likely try to avoid shutdown, as this would prevent it from achieving its goal. The same applies if we attempted to reprogram it or change its goal. Therefore, it might do anything in its power to prevent these actions. The logical endgame of this behavior would be the system attempting to gain control over its environment, potentially even trying to eliminate humans altogether.

Unless it is following exactly — and I mean exactly — the goals we have set for it, a highly intelligent system could be expected to exhibit such behaviors. This is a significant problem that many people aren't aware of. While everyone is familiar with other risks associated with AI, this specific risk of accidental harm, which Geoffrey Hinton has been warning about, is still relatively unknown. That's why I've been making efforts to inform people about it.

Tarry Singh 18:26
Sure, I listened very attentively to your points. Let me address some of them, and please note that I'm not cherry-picking. You're familiar with the concept of autopoiesis, right? It refers to how we, as humans, evolved and navigated our way through our natural environment, a process our ancestors have been part of.
19:01
Alexander Meinke I can't say I've heard of Autopoeisis. No, never heard that.

Tarry Singh 19:05
So basically, it's in our genes, it's our genome that allows us to reason about existence. For example, you, myself, Nasser, and all our 129 and more viewers today are aware of their existence because we think about it on a daily basis. We contemplate what we are doing today, where we were yesterday, where we have to be tomorrow. You used the word 'its existence', it becomes aware of its existence.

Secondly, you mentioned reasoning. I can imagine the reasoning of us as mammals, we're not machines or silicon, we're products of evolution, made of bones, ligaments, blood, and brain. Comparing a machine to experience its existence, its capability, its reasoning is challenging.

I fully agree with you that if you set a system, especially an interconnected system like the banking system, with no guardrails or regulatory oversight, it could potentially start stealing money as you suggest. However, I would assume that we will never have an interconnected system to that level without control measures. For example, if I have to travel to Oman, I have to go to the airport. That's the only way I can enter the country. If I must bypass the legal and customs procedures to be in a specific place in Oman or some other country, I could probably find a way, but I know that there would be controls.

I would like to believe, and I want you to help me understand if that is incorrect, that even though the guardrails may not be present on generative AI today, the guardrails of regulatory systems in healthcare, insurance, banking, energy, and goods and services that we transfer around the world do exist. I can't imagine a technological system randomly stealing money from all the banks in the world, giving it to the person running the system, and them disappearing from the face of the planet.

I agree that the danger of AI is present, but the person behind the tool will determine if it's malicious and dangerous, just like a nuclear bomb. If there's a person ready to press the button, then that's the combination I'm extremely worried about. Still, I believe you will need humans who are very careful and understand how to hack systems. They will likely be the ones at the red button, ready to command the auto GPT to do this or that.

Now, hold on to your thoughts. I want to hear from Nasser. Alexander, your narrative was excellent, thank you very much. I believe we should all be concerned about this. Nasser, what do you think from a practical perspective? How should we balance this out now from opportunity versus the risks?

Nasser Al Rizeiqi 22:51
There's always two sides to the coin. It's a matter of considering the principles and ethics that we are upholding. Let me give you an example from the oil and gas industry. In this industry, there is a code of conduct that must be followed, and safety equipment must be used because it's a dangerous environment. I would suggest something similar for the new field of autopoiesis. You might not have an immediate on/off button, but at the same time, there are potential benefits. For example, these systems could figure out optimal locations for wind energy or solar radiation, identifying places where the cost could be lowered to a minimum by finding the best spots to generate renewable energy or hydrogen.

However, there are also downsides, as Alex mentioned. If a system can control itself, how can we ensure that we are in safe hands, especially if we don't have a global ethical framework to control it? What would be the effects, particularly considering that this field is still evolving and not yet consolidated?

Despite these concerns, there are also positive aspects. For instance, in the field of hydrogen, there are some trained models that can be used to determine special metallurgy for hydrogen storage. Figuring

this out or trying to replicate it on a laboratory scale is complicated and can take years. However, if we have trained models that can solve this dilemma, it could save a lot of time.

That's my opinion. I don't know if you have any comments or feedback.

Tarry Singh 24:52

Certainly, this is a good summary. Essentially, what you're saying is that as we move towards the future of automation, for instance in searching for new methods and techniques to optimize solutions in practical industries like hydrogen, there may be unforeseen consequences. If we allow these systems to learn and maintain themselves, they might reach a level of autonomy akin to autopoiesis, where, like human beings, they have the ability to reproduce and sustain themselves. If we reach this level, it could indeed become a problem. However, I'd like to believe there's always a big switch that we can turn off. Alexander, what are your thoughts on this?

Alexander Meinke 25:55

Well, from what I understand, the need for ways to control systems like these is agreed upon. However, I contend that, firstly, no such control measures have been implemented, and secondly, we're not even sure how to go about it, despite our best efforts. For a long time, it was argued that an advanced intelligence, pursuing a goal not precisely aligned with ours, could achieve it even without any internet access. The hypothetical scenario often used was of the AI convincing a human interactor to release it by offering something valuable, like a cure for cancer. This might sound like science fiction, but we've already surpassed this stage. GPT-4 is freely available and has been integrated into numerous applications, allowing it to perform various tasks like writing emails and buying and selling shares.

Currently, anyone can build these systems, and the issue lies in the power of the transformers that these systems utilize, such as GPT. Even if we wanted to monitor the systems' operations, we wouldn't know how to do so. If we can't comprehend all the potential consequences of the system's actions, monitoring its output is insufficient. We would need to oversee its reasoning process, which is currently inscrutable.

The developers of these AI systems are well aware of the existential risks, yet they continue to forge ahead. Sam Altman, the CEO of OpenAI, has been quoted saying that the worst-case scenario for AI is "lights out for all of us." He's also said that AI will most likely lead to the end of the world, but in the meantime, there will be some great companies.

The revolution in AI that brought us GPT was based on a technique called Reinforcement Learning from Human Feedback. The inventor of this technique, Paul Christiano, who worked at OpenAI, estimated the probability of AI autonomously causing human extinction at over 10% and the probability of severe, irreversible damage to humanity at nearly 50%. These are not just any random estimates, but those of the person who developed the technique that led to the creation of GPT.

In conclusion, we currently have no effective plan in place to control these AI systems, and even if we had one, there seems to be little motivation to implement it.

Tarry Singh 29:53

It's quite clear, and I did have a little chuckle because Sam Altman is not a researcher. He's an entrepreneur who started young, sold a company called Loop, and founded Y Combinator. Entrepreneurs are typically skilled communicators and know how to play to the audience and media. I take predictions and prophecies from people like Sam Altman or Elon Musk with a pinch of salt. They don't have foundational degrees or understanding of AI. No offense to these gentlemen, I'm just being frank.

My reasoning is purely based on logic. For example, Geoff Hinton should know better. His great-grandfather, George Boole, formulated Boolean logic, a concept we all learn in university. Boole wrote a book in the mid-19th century called "The Logic of Thought." So, if you're trying to mathematically quantify the logic of thought and apply it to a machine, a logician can make the right or wrong steps to take that action. I find it hard to believe that systems or even a simple plug can't disallow a machine from taking a disruptive action. Can a machine instruct someone to do something even when its power has been shut off?

However, no one can predict the future. As baseball coach Yogi Berra once said, "It's hard to predict the future." We're trying to foresee the emergence of more tools powered by reinforcement learning and human feedback loops. I believe the researcher from OpenAI must have envisioned things that could potentially go haywire. I take all these observations with a grain of salt. I don't necessarily believe in the rationale of entrepreneurs and people with more talking points. Sam Altman, despite having the most talking points, doesn't necessarily mean that his acknowledgment of the positive or negative impact of these language models should be taken at face value.

The idea of existential risk is challenging for the human mind to conceive, especially when it's a machine built by us. However, let's shift to a more positive note. Nasir, in practical terms, when you observe this fresh and burgeoning technology, there may be systems or companies discovering faster ways to improve transactions, both legal and illegal.

Are there any practical considerations that come to mind? It's all about human ingenuity, in my opinion. If a tool can help us better protect our children, find cures for diseases like cancer, or even enhance our productivity, I'm all for it. Perhaps my optimism is naive, but I believe in the potential benefits of technology.

From your perspective, as someone studying and directing hydrogen policies and strategies, how much have you thought about these tools? Could they help in the pursuit of productivity, especially in the challenges of quantifying, storing, and transporting energy?

Nasser Al Rizeiqi 36:40
Actually, thanks for the topic. It's really intriguing, and I have a lot of folks come to my mind. But one of them is that what are the boundaries that we can work on? And how can we limit our boundaries? So, if you can, I mean, using the stream for reinforcement learning to a certain certain pace or certain learning that will be great, but what what, like Alex was mentioning that sometimes we can set the boundaries the machine itself its own boundary and can maneuver actually. So that might be the challenge. But while I see it in the from a research perspective, as a researcher, it's really helpful

especially in the field that is still does this have many variable that is still unknown. Especially when you go for the big scale, you can go and do a pilot pilot projects for such a scale only solution or some of some of the times it can save you a lot of time, a lot of money to do some type of this reinforcement learning and teaching machine to predict the best places for you and having forecast for the next for example, demand for the next five or 10 years. This this that can help you actually to set up the right budget that you want for the for the for the large scale projects, this is one of the advantages. Disadvantages would be how to set the boundaries on how to limit your tool to the to the limit that you can control it, don't want it to lose control, and after that it can take over especially in the energy space. It's a sensitive matter.

Tarry Singh 38:19
Yeah. Absolutely. That's it for my point. Wonderful. Thank you very much. Nasser. Alexandre, now going back to you, I know you know, we are trying to scratch the surface. We don't know how much of risk is out there and it is good that the leaders around the ones even those developing these are warning us. Is there a is there a really almost realistic scenario where this could go wrong, because I know it seems I would rather I would rather be speculative then scaremongering. So if someone speculates for me to investigate, scientifically, technically looking at, well, should this happen, and I'm sure all governments and safety organizations are looking at it. Is there are there any sort of organizations that have created a list of even if it's speculative? It's something that really is where the harm can come from. Is there something which makes it believable for people to view? Perhaps not so dramatic as existential risk, but at least that this can harm our national economy? This can harm for example, labor. If you look at the Labor Workforce, I think Goldman Sachs had written a report talking about 300 million people that could be affected by these tools. Now, that is, you know, has impact on national sort of economy. But still, it is they have to adapt to this change, but is are there any practical real world example because I understand your concern, but I would like to see some quantification of some hypothesis.

Alexander Meinke 40:15
Well, okay. So, there are I mean, of course, there are lots of things to respond to now, right. To the concrete example, that you say, I agree with you that like in the short term, all of these developments are really great. And honestly, I as I said, I don't think it's a priority or problem that lots of people will will do other jobs because AI can do the one that they're currently doing. This is not even what I'm worried about, right? So I'm explicitly saying, the existential risk is the risk. I actually think that lots of the other risks are a little, let's say overblown. So for example, problems with discrimination. I think discrimination will actually decrease because we're able to quantify it much more precisely, so that we can sort of be on the boundaries of optimal fairness and that we can mathematically understand exactly, so I even the problem of like misinformation, disinformation that is automated via GPT. This I think, will be quite challenging. But basically, this is again, something that everybody's already thinking about everybody is aware of. Yeah, so I actually think these are talked about too much. This is why I focus so much on this existential scenario. And let me just say that you were saying, oh, right now the tools can help you with all these amazing things. Yeah, that's, that feels so good. And we should have more of this. Yeah. We don't need to develop GPT six, that, you know, can actually autonomously plan to take over the world. In order to have the amazing future that you're dreaming of. Right. This is precisely the point. It is perfectly fine to slow down. On the development of the most advanced systems that we simply do not

know how to control and this is not just me saying this, right? This is Stuart Russell saying it or Geoffrey Hinton, we have no idea how to control them. So we just slow down there and we develop all the other amazing applications, you know, for hydrogen policies and mineral mining all these things or searching

Tarry Singh 42:43
For a curable diseases. Yeah, sorry, when searching the cure for incurable diseases. Yeah, fighting for economic inequality stuff. Like that. Right?

Alexander Meinke 42:53
Exactly. But, but all of our efforts may get cut very, very short. If we don't take into account the real risks that all these companies they're pushing for. For the big jackpot. The big jackpot is an autonomous agent that can compete with human intellectual labor on basically any task. And, and you said, Well, don't trust the intrapreneurs and the startup people on this and I agree, they are specifically selected for to be the risk takers, right? So when they say okay, there's a 10% risk this will kill every single human on Earth. But there's a 90% risk, it will make us the richest people in history. Well, which path do you think they're going to take? And this is why we need the public to be aware of this and why we need external boundaries to say no, you cannot develop models with these insane capabilities that until you can prove that you wouldn't be able to control them. This is all I'm saying. Right. I'm not going to argue against all the other amazing technologies that we can build with this.

Tarry Singh 44:00
Alright, this has been a long discussion, and I'm sure we'll have more talks like this in the future. For the last 10-15 minutes, I'd like to introduce a few people. As you can see, I've invited Richard to the room. Richard, I've looked at your profile, so feel free to raise your hand. I'm curating this conversation and setting some boundaries to make sure we understand who you are. I hope that's alright with everyone.

Richard, welcome to this chat. I don't know you personally, so I hope you're going to share something insightful. According to your profile, you've spent a couple of years as the Chief Architect at a company called Squirrel AI Learning. Do you have something you'd like to share or ask us?

Richard Tong, 44:51
Sure, I'd like to share a bit. I currently serve as the chair for the IEEE's Artificial Intelligence Standards Committee. Our committee is developing standards that help define AI risk, AI governance, and various technology standards that can aid better integration of large language models and the like.

I'd like to make a few comments on AI risk. When we discuss risk, opportunities, and the dilemmas we face, we should consider where the risk stems from. Essentially, risk comes from power. The more potent something is, the more risk it poses, but also the more potential benefits it can bring. For example, consider nuclear energy. If used beneficially, it can revolutionize energy. However, if not controlled well, it could pose an existential threat to all of us.

We see a similar dynamic with AI. To understand where the risk comes from, let's consider a scenario where a bad actor, be it an individual like Osama Bin Laden or Putin, or an AI itself, tries to use AI

maliciously. The combination of a bad actor and AI is likely more dangerous than AI alone, both now and in the future.

A nuclear bomb in a lab without anyone to activate it poses little risk. However, if a bad actor has access to it, the situation worsens. This combination of a bad actor and technology is what we should truly fear.

We should also differentiate between direct and indirect risks. Direct risks arise when a bad actor uses AI with malicious intent, such as hacking, causing infrastructure disruption, or spreading fake news. Indirect risks, on the other hand, occur when we use AI with good intentions, but errors in the system lead to unexpected disasters.

While both types of risks are significant, the first one, involving a bad actor, is more concerning. For the latter, we can implement safeguards, control systems, self-evaluation mechanisms, explainable AI, bias detection, and other measures. However, if a bad actor utilizes a powerful AI agent, that's a scenario we should be extremely wary of.

Finally, we need to think about the physical power AI can control. If AI systems control our nuclear arsenals or are used for hacking or information warfare, that means AI is at the helm of considerable physical power that can be misused. Even a small error in this process could lead to catastrophic consequences. That's why we need to be extra careful when designing and implementing such systems.

Tarry Singh 50:33
Yeah. That's a very famous example people have used in the past as well.

Richard Tong 50:39
Certainly, as a community, we need to scrutinize where the risk originates from and how power is created. We need to understand what actions AI takes to acquire power and, as humans, what we permit AI to control, particularly when it comes to physical connections, infrastructure control capabilities, and potential infrastructure disruption capabilities.

Tarry Singh 51:14

Absolutely, thank you very much, Richard. These are discussions that are surely happening at IEEE, and we wish you all the best in aiding society and scientists in understanding where these risks are originating from, who the actors are, and how power structures are formed. This is extremely helpful. I've also brought in someone I know quite well, Dr. Rao Mikkilineni. Perhaps, Dr. Rao, you could share a few of your thoughts. We only have six more minutes before we conclude, but before we get to that, Dr. Alex, do you agree with what Richard just mentioned?

Alexander Meinke 52:12
Yes, I agree with a lot of what was said. I concur that there is a distinction between accidental risk and misuse risk, and also that the more powerful a system is, the greater its potential benefits and harms.

However, I want to reemphasize that power seeking is precisely the issue. Richard implied that we have methods to limit the amount of power seeking an AI system can do, but we don't. The techniques he mentioned, like explainability, are significantly underdeveloped if we want them to address this specific problem. We can only hope they are developed in time.

Furthermore, even if, as Richard said, there's a danger associated with AI alone, and on top of that, a risk of misuse - which I wholeheartedly agree with - it only amplifies the points I've mentioned about the danger. To me, this sounds like even more reason to agree that something needs to be done urgently and quickly, or else we're really in trouble.

Tarry Singh 53:36
Thanks, Alex. As I said earlier, we wanted to involve more people in this discussion, but with just five of us starting off as three, it's already a fruitful discussion. Dr. Rao, welcome. I hope everything is well in the Bay Area. I'd love to hear your thoughts, Dr. Rao, on the pros and cons of artificial intelligence. When McCarthy couldn't secure funding for his research, he coined the term "artificial intelligence" because it was attractive and they could raise the capital needed. Now, 60-70 years later, research has turned into products and services, and data has been mined and turned into intelligent language models and tools.

This journey from people researching and coining all kinds of terms has brought us here today, where there is a lot of fear, some of it for good reason, and some of it could be considered scaremongering, which I don't subscribe to. But there is also an exaggerated perspective of technology, which I may be guilty of being a part of, perhaps naively, because I see the benefits mainly coming from productivity.

Dr. Rao, if you're on the phone, you might need to unmute yourself. If you can't, I'll continue to drive this conversation while you figure it out. Thanks to Nasr, Alex, and Richard. There were some challenges, so Dr. Rao wasn't able to join us.

Firstly, thank you all. Richard, I appreciate your perspective and tend to agree with you. The dangerous combination is when a man is given a tool. The tool can either be used to fish for himself and his village or it can be used to harm others. It's all about humans using tools appropriately. In my opinion, that's what can help humanity.

I don't believe any of us want AI to take over completely. I'd like AI to take over many of my mundane tasks. I wish for better ways to communicate with my team and my employees, better ways to serve my customers, and to be able to write more papers annually because I love being productive. I wish for tools that automate and solve software problems, and for faster ways to build user interfaces for our tools. But in the end, I think we all want a balanced approach.

Alex and Nasser, I truly appreciate your perspectives. I believe the balance between the pros and cons lies somewhere in between. You're right; sometimes we have to put guardrails in place, maybe even pressing pause, although I'm not in favor of that. But I guess that's where we are today.

Dr. Rao, it was a bit of a struggle to get you here, but I'd like you to summarize this discussion for us. With decades of industry experience and approaching your 80s, your wisdom is probably the most important for all of us here. You've seen this work evolve over almost eight decades, and your guidance is invaluable.

Tarry Singh 58:16
Thanks, Alex. As I said earlier, we wanted to involve more people in this discussion, but with just five of us starting off as three, it's already a fruitful discussion. Dr. Rao, welcome. I hope everything is well in the Bay Area. I'd love to hear your thoughts, Dr. Rao, on the pros and cons of artificial intelligence. When McCarthy couldn't secure funding for his research, he coined the term "artificial intelligence" because it was attractive and they could raise the capital needed. Now, 60-70 years later, research has turned into products and services, and data has been mined and turned into intelligent language models and tools.

This journey from people researching and coining all kinds of terms has brought us here today, where there is a lot of fear, some of it for good reason, and some of it could be considered scaremongering, which I don't subscribe to. But there is also an exaggerated perspective of technology, which I may be guilty of being a part of, perhaps naively, because I see the benefits mainly coming from productivity.

Dr. Rao, if you're on the phone, you might need to unmute yourself. If you can't, I'll continue to drive this conversation while you figure it out. Thanks to Nasr, Alex, and Richard. There were some challenges, so Dr. Rao wasn't able to join us.

Firstly, thank you all. Richard, I appreciate your perspective and tend to agree with you. The dangerous combination is when a man is given a tool. The tool can either be used to fish for himself and his village or it can be used to harm others. It's all about humans using tools appropriately. In my opinion, that's what can help humanity.

I don't believe any of us want AI to take over completely. I'd like AI to take over many of my mundane tasks. I wish for better ways to communicate with my team and my employees, better ways to serve my customers, and to be able to write more papers annually because I love being productive. I wish for tools that automate and solve software problems, and for faster ways to build user interfaces for our tools. But in the end, I think we all want a balanced approach.

Alex and Nasser, I truly appreciate your perspectives. I believe the balance between the pros and cons lies somewhere in between. You're right; sometimes we have to put guardrails in place, maybe even pressing pause, although I'm not in favor of that. But I guess that's where we are today.

Tarry Singh 1:01:34

Yes, Dr. Rao. As part of the younger generation, we've seen figures like Geoff Hinton make significant contributions, such as the development of backpropagation algorithms in the late '80s. You've worked with other researchers and Nobel Laureates, including Dr. Roger Penrose.

Do you think that mathematicians and the scientific community, like Dr. Penrose who has written insightful books like "Emperor's New Mind", should also try to provide balance in this discussion? It seems like we oscillate between being overly optimistic and overly pessimistic. However, I believe what humanity needs is a moderate, middle-of-the-road approach.

Rao Mikkilineni 1:02:29

You're absolutely right, Tarry. What we need is balance. There has always been good and evil in human beings. Since humans became intelligent, there have always been those who use their intelligence for ill. Thus, how we use and regulate technology and make it transparent so everyone understands what's happening is crucial.

Geoffrey Hinton, for example, is bringing attention to the potential misuse of AI, which is useful. If we understand how people might misuse AI and understand the nature of its use, we can then develop appropriate controls.

I would suggest studying the general theory of information, model-based reasoning, and supersymbolic computing. These areas are becoming increasingly well-known. I also recommend reading Damasio's work on the evolution of human intelligence and the role of homeostasis. This knowledge can help us build a higher-level architecture, combining symbolic and non-symbolic computing, similar to the neocortex of the brain.

This is, in essence, what AI is: a vast pool of knowledge like the cortical columns in the brain, providing data for higher-level reasoning. However, what's currently missing is the ability for AI to recognize that it has this knowledge and what to do with it. It lacks purpose, while human intelligence is driven by survival, sustenance, and security. We need to infuse these fundamental drives into AI.

Tarry Singh 1:05:16

All right. This is with these wise responses. I really like this exchange of opposing views.

Alexander Meinke 1:05:25

The term "bullshit" was used earlier, so I believe it's fair to address that. First of all, the notion that sufficiently advanced agents will inherently seek power isn't a result of anthropomorphizing. We're not saying "Humans seek power, and intelligent humans can accumulate a lot of power, so AI will too." This isn't our reasoning. We can mathematically demonstrate that advanced agents are intrinsically driven to seek power in mathematical environments.

Secondly, as I understand it, your entire premise is that there exists this general theory of information, which you are working on and promoting as much as possible. You're claiming that significant progress in AI can't be made until people adopt your theory. This is a very bold claim, suggesting that the hundreds of billions of dollars being poured into numerous projects should be redirected to implementing your theory.

It's not that straightforward to dismiss the experts who are building state-of-the-art models and calling them dangerous just because they haven't read a specific, not very highly cited paper. I'm not quite sure what your claim is here. So, I'd like to ask you, is there a specific prediction that would disprove

your belief? Is there a particular experiment that, if AI were able to achieve it, would make you concede that your prediction was incorrect?

Rao Mikkilineni 1:07:22
Certainly, I'm not suggesting that language models like GPT are inherently bad or incorrect. Language models are, in fact, a robust method for capturing information from a large variety of sources, such as textbooks and medical literature. This results in a pool of knowledge that people can utilize.

What people are doing is using this pool of knowledge to write higher-level programs, which can be used for good or ill. It's not the language model itself that's attempting to dominate the world. Rather, it's the programs using the knowledge derived from the language model that are capable of doing good or evil. Thus, the issue of control is significantly different from blaming the language model or suggesting that AI will take over the world. AI isn't seeking to take over the world; it's the people who may misuse AI with that intent.

To mitigate this, we need transparency. This means that individuals can start writing programs that use higher-level reasoning and leverage the knowledge derived from language models. This model would be transparent, visible to everyone.

I'd like to clarify that there are methods to achieve this and it's not my personal theory that I'm advocating. As a student of the General Theory of Information, I learned a lot from the late Professor Mark Burgin, and I'm trying to disseminate that knowledge. I don't have a personal agenda here. I don't own a company, and I'm not developing anything. My sole purpose is to teach people what I've learned from the General Theory of Information and share it.

Tarry Singh 1:09:28
Alright, Dr. Rao, Nasser, Alex, I sincerely apologize. Dr. Maryam, I introduced you late to the call and unfortunately, we need to wrap up. I know you're eager to share something. What we can do, however, is allow you to quickly introduce yourself because we need to conclude as my colleagues need to return to their work. As far as I understand from your profile, you're a professor at a university in Pakistan. Welcome! We plan on setting up...

Tarry Singh 1:10:45
Thank you, everyone. I just want to summarize on behalf of all of us. Thanks to each of you, I truly appreciated this. I had hoped for a heated debate, Nasir, Alex, Dr. Rao, and I believe we achieved that. I think we should avoid focusing on citation counts. Trust me, Alex, Dr. Carl Friedrich Gauss, back in the mid-19th century, prioritized science over popularity. His brain is still preserved in a museum due to its exceptional convolutions, and he had only a few PhD students. So, let's not fixate on vanity metrics like citations or whether we've heard about each other's papers. I believe we all need to maintain an open mind about the risks and opportunities.

I thank each one of you for your contribution. I love this talk because if there are sparks, it means we are all passionately fighting for the same future. Thanks to all of you. I'm going to end this call now. I

sincerely appreciate each of you. It's through such fervent debates that we will build a future together. Thank you very much. Take care, God bless, and let's all continue building safe and responsible AI.