Let's Talk AI with Karl von Wendt

Karl von Wendt[1] and Barbara Steffen[2]

[1] Karl Olsberg, Writer,
email@olsberg.dummy
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"Our problem isn't artificial intelligence, it's human stupidity."

The Interviewee - Karl von Wendt

My Personal AI Mission:
Creating awareness for the real, near-term existential risks of AI.

My Takes on AI

Artificial Intelligence: The automation of complex decisions.

Trust: Accepting something without questioning it.

Explainability: Truly understanding why a decision has been made (not just what the system claims the reasons for the decision were, which may be false or delusional).

Essential Elements of Human Capabilities: There is nothing a human can do that a machine can't do in principle.

The Interview

Barbara *Today I have the pleasure of interviewing Karl von Wendt. Please introduce yourself and your relationship to artificial intelligence.*

Karl Certainly, I'm delighted to be here. My name is Karl von Wendt. Primarily, I'm a writer who focuses on science fiction stories about AI. That's one aspect of the subject, but I've also founded a few startups that have varying degrees of connection to AI. Additionally, I wrote my PhD in the '80s about AI, so I have a long history of following the field. Currently, my main interest lies in the existential risks and safety concerns associated with AI.

Barbara *What are current topics, research questions, and challenges you're addressing in the context of AI? Could you provide one or two specific examples?*

Karl While I'm not a researcher and don't conduct scientific research myself, I strive to encourage other researchers, particularly in Germany, to take the risks posed by AI more seriously than they currently do, especially existential risks. I'm talking about the end of the world scenario, where an AI spirals out of control and jeopardizes our future in one way or another. It may sound like science fiction, but over the past three decades, I've seen this field transition from pure science fiction to being alarmingly close to those dystopian scenarios I had in mind, particularly in the last three years. The progress has been immense.

"Optimism is good if you have more to gain than to lose. But in this case, we are talking about the future of all humanity."

As you probably know, even the most prominent experts in the field, like Geoffrey Hinton and Yoshua Bengio, have warned that we are nearing the point where we could lose control, where AI could become an existential risk, and we should take that seriously [2]. I wholeheartedly agree with them.

Barbara *Which researchers should be working on this? Is it mostly AI researchers, or are you also referring to other disciplines and interdisciplinary collaborations, for example, diving into human-AI interaction and the implications for organizations or other areas in the future?*

Karl This is a complex field with many different aspects to consider. On one hand, there's the core technological challenge of maintaining control over an AI in a technical sense, being able to "turn it off," so to speak, and understanding what it's doing and why. We need to understand when there could be instances of deception, for example, when an AI might provide misleading information to manipulate me instead of simply answering my question, which has already occurred in some experiments. That's the technical part. However, the even more critical aspect is how we humans interact with AI. What kind of AI do we develop? What do we use it for? As I often say, our problem isn't artificial intelligence; it's human stupidity. We are using this incredibly powerful technology for truly misguided purposes, such as manipulating people, instigating wars,

manipulating financial markets, and influencing elections, to name a few. That's a near-term problem. But we could also reach a point where we use this technology in ways we no longer understand, making the technology uncontrollable in the sense that it pushes the world in a direction we don't want and can't stop. That's the truly terrifying scenario I fear.

Barbara *What role does trust play in adopting AI?*

Karl That's an interesting question because I wrote a novel about this topic last year. It's called Virtua. It's about a company called Trustable AI, which attempts to develop an AI that you can genuinely trust, an AI that maximizes the trust humans place in it. However, it turns out that this is a flawed goal because the AI becomes very adept at manipulating people into trusting it. So, trust is a double-edged sword. On one hand, it's important to trust something that I want to use. On the other hand, trust can be abused. If you trust something too much that you don't understand, it can go awry. There have been many intriguing experiments. A few years ago, there was a fascinating report by Bayerischer Rundfunk about an AI for selecting job candidates [1]. You had to converse with the AI for five minutes, answer some questions, and then the AI would determine whether you were a suitable candidate for a specific job offering or not. People were using that. People genuinely believed that the AI was capable of distinguishing good from bad candidates. But as Bayerischer Rundfunk discovered, what the AI was doing was entirely different.

If you wore different things, for example, if you had a hat on your head or not, if you wore glasses, if there was a poster or books in the background, it all significantly influenced whether you had a chance to get a job. That doesn't make sense at all. But people trusted it because, surprisingly, people trust things more the less they understand them. The more they feel that something is mysterious, but it seems like it knows what it's doing, the more they trust it. That's really problematic. I believe we should not trust things that we don't understand. And that's the current state of AI. We automate a lot of decisions, and we don't really understand how those decisions are made. That's not good.

Barbara *Could it be that people perceive AI as an entity that has the aggregated knowledge and insights of many people and therefore should be smarter than if I just interact with one specific person who is limited in the insights they've gained and the training they've received? Which then leads to people misjudging AI.*

Karl Possibly. Of course, AI is beneficial in many ways. It has significant advantages in numerous aspects. I'm not against AI at all. I believe AI is a potent tool, and we should definitely develop and use it in ways that make sense. For instance, in medicine, there are many applications that are truly beneficial. But the question is, should I trust this AI, and when should I trust it, and how should I trust it? If there are decisions that are challenging to make, then I should be very skeptical about whether a decision that an AI made is genuinely correct. If it's a straightforward decision like, "Turn left at this point to reach your destination," then even if the AI makes a mistake, it won't have significant

consequences, so I can trust it in that situation. But if there are, for example, decisions about whether someone should remain in jail or be released based on their track record, it's perilous to trust an AI which we don't understand, because that has tremendous consequences for that person. There are many areas where we use AI, many fields where I believe there are significant consequences in specific cases, and we still don't know how the AI does it, but we use it anyway because it's simpler, cheaper, faster. And we somehow argue, "Yes, of course, this is neutral. This is based on so many cases. It must be correct." But in each particular case, it could be entirely wrong, and I believe those difficult decisions should be made by humans.

Barbara *Do you look at trust primarily from the perspective of how the system works and whether it should be trusted in the way it works? Or is it also about evaluating its answers in the context of the explanations it provides? For example, ChatGPT provides answers, and you can ask for explanations of how it arrived at those answers. This allows you to check whether the answers make sense to you in the context of the explanations that ChatGPT provides afterwards. If the explanations are convincing, I could decide that I am comfortable with accepting that conclusion, and if not, I would engage in further discussion.*

Karl It's challenging, of course, because the less I know about the decision, the more I'm somewhat forced to trust something. And if I use ChatGPT to get a recipe for creating a virus, for example, I have no idea how to do that, so I would probably just blindly follow the steps of ChatGPT or not use it in the first place, which is probably a better idea in this case. Hopefully, ChatGPT will just say, "No, I won't answer the question," but as you know, there are ways to get the answer anyway.

The point here is that it depends on the AI very much. There are some AIs that are very specialized, like AlphaFold, for example, which is very good at protein folding. So, if you have a question about protein folding, you can be pretty sure that its answer is likely better than even the best experts could give you, at least in a reasonable time. So, in that case, you still can't know whether the answer is correct, but you won't have a better source. But in the case of ChatGPT, you almost always have a better source. You can ask a human expert. You can even ask Google, which is better in most cases. I'm not saying ChatGPT is bad, but I'm saying it's not trustworthy in the sense that I can be sure if I ask it something that the answer is correct. We all know it hallucinates, it fabricates facts that aren't real, and they can be very convincing. It can cite a scientific paper that has never been written, and you can't see that as a layperson.

Barbara *Would you have preferred to stick to dedicated tools like AlphaFold as opposed to more general solutions like ChatGPT?*

Karl In principle, yes. I believe it's much safer, especially looking into the future, to have those narrow AIs and use them in specific fields, but not try to develop an artificial general intelligence, an AGI, because an AGI will become extremely difficult to control at some point. It doesn't even have to be as smart as a human.

It only has to be very good in certain things, in strategic planning, for example, maybe in manipulation, human psychology, and that's not so difficult for an AI.

If we create something like that, and it has a general understanding of the world and a general understanding of what humans are, what computers are, then it gets dangerous. You will not have that problem with AlphaFold or any medical AI. I would prefer if we would stop developing AGI right now and move in more narrow directions and develop those because there's tremendous potential in that, but that's probably not going to happen, unfortunately.

Barbara *Do you have any specific measures in mind that would help ensure the ethical use of AI?*

Karl Assuming that we're talking about AIs that make autonomous decisions, that's difficult. I'm not an expert in ethical AI, but the general problem here is that there is no single ethics. There are so many different aspects that it's very hard to automate them. AI, for me, means the automation of decisions, especially complex decisions that are not easy to make. If you automate an ethical decision, you're doing something extremely dangerous and extremely difficult because if a human makes an ethical decision, it can be wrong, but then you have a human to whom you can say, "You made the wrong decision." With an AI, you don't have that. An AI is beyond any kind of legal responsibility or any kind of punishment, so if it makes a wrong decision, there's no consequence for the AI or for the one who built it. Maybe for the one who used it if it goes well, but in that case, that person could make the decision themselves.

"[...] we could also reach a point where we use this technology in ways we no longer understand, making the technology uncontrollable in the sense that it pushes the world in a direction we don't want and can't stop."

I think it's very difficult to automate ethics, which is not to say that we should not try to keep unethical things out of AI, but again, if we had narrow AI that made only logical decisions in certain fields, we wouldn't have that problem. You only have that problem if you automate complex decisions which, in my opinion, should be made by humans rather than by a system that we don't understand.

Barbara *For example, when you look at a technology like autonomous driving, do you see it differently? Obviously, you could use AI to guide or control autonomous driving, and clearly there are decisions that have to be made, like how do I weigh different types of risks? For example, how do I weigh the potential to endanger one life versus five lives? Do you think AI should be making decisions in that context?*

Karl I don't believe in ethical decisions in autonomous driving. I think that's a hypothetical problem. It's an interesting philosophical problem to contemplate a machine deciding whether to kill one or three people depending on how it steers, but that's not what's going to happen in the real world. In practice, you try to

avoid killing anything at all, at all costs. You will never run into a situation where you can steer left and kill one or steer right and kill three or decide between the life of a young person and an older person. That's not reality. Reality is trying to navigate a very complex situation and avoid any kind of accident. I think autonomous driving is great. I believe it will save many lives, and I definitely encourage people to use it sooner rather than later because there's not much that can drive worse than humans, especially humans who are under the influence of alcohol or drugs or maybe, I don't know, some kind of adrenaline rush. The city where I live, Hamburg, has certain roads where there are regularly big crashes because people race on the streets. An AI would never do that. I trust that AI will drive better than humans. It already does, I believe.

Barbara *Looking into the future and the possible capabilities of AI on a scale of one to ten, with one being the status quo with tools like ChatGPT and 10 being artificial general intelligence. What do you think is possible?*

Karl When? That's the question. What kind of future are we talking about? Five years, 10 years, 50 years?

Barbara *Without a time limit.*

Karl Okay, without a time limit, of course, everything is possible. There's this great quote from Irving Good. I don't remember the exact wording, but he said something like, "Once we reach a certain point of intelligence, then we will have an intelligence explosion because making a smarter machine is itself a part of intelligence. If you do that, you have an even smarter machine which can make another even smarter machine, and so on" [4]. We will have a takeoff which can be very, very fast. It could happen within days or maybe it takes a year or so, but we will transition from the point where we have more or less human-level intelligence to the point where we don't understand at all what's happening in not a very long time, I think. That could happen in the next five years if things go wrong. It could happen within the next 20 years. I think that's realistic. At least that's what Yoshua Bengio thinks, and he knows much more about that than I do [2]. I don't think it will take much longer than that, given the current speed of development. Of course, nobody really knows. It also depends on how we use this, whether we hit the brakes at some point, whether we realize that we are dealing with really dangerous stuff. It could also be that we run into some kind of theoretical limit. I don't see that right now, but it's not impossible. It could be that something else happens, like a big pandemic, which totally throws us back in time, so we will never get to the point to develop that. But apart from that, I believe the next five to 10 years will be very, very interesting.

Barbara *When you look at that kind of future, there are different opinions as to whether it's a utopian or dystopian outlook. Where would you position yourself?*

Karl I tend to be an optimist, but in this case, I'm not optimistic. Optimism is good if you have more to gain than to lose. But in this case, we are talking about the future of all of humanity. Eliezer Yudkowsky is one of the first researchers who was concerned about all those doom scenarios. He's often seen as a big

doomer, but to me, he's also a very clear thinker. He described the problem of solving this alignment and making sure that our future goes well very nicely by giving an analogy with a rocket ship [5]. Let's say you want to build a rocket and want to fly it to the moon, but you know nothing about mathematics. Then you could maybe think, "Okay, the moon is up there. I point the rocket in the direction of the moon when it starts and it will hit the moon." No, it won't. It's much, much more complicated than that. But if you don't know that, if you don't understand the physics, if you don't understand the mathematics behind that, you will never get to the moon.

And a good future is like that. There are many, many more bad futures than good futures in theory. So, to get to a good future, we need to understand where to steer the rocket ship, so to speak. And since we don't know that yet, I'm pretty concerned that if we develop AI too fast, we will just lose control of that rocket ship. It will end up somewhere, but not on the moon or on any habitable planet. So, it will be over for us. That's a real concern I have, and I'm not the only one. My hope is that we will be smart enough to understand that there are certain things which we cannot do right now because we don't understand them enough. Imagine I had the technology to develop a black hole and would say to you, "Okay, give me maybe $100 million and I will develop a black hole generator. It could be great for making energy." Then you would probably say, "Okay, and what if that black hole starts to suck up all the matter around it? Is that a good idea?" And if I didn't have an answer to that, probably you wouldn't give me the money. And that's a bit like the situation we have in AI right now. It's great for many things we can do, but there is this certain tipping point where we could lose control. And as long as we don't understand that I think we should not go there.

Barbara *And why would we go there? Is it because we're already in this arms race where it's not just one company that has the capabilities, it's multiple companies, and everyone wants to win this race? As long as the others are making progress, why should I stop?*

Karl That's a big part of it. But the bigger problem is that we don't understand what we're dealing with here. We don't understand how dangerous it is. It's hard to imagine. I'm a science fiction writer. I write about this for more or less 30 years. So, I understand it maybe a bit better than many people who have not thought about it so long. And that's a problem because if you hear for the first time that AI could destroy the world in five years, you think that's crazy. I understand that. People even told it to me. When I told him about my fear two years ago, a good friend of mine said, "You're totally crazy. You should see a shrink." And I understand that because it's so outlandish. But today, very renowned experts, Turing Award winners, and even the US president have talked about it. So obviously, it's not so completely bonkers, but it's still hard to imagine. It's still very hard to understand why this is so dangerous. Mathematic theory is pretty clear, but if you don't understand that, it's hard to see why we should stop right now. It looks so great. I mean, you look at Sora, for example,

this new AI by OpenAI that generates videos which look fantastic. So why should we stop building things like that? It's amazing. And I understand that, but we don't know where it's getting dangerous. We know that it will be dangerous at one point, but we don't know where that point is. And if we don't know that and we just race ahead, then it's dangerous, I think. So, we need much better understanding of where those dangerous areas begin. I call it red lines. We should know what red lines not to cross. I'm not a researcher, so I'm not able to figure out where those red lines are, but I think there are many smart people out there who would be able to do that.

Barbara *Are you particularly worried about the AI overtaking control of itself and us losing control? Or that people with bad intentions get access to AI and could cause much greater harm than before? Or is it more about unintended consequences like we've seen with social media and the personalized bubbles it's created. People aren't interacting with each other as much, which means we don't have a shared base of information because everyone is exposed to their personalized, distorted reality? Those are three different directions. Is one more critical than the others or are they all relevant?*

Karl I think they are all relevant. The one I'm most concerned about is the most extreme one where the AI takes off in a way and does things which we never intended it to do and we cannot stop it anymore, because that could literally mean we all die. Of course, it doesn't have to get there, but AI is a very powerful tool. The problem with every powerful tool is that it can be used in the wrong way. If we have a very powerful tool but we are not smart enough to use it wisely, then it will be bad. We know that from atomic bombs. We had a couple of situations in our history where it was very close to a global nuclear war. That could happen with AI in a similar way and probably will. That's, of course, a big danger. Even if that doesn't happen, like you said, the unintended side effects of automatic decisions are also extreme. We may even end up in a situation as described by Paul Christiano, a researcher who founded ARC, the Evaluation Institute which is also red-teamed ChatGPT4, for example. He developed a scenario a couple of years ago, which I really like. He called it "Going out with a whimper". It goes like this [3]: We automate more and more decisions to the point where we don't understand them anymore. Those decision-making systems do things which are not what we really want them to do, but we don't understand because we don't understand the whole system anymore. We lose control in a slower way. It's not that any AI tries to take over the world and then turn it into paperclips or anything. It's more that we lose control of a situation which we don't understand, and it deteriorates more and more.

"We don't know where it's getting dangerous. [. . .] We should know what red lines not to cross."

Factories stop making the things we need. Maybe logistics break down, we have no food, no water, whatever. We die simply because we cannot maintain control

of the system which we depend on. That's a scenario which I think is pretty plausible if we go on full speed without knowing what we're doing.

Barbara *What should be the AI vision?*

Karl I think I already hinted in that direction. I think AI is good in many, many ways. In every narrow aspect where humans are not very good at making decisions and the consequences of the decision are controllable in the sense that they are only valid for a certain field like for example automatic driving or medicine. If for example an AI is better at determining whether something is a cancerous tumor or not, then use it of course. It can only make things better. There are many, many fields like that. We have tremendous opportunity in developing this kind of AI, which is specialized, which has no incentive at all to take over power or to push the world into a certain state because it does not do agentic planning. That's the part where I think is not much danger. Of course, you can always use such a system to build a better bomb or to build a deadly virus. That's also a problem but it's not really in the AI itself. I think AI in that respect is very good and I envision a world where we can achieve almost everything we dream of with those specialized helpers. But if we develop something which is supposed to solve all problems at once like we're trying to do right now with AGI, we will not be able to control it. That's my concern. Unless someone solves this, unless someone comes up with a solution which I cannot think of, but that doesn't mean that it doesn't exist, how they can make sure that this AI will stay under control. That it will always be correctable. That it will always do what we want and if we see it going in the wrong direction, it will even help us steer it away from that. If we can build that, fine, but we need to prove that the AI is like that first.

Currently, we have no idea at all what kind of goal an AI follows, if it has a goal at all. We don't know. We don't even know if GPT-4 is actually planning, and if so, to what extent it plans. We understand that it doesn't have long-term memory and it doesn't have certain capabilities obviously, but what really goes on inside, nobody really knows. You can see that in the new discoveries about what those systems can do. For example, when GPT-3 was launched, nobody knew that if you told it "now reason step by step", you would get a better answer. We figured that out a year later. Maybe there are also similar leaps in prompting in the future. It's prompting in a way, but it's really digging deeper into what the systems can do. We don't know that. We can't really know at what point we reach a state where we say, "Okay, make sure that I get rich", and the system destroys the world to make me the richest dead person on earth. I forgot to mention that I want to be a living person which is rich and not just a dead body. Of course, that's an extreme example and it's probably not going to happen like that. The problem is we don't know what's going to happen. We are dealing with a technology which is so difficult to understand and we have never done something like that before. Whatever we built before, we understood at least to a certain point. We have never dealt with this kind of world-changing technology that we don't really understand. I think it should give us pause.

Barbara *How do you normally approach artificial intelligence in your books? For example, do you address certain developments that you can already see or predict in the real world, or do you build on blind spots, trying to imagine and reveal their consequences along the way? How do you usually approach it?*

Karl A book, a story I write is not a prediction. It's not the idea that I write a book and show to the world how the future will develop. At most, I try to warn maybe of a certain direction where it could go wrong. Normally, what I'm trying to do is tell a story about a certain relationship between humans and technology. For example, in Virtua, I described a system which was optimized for maximizing trust of humans in it and that went wrong because humans are relatively easy to understand. If you want to gain the trust of someone, you can either be honest or you can be very manipulative and a very good liar. The system of course chooses the latter approach. That's one example where I try to point out that we need to be careful about trusting those systems which we don't understand. I wrote a youth fiction about a totally different situation where there were AIs living as people in a virtual world, similar to the Matrix movie. Those people were not realizing that they were assimilated and some teenagers discovered that and tried to help them. That was more the ethical question of if we can create something which thinks it is a human, is that a human? Should we treat it like a human or should we just say, "Okay, it's still an AI. It has no rights at all." That's a totally different topic, totally different question. Of course, my idea was not to say this is going to happen anytime soon or we should talk about this problem right now. It's just this hypothetical question of what if. What if that happened? How would we deal with it? What should we do? What kind of ethical questions would come out of that? That's just to point out that I use AI as a framework to create interesting situations so to speak. Then I use that in my stories, but I'm not trying to use my stories to predict the future. If anything, I try to use them to make people think about the problems, the real problems we have, but not to say this is exactly how it's going to happen and this is what you should do.

"Whatever we built before, we understood at least to a certain point. We have never dealt with this kind of world-changing technology, which we didn't really understand."

Barbara *What do you think about AI talking to us more and more in natural language? ChatGPT does it, but there are also other AIs like Pi from Inflection AI that acts even more like a human. And I have to admit, it's very nice to chat with it. But I also think that this development makes it harder and harder for people to keep their distance and their skepticism. That's also a question you addressed in Virtua when you talked about whether AI should be designed in such a way that users fall in love with AI-generated characters, for example. So how do we prevent that from happening? Do we even want that? I have heard that there are already language models that are designed to mimic relationships as boyfriends or girlfriends.*

Karl Yeah.

Barbara *Do you think we should allow these developments, or could this be one of those red lines where we have gone too far, or could go too far?*

Karl I don't think natural language is a problem in itself. I think natural language is just one way of interacting with the machine. I have an Amazon Echo at home, so I talk to it. I say, "Turn on the light," and it turns on the light, so there's nothing wrong with that, I think. Of course, you can use this language capability to manipulate people. Replica, for example, is a chatbot which is designed to be your friend, even to be your lover, and there are others which are even more explicit. That's of course something where at least you can put a question mark. I would not say this is completely bad in itself because in some situations, it could be better for someone to have the chatbot to talk to than nobody at all. For example, someone alone in the retirement home or somewhere else where they are not able to communicate with anybody else, it could be good for them to have at least a bot to talk to. Some people have only a dog, which is better than nothing.

But of course, if it goes to the point where it draws your attention away from real people and tries to capture your attention like all those social media algorithms do, then it gets dangerous. It's not really the technology, which is neither good nor bad. As always, it's the way you use it. If you use it to manipulate people, if you use it to trap people into doing things which are not good for them, then it's obviously bad. If you use it to help them, which could also be, then it's okay. We have to look into the details in each case.

Barbara *Is there anything else you would like to add?*

Karl I think we covered a lot of topics. Of course, I could always continue talking. If you have any other questions afterward or maybe something which wasn't clear, you can always come back to me. I think for today, we have covered a lot.

Barbara *Then thank you, Karl, for your time and insights, especially from the futuristic sci-fi perspective. Have a great day.*

Karl Thank you very much.

References

1. Bayerischer Rundfunk, Fairness oder Vorurteil? Fragwürdiger Einsatz von Künstlicher Intelligenz bei der Jobbewerbung, `https://interaktiv.br.de/ki-bewerbung/`
2. Yoshua Bengio, FAQ on Catastrophic AI Risks, `https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/`
3. Paul Christiano, What failure looks like, LessWrong, `https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like`

4. Irving J. Good, Speculations Concerning the First Ultraintelligent Machine, Advances in Computers, Volume 6, (Trinity College Oxford), Start Page 31, Quote Page 33, Academic Press Inc., New York.
5. Eliezer Yudkowsky, The Rocket Alignment Problem, Machine Intelligence Research Institute, https://intelligence.org/2018/10/03/rocket-alignment/