Let's Talk AI with Thorsten Helfer

Thorsten Helfer[1] and Barbara Steffen[2]

[1] Saarland University, Department of Philosophy, Germany,
[2] Thorsten.Helfer@uni-saarland.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"Most ethical problems about AI have been there before. They just seem more pressing now."

The Interviewee - Thorsten Helfer

My Personal AI Mission:
Education about AI and its societal impacts.

My Takes on AI

Artificial Intelligence: A system instantiates artificial intelligence if and only if it is artificial and has certain general or specific abilities to interact with its environment in a way that appears like human intelligence.

Trust: A trusts B if and only if A has confidence in the dependability of B.

Explainability: A system is explainable if and only if it can provide relevant explanations sensitive to contexts.

Essential Elements of Human Capabilities: I think there are no essential elements of human capabilities. There are relevant stereotypical features of humans like reasoning or consciousness.

The Interview

Barbara *Today I have the pleasure of talking to Thorsten Helfer. Please introduce yourself and your relationship to artificial intelligence.*

Thorsten Certainly, as you mentioned, my name is Torsten Helfer. I am a philosopher at Saarland University and currently involved in a project called „Explainable Intelligent Systems“. I've recently joined this project, just about six months ago. Here my focus lies mainly within the ethical implications of AI systems and the ethical need for explanations of these. Additionally, I am involved in a project set to start in January next year within the association Algoright e.V. In this capacity, I will be an ethical adivsor on projects within healthcare and digitalization in Saarland.

Barbara *What are some of the specific AI-related research questions that you are working on?*

Thorsten At present, my primary interest lies in the concept of the 'human in the loop'. I am exploring questions such as: Under what conditions do we want a human in the loop? Why do we want a human in the loop at all? Could it be that under certain conditions, the human is merely a scapegoat? It seems odd to want a scapegoat, but sometimes I wonder what other reasons we might have for wanting a human in the loop. At least in part, we want to introduce AI systems because they are more accurate and it is also unclear whether humans are less prone to unfair or biased decisions than AI systems. I do agree that a human-in-the-loop is relevant for trust issues but other than that I think we should be more critical and really examine in what scenarios a human-in-theloop is helpful and in what scenarios they are a waste of resources or simply a scapegoat.

"The question should be which AI is trustworthy, not which AI do we trust."

Barbara *In your opinion, what role should trust have in the adoption of AI?*

Thorsten This might be an unpopular opinion, but I don't believe that trust inherently holds value in this context. Trusting AI might make us feel better about interacting with it, but that doesn't mean the AI is ethically sound. The question should be which AI is trustworthy, not which AI do we trust. We could trust an AI system that is actually harmful, simply because it interacts with us in a certain way. However, trust could hold instrumental value, as most of us will likely only interact with AI if we trust it. So, while I'm unsure about the intrinsic importance of trust, I believe it holds instrumental value (vgl. [1]).

Barbara *Are there specific metrics or frameworks that already distinguish between high-stakes and lowstakes environments, such as a list of criteria whose importance is weighted according to context? Or are we just looking at specific scenarios? And if we only look at specific scenarios today, would it be possible to derive a general evaluation framework or something like that later?*

Thorsten Determining whether something is high risk or low risk largely depends on your ethical viewpoint. From my perspective, it depends on what is affected in the end. The number of people affected, the potential impact on people's welfare and well-being, and how people's rights, autonomy, and freedom might be affected, are all important considerations. These factors determine how high risk or low risk a certain AI system is. It's not as simple as categorizing certain sectors, like healthcare, as high risk. For example, an AI system in healthcare that applies Band-Aids might not be high risk. It depends on what's at stake within that area. Additionally, predicting how an AI system will impact values in the world, people's well-being, autonomy, and freedom, are empirical questions that, as an ethicist, I can't answer.

Barbara *Is it of ethical merit to think about future scenarios, for example, to establish what-if understandings? If we allow X to happen now, what will the consequences be? Or is ethics more focused on today's reality rather than on possible future scenarios?*

Thorsten I'm not entirely sure I understand the question. Are you asking if it's ethically justified to have a general AI?

Barbara *My question is whether we should look at future possibilities through an ethical lens to prepare for them now. For example, if something like "X" could happen in the future, should we act differently or prepare for it today? Essentially, should we map out different possible future scenarios and have a plan ready with strategies for how to deal with them if they do happen? Or do we wait and react only when these possibilities become reality?*

Thorsten Well, I don't think we should address it only once it's there. We should address it when people start to think about developing it, when it's in a planning phase. There's been a lot of philosophical discussion since the 1970s about experience machines, where you can plug in, like the whole matrix idea [2]. People have thought about that already. A lot of movies about AI have already been made. So people are already thinking about a lot of this stuff. But I'm not sure whether all of these ideas are that helpful. There's also this whole discussion about fear mongering among the debate about techno-optimism. Should we be more optimistic about everything? Should we be scared about technology in that sense? Usually, these ideas are depicted in a more dystopian way. I'm not sure whether these ideas are that helpful. Of course, we should be prepared for some bad outcomes, but it should be proportionate to what could actually happen. In this respect, politicians, ethicists, developers and computer scientists should work together and figure out what could actually happen and what should be done.

"I'm not even sure whether specifically philosophers are needed, but you need somebody who will look at the ethical and societal impact of what AI products might have in the end. Some people who will have a bigger-picture view of things."

Barbara *Do you feel that philosophers are sufficiently involved in the development phases? Is there enough collaboration to ensure that you can address new developments and challenges in time?*

Thorsten In my recent acticvities I was actually quite surprised, because I've been talking to computer scientists or developers, and I thought they would be really not interested or even averse to an ethical perspective. But usually, if you approach them in a way that you want to help them make their product within society better, they're usually quite happy to hear your opinion on everything. So I'm not sure whether it should be institutionalized that philosophers are involved. I'm not even sure whether specifically philosophers are needed, but you need somebody who will look at the ethical and societal impact of what AI products might have in the end. Some people who will have a bigger-picture view of things. And usually, it seems to me, that philosophers are pretty good at that.

Barbara *Regarding the different future scenarios that are being discussed, ranging from dystopia to utopia. Where on that scale would you place yourself?*

Thorsten I would be rather on the optimist side, I think. I want to distance myself clearly from all of this. Marc Andreessen, techno-optimist view, where everything has to have this kind of religious touch, where technology is basically the new religion, where everything is good and you don't need any kind of regulation, and the market will take care of everything, we don't need any kind of regulation, and all techno-ethicists are just fear mongers [3]. But as I said already, how I at least want to approach things is more from a supportive kind of view. I think ethicists and critics of AI systems should work together with developers and computer scientists in order to figure out the best route to go. AI systems have a huge potential, just see the potentials of personalised medicine for example, but the same system that could positively revolutionise medicine could create the deadliest toxins. So, I want to see it as a more optimistic side, but certainly there are risks involved and they should be worked on together.

> "Trust is something very different than trustworthiness. Look at the evidence from psychology. The best step to improve trust in an interactive robot is if it hands out flowers at the beginning of the interaction."

Barbara *Reflecting on the past days, what insights from other disciplines were particularly interesting?*

Thorsten This has been quite a ride, actually. I've never been to an interdisciplinary conference that was so productive. Usually, you have to translate a lot from one discipline to the other. What I do sometimes, people from law, people from computer science, or people from psychology don't understand all the terms that I use, and I don't understand all the terms that they use. So you have a lot of catching up to do with the other disciplines. But somehow this was really productive and really helpful. I learned a lot about how different kinds of explanations can work for different kinds of trust, and the framing of explanations, and how that can impact the kind of trust. Then I have learned a lot about the ethical groundwork of law. I always thought there must be some ethical basis on which specific laws grounded, but it I realized here that sometimes you might have reasons to put laws in effect that have nothing to do with morality in the end. It's just to incentivize certain behavior in order for a more productive society. So I found all of that pretty interesting.

Barbara *Is there a specific research question or topic that you would like to see addressed from an interdisciplinary perspective in the future?*

Thorsten All of this is very interesting. I'm still hung up on that whole human oversight thing. Before I started working on this, I thought there must be more than enough literature about this. There must be a lot of lawyers who have thought about human oversight. There must be a lot of psychologists who thought about human oversight and trust, and philosophers thinking about under what kind of circumstances we actually want to have human oversight. But it seems to me that that has just stated. It's just said claimed that we want to have some kind of human oversight. We want to have people in there having the last decision. And it was really strange for me to realize, not a lot of people have thought a lot about the specific conditions of human oversight and a human in the loop. What does that entail? When do we want it? How does it relate to trust? And, as I said before, are there cases where we especially do not want a human in the loop?

"Why do we want a human in the loop at all? Could it be that under certain conditions, the human is merely a scapegoat?"

Barbara *Why should I trust the human more than the AI? Does the human look at the aggregated information and see if it makes sense, adding credibility? Or do we automatically trust a human more just because they are human?*

Thorsten I don't think that you should trust it more, necessarily. Look at autonomous cars. It seems to me that autonomous cars will very quickly kill fewer people than people do and will be better drivers. So I think for some cases, the AI systems are better than humans. So I don't think that you should trust humans more, but, as far as I know, the empirical evidence shows us that many AI systems are trusted more when there is a human in the loop. Trust is something very different than trustworthiness. Look at the evidence from psychology. The best step to improve trust in an interactive robot is if it hands out flowers at the beginning of the interaction. This might be relevant for trust but not necessarily for trustworthiness. So it's not necessarily that you should trust people more, but you do.

Barbara *From your personal perspective, what should be the AI vision?*

Thorsten I don't understand the question. What is the AI vision? What the AI will bring in the future? What it should bring in the future?

Barbara *Yes, what AI should bring in the future.*

Thorsten Well, from a very pedantic, philosophical point of view, more well-being, whatever that means. I have no idea what that will bring in the end. I'm personally hoping for some kind of experiencemachine, where it could simulate a lot of experiences and then figure out what I want to experience and under what conditions, but that is far in the future.

Barbara *When you think about well-being, do you mean well-being in the present, well-being in the near future, or well-being in the distant future? Does it make a difference? And how can we approach the right context of well-being for society?*

Thorsten You mean, should we sacrifice a certain amount of well-being now to have more well-being in the future?

Barbara *For example.*

Thorsten Yeah, this has many dimensions. If you think from a purely utilitarian point of view, where you ask what the best world or outcome is, then clearly, it's better to sacrifice some well-being now to get a lot of well-being in the future. And there is a certain discussion within philosophy about longtermism and whether we should try to avoid even the slightest risk of human extinction, even if it costs us a lot of effort right now, and even if we have to sacrifice a lot of well-being right now [5]. But that seems to have a lot of problems for- I mean, if you look at the real world right now and how we make decisions, and at least in western and democratic countries, we decide based on a democratic system. And as it is right now, we only have people living right now actually deciding upon that. Future generations do not decide on matters that seems to influence them. Usually a lot of people right now are out for their own well-being or maybe for the well-being of the close people around them. The far in the future generations are not represented in all of this. And it's the question whether they should be represented in all of this and therefore in the democratic system. I'm not so sure what the answer is there. It depends on all kinds of ethical questions, the ethical kind of view about long-termism, the ethical kind of view about utilitarianism at all, about future generations, about how you want to deal with democracy [4].

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

Thorsten I'm pretty good. This and the whole conference was a lot of fun. Like I said, the interdisciplinary work was very new but rewarding for me. Learning more about AI from a computer scientist's perspective, from legal, from psychologist perspective, all a lot of fun. I would love to do something like that again.

Barbara *Perfect, I do, too! Thank you very much for your time and ethical perspective on AI and the present, Thorsten. Have a great day!*

Thorsten Thank you.

References

1. Andreessen, M. (2023): “The Techno-Optimist Manifesto”, https://a16z.com/the-techno-optimist-manifesto/.
2. Friedman W., (2006) “Deliberative Democracy and the Problem of Scope”, Journal of Public Deliberation 2(1).
3. Greaves, H. & W. MacAskill (2019): “The Case for Strong Longtermism”, Gpi Working Paper.
4. Nozick, R. (1974): Anarchy, State and Utopia, Basic Books, New York.
5. Reinhardt, K. (2023): “Trust and trustworthiness in AI ethics”. AI Ethics 3, pp. 735–744.