Let's Talk AI with Eva Schmidt

Eva Schmidt[1,2] and Barbara Steffen[3]

[1] TU Dortmund, Department of Philosophy and Political Science, Germany,
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany,
eva.schmidt@tu-dortmund.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"Developers of AI systems need to be aware of and informed about the ethical and societal impacts of their products."

The Interviewee - Eva Schmidt

My Personal AI Mission:
Applying concepts and tools from philosophy to contribute to the development of ethically unproblematic AI systems.

My Takes on AI

Artificial Intelligence: I find the distinction between weak and strong artificial intelligence (AI) from philosopher John Searle to be the most relevant here. We can say that a system is a weak AI system when it is able to (merely!) simulate mental abilities, especially abilities to solve specific problems. For example, some systems are able to correctly classify cat images. Strong AI systems have a broad range of genuine mental abilities such as understanding language, solving problems intelligently, or playing games. So far, no strong AI systems exist.

Trust: I am more interested in reasonable trust than in mere trust. Regarding autonomous AI systems, my view is that a user reasonably trusts such a system only if the system is trustworthy and she is in a position to know this; and it is trustworthy for her only if it shares her goal and pursues it competently on the basis of the information relevant in the context.

Explainability: An AI system is explainable in a certain respect, given a certain context, just in case there is information available to a relevant stakeholder in the context, which can positively affect the stakeholder's understanding and thereby contribute to the fulfillment of the practical interests of the stakeholder, or to the fulfillment of certain societal desiderata more broadly.

Essential Elements of Human Capabilities: According to philosopher Helen Steward, what is specific about human actions is that they are exercises of so-called *two-way powers*. These are powers that a human agent is able to exercise – *or not* – at a particular moment. (Think of yourself standing in front of the open fridge – you can reach for the apple juice, or you can decide not to do so.) By contrast, one-way powers are manifested by the object that has them whenever it gets into the right conditions. (Think of a fragile glass, which breaks whenever it gets into the condition of being struck.) Plausibly, current AI systems do not have two-powers, as humans do.

The Interview

Barbara *Today I have the pleasure to talk to Professor Eva Schmidt from TU Dortmund University. Could you please briefly introduce yourself and your personal relationship to artificial intelligence?*

Eva Yes, thank you. I am a professor in theoretical philosophy at TU Dortmund University, in the Department of Philosophy and Political Science. I have been working in the field of explainable AI for some time now. My academic career started with studying the philosophy of perception and reasons, as in reasons to believe things or act in certain ways. A few years ago, Kevin Baum approached me and asked if I would be interested in delving deeper into the area of explainable AI. Given the importance of this topic, I found it interesting to contribute as a scientist and specifically as a philosopher. Since I was already working on explanations of human behavior, particularly human actions [5], I thought it would be an interesting challenge to explore whether we could apply explanations of human actions to explaining the outputs or actions of artificial intelligence systems.

Barbara *Can you give examples of one or two specific topics or questions your AI research is currently addressing?*

Eva I am especially interested in the explanatory information provided by explainable AI and how we should interpret it. What makes for a good explanation? How does it relate to understanding? At this conference, I presented parts of a paper that considers which contextual factors determine whether a person truly understands something [9]. I believe that it's easier to gain understanding when you have to make an unimportant decision and harder when the decision is significant. For example, consider a judge who must decide whether a convicted criminal should receive a harsh or mild sentence. This is an important decision, as it determines whether the individual will be imprisoned for a longer or shorter time. If the judge receives a high-risk score for the convict from an AI system, suggesting that the person is likely to commit another crime in the future, the judge might want to incorporate this input into their decision and impose a harsher sentence. In such cases, I believe it's essential to understand why the system provided this score. We know that some of these systems exhibit racial bias, assigning harsher scores to people of Color than to White people. In these situations, it's necessary to understand why the system gave you this score, but it might not be easy to understand due to the importance of the decision and the high stakes involved.

> "Can people still make responsible decisions when their decision-making is based on the output of AI systems? Are those AI systems fair? Do they comply with certain legal regulations? Do they violate any human rights? These are ethical questions."

Barbara *Do you think there's a difference between the tools and explanations needed for judges to support their sentencing or doctors to support their diagnosis versus patients who don't have the expertise like someone with a medical background and therefore need simpler or maybe completely different explanations? Also, do you think that patients might trust these tools differently than experts?*

Eva Absolutely. I believe that the context is extremely important when deciding whether something is a good explanation for the output of an AI system or for how a system functions as a whole. One factor I've already discussed is the stakes involved in the context. Another aspect is the individual we are considering. What's interesting is their level of knowledge. How much do they know about the field in which the system is applied? For instance, a medical doctor has extensive knowledge about medical care, while a patient may not. Additionally, how much knowledge or expertise does this person have in computer science or the theory of how AI systems work? I think these factors make a significant difference because if you have more expertise in a certain area, you can better integrate an explanation you get with your existing knowledge [7].

Barbara *What role does trust play in the adoption of AI?*

Eva Trust is an interesting topic, especially from a philosophical perspective. There has been a broad philosophical debate about trust for decades, but it typically deals with trust between persons [1]. It's interesting to consider whether we can apply some of these theories or approaches to trust in technological systems or perhaps combinations of technical and social systems. It's often not just the system by itself. Many people treat the issue of trust as a central concern, asking when can we trust a system? When do people trust AI systems? But also, when are such systems trustworthy? These are distinct questions [8]. What does it take for an AI system to be trustworthy? And what does it take for people to be good at judging whether a system is trustworthy? These questions are all related, but they are different. Some people consider this to be a very central issue. For example, the high-level expert group of the EU discussed trustworthy AI in their paper on this subject [4]. I do believe that trust plays an important role if you're interested in the appropriate adoption of these systems. However, I think it's only one of many factors. It's also important to consider whether people can still make responsible decisions when their decision-making is based on the output of AI systems. Are those AI systems fair? Do they comply with certain legal regulations? Do they violate any human rights? These are ethical questions. There are many other questions that are at least as important as the question of trust. As for the ethical component of your question...

Barbara *What measures do you think are essential to ensure ethical use of AI?*

Eva That's a very broad question. To ensure that AI systems respect people's rights and meet ethical requirements, I think we need to consider different contexts. One crucial aspect is having regulations that make sure the use of these systems doesn't violate any rights or harm anyone. Another approach, which I find very interesting from a philosophical perspective, is to see if we can embed ethics into these systems from the inside, rather than just observing how they are deployed and trying to enforce compliance from the outside. I don't work in this area specifically, so I may not be up-to-date with the state of the art. However, my impression is that there haven't been any programs yet that handle ethics very well.

Barbara *We have heard some presentations in the last few days saying that generative AI learns like children by observing, interacting, and adapting to the environment. Are you comfortable with that from an ethical point of view? Do you have enough confidence in the ethical behavior of humans to be good teachers for AI, or could that be worrisome?*

Eva I believe that we are generally quite good at teaching children ethical behavior. We invest a lot in this, either by being good role models or by telling children what's right and wrong. However, this approach can break down in larger contexts with many people interacting and various social pressures pulling in different directions. We see this all the time when people start wars or treat others badly. So, it doesn't work as well on a societal level as one might wish. The question then is, can we train an individual AI system to work within ethical norms? We should be able to do this to some extent. But if the system is trained, we can never be sure that it will follow the rule in some unforeseen situation. So there is a limitation there for sure. Another question, similar to humans messing things up when we act in larger groups, is what the effects would be if we had many systems interacting with many people on a larger scale. I haven't thought about this question before, but I find it very interesting. In these cases, I would be more worried about potential negative effects, even if we manage to train the systems well with respect to ethical norms.

> "When you consider AI as emulating human intelligence, I don't think there is a fundamental limit to the systems achieving all the intelligent capabilities that we possess."

Barbara *Regarding the technical capabilities AI might have in the future: On a scale from 1 to 10, where 1 describes the artificial intelligence systems we know like ChatGPT, and 10 stands for artificial general intelligence systems that surpass human capabilities on a global scale. What do you think will be possible in the future?*

Eva That's hard to say. I must admit, I feel quite modest about my ability to predict these things. But, when you consider AI as emulating human intelligence, I don't think there is a fundamental limit to the systems achieving all the intelligent capabilities that we possess. My view of humans and how we acquire our capabilities is very naturalistic. So why shouldn't it be possible, in principle, for an artificial system to have the same capabilities? Another question, for which I have no good answer, but I believe is very relevant here, is whether these capabilities are grounded only in the causal relations and functions being computed in the brain, or whether they also involve the biological substrate [3]. If it's something about our biology, then obviously AI systems have some limitations. But if it's all about the functions, then I think it should be possible to transfer everything completely to these silicon-based systems. That's the question in principle. The other question is whether we will direct research in the way to achieve all the results that I think we could, in principle, achieve. We've been making significant progress, so it seems likely that we will do that, even in my lifetime. But after that, I'm not sure.

Barbara *If you had to choose a number from 1 to 10, what would you choose?*

Eva Assigning a number is difficult. Let me mention one more limitation. I'm not sure if we can truly have consciousness in AI systems. We're not talking about intelligence as such, but will these systems be able to feel pain when we step on their toes? Will they really be able to perceive colors like a bright pink and have the same pink experience that I have when I look at a pink blotch? I'm not sure about that. So, I'm more cautious if we consider that. And then the question is, does intelligence in any dimension rely on this consciousness aspect? For example, John Searle, in his paper where he presents the Chinese room thought experiment, believes that understanding, as an aspect of intelligence, is somehow tied to a conscious perspective [10]. If we consider that, then I'm more cautious. And then, between all these things, what's the number? I'll go with five, just to stay in the middle, but it's mostly because I think it's a very hard question to answer.

Barbara *It seems that one aspect that makes it really difficult to answer this question is the fact that we don't have a clear definition of what makes us human. So it's not like we have a list of criteria that basically defines different levels of capabilities, for example, divided into three categories: Level 1 describes everything below human capabilities, Level 2 describes everything in the range of human capabilities, and Level 3 describes everything beyond human capabilities.*

Eva That's right. One of the problems is that we don't have a clear enough understanding of some of the relevant capabilities. If we don't really know what is needed, then it's harder to say whether we will be able to build it. I would say that as far as having some sort of observable generally intelligent behavior, like being able to handle different problems and tasks, it seems clear that we will reach that. But to the extent that we want real understanding and a real mind, I think that might all be tied to consciousness, to having a subjective perspective on the world tied to phenomenal experiences. There, I'm more skeptical about whether we can build this with the tools that we have for AI. But who knows? The future may show us, or maybe we won't even be able to tell.

Barbara *How important is it for AI tools to actually feel pain or perceive color? Is it necessary or important for AI tools to have the same experiences as humans? For example, they might not feel pain, but if they stub their toe, they might have the right reaction and say that it hurts. They may not feel emotions like sadness or happiness, but they will act them out perfectly at the right moments. So does it matter to the human interacting with the AI whether the AI actually feels the* *pain or emotion, or whether it is just acting the "right" way to feel natural in the interaction?*

Eva What you're talking about, even though I don't know if you've heard about it under this name, is the concept of a philosophical zombie. A philosophical zombie is someone who, for example, would be my perfect twin, who looks like me, acts like me at all times, and also has the same functional states that I have. But this twin lacks the phenomenal experiences of seeing this bright pink or feeling pain when stepping on a rock [2]. So one could say maybe that everything those AI systems could ever be are philosophical zombies, in the sense that they have all the behavior and all the functional states of someone who has phenomenal consciousness or maybe even of someone who has a subjective perspective tied to those phenomenal experiences. Then the question is, does it matter if we reach the real thing or not? I would say on one side it doesn't matter, because if they have just the zombie status, I think that is good enough for them to change our society in many ways. Imagine that we interact with them, say, on the internet and they have this as-if character. This may help them to manipulate us in many ways. So that will make a big difference already. But another interesting aspect, which I find very difficult to say something definitive about, is how we should treat them morally. Do they count morally like people do, or maybe at least like some of the higher animals do, or do they not count morally at all? I think if they have no consciousness, no subjective perspective, we can disregard them morally and use them for whatever we want. But if they were to become conscious, I would say that's the point where we really have to think about how we respect their rights, or how we avoid causing harm to them. I find that a very scary prospect, actually, that we might get there. Because we use them as our personal slaves at this point. But it would be really wrong and evil to do that if they were actual conscious beings [6]. So we would have to completely change our behavior towards those systems, and I worry that we wouldn't. So that would be really horrible, I think.

> "[...] the business landscape or the market should be organized to [...] support individual companies in not just doing some ethics-washing, but really having business models that respect ethical constraints."

Barbara *So it would be important to develop tests that allow us to distinguish real consciousness from well-faked behavior in order to set the right moral framework?*

Eva Indeed, that's the next step. How do we differentiate this? Some philosophers express concern that we may not be able to discern whether a certain system has achieved consciousness or not. This uncertainty could already be a problem. When should we start treating these systems as morally significant, just as a matter of caution?

Barbara *One of the talks today was about the extent to which we should allow this human-like interaction with artificial intelligence. Should AI tools be allowed* *to act human-like? As soon as people chat and talk a lot with AI tools, it becomes increasingly difficult for them to distinguish between humans and AI, and they lose a potentially critical distance to these tools. Do you think this is something we should be thinking about more? Should we steer development in a direction that ensures that humans remain skeptical?*

Eva Absolutely. We need to consider what benefits individual humans and our societies overall. I think, we must examine all the different contexts in which AI applications interact with us in a human-like manner. Is this beneficial? Is it harmless? Or is it detrimental that they're used in this way? We need to consider how these AI systems or bots might manipulate people on a large scale, potentially undermining our democracies or influencing voting outcomes. We need to think about how best to regulate this. One possible solution could be to enforce a rule that all bots must be clearly labeled as such. Another important point is how readily available non-human interaction partners might undermine people's ability to interact with other people. This is an important aspect of human well-being. If that is compromised, it could lead to a poorer quality of life. There are studies showing a correlation between increased online interaction and a higher prevalence of depression among teenagers. This is one area where this could be problematic. However, there may be many other areas where having these kinds of bots is completely harmless. For example, in customer service. It might also be beneficial to have easily accessible, low-key interaction partners in certain social contexts. We need to examine how such technologies truly impact people.

"But if [AIs] were to become conscious, I would say that's the point where we really have to think about how we respect their rights, or how we avoid causing harm to them."

Barbara *In light of all this, there are many different visions of the future being discussed, from utopia to dystopia. Where would you position yourself?*

Eva I would place myself somewhere in the middle. I don't believe AI will solve all our problems, nor do I think it will drastically disrupt our society in a negative way. Some existing problems might grow, but I also believe we have the power to steer things in a positive direction. We need to make small changes to the current processes, and those of us in influential positions should try to do that.

Barbara *Reflecting on the last few days here at AISoLA, was there an insight from another discipline that was particularly interesting to you?*

Eva There were so many insights, it's hard to choose just one. However, one thing that struck me was the importance of empirical studies, particularly from psychology, in understanding how explanations may or may not influence people's judgments of a situation. It's very helpful for philosophers to engage with psychologists to determine what we should research and how the results of that research should influence our thinking.

Barbara *Is there a specific topic or research question where you see an important angle for interdisciplinary work?*

Eva One interesting area could be understanding an AI system or its output. I propose in my paper that the stakes of a situation impact someone's understanding. It would be great to collaborate with psychologists to study whether people's perceived understanding of why a system produces a certain output is affected by the stakes of a situation. This conference has shown the need for perspectives from all social sciences. AI systems are increasingly influencing our lives, and it's crucial to consider all the different aspects if we want to guide the direction of these changes.

Barbara *At this conference, we had a mix of computer scientists, psychologists, legal experts, and philosophers. Are there other disciplines that should get involved in the future?*

Eva Definitely. We are missing political scientists and sociologists.

Barbara *What about business or management people?*

Eva Yes, business people too. I haven't worked much with them, but I'd be interested in learning more about potential intersections. Political scientists could help us understand how politics and our democracy might be impacted by the use of AI systems. Sociologists could provide insights into how certain areas of society are affected.

Barbara *I've been thinking about the rapid pace at which companies like OpenAI and Microsoft are advancing AI and releasing all these innovative tools. It's fascinating to think about the motivations behind such rapid progress. If the focus is solely on making quick profit, we will have problems along the way. We need to explore alternative business approaches, innovative business models and strategies that not only bring short-term benefits to customers, but also benefit society. It's important to combine profitability with ethical practices and doing the right thing. And this requires business experts who can translate social benefits into long-term business strategies.*

Eva That makes complete sense to me. I mean, one of the issues that I see is a lot of power being concentrated in the hands of just a few businesses. It makes sense to think about how the business landscape or the market should be organized to get better results and how we can support individual companies in not just doing some ethics-washing, but really having business models that respect ethical constraints. So it seems like a very good idea to have people working in this field included in this kind of discussion.

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

Eva My vision is that AI development should not be a matter of throwing new tools at society and seeing what happens. Instead, we should consider the interests of all societal groups and determine our goals as a society. Then we should think about how we can use AI tools to improve our lives. This approach is not being followed at the moment. For example, as a university teacher, I suddenly had to deal with ChatGPT and figure out how to ensure that my students still gain a deep understanding of philosophical issues when they could easily fake many tests using tools like ChatGPT. We should first identify our needs and then design tools to meet those needs.

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

Eva I believe it's extremely helpful to interact with people from different backgrounds. The questions I'm interested in cannot be answered by philosophers, computer scientists, or psychologists alone. Interdisciplinary exchange is really valuable in this area.

Barbara *Thank you very much, Eva, for your time and your thoughts on AI from a philosophical perspective. Have a great last couple of days at AISoLA!*

Eva Thank you for interviewing me!

References

1. Annette Baier (1986). Trust and antitrust. Ethics 96 (2): 231-260.
2. David J. Chalmers (1996). The Conscious Mind: In Search of a Fundamental Theory, New York and Oxford: Oxford University Press.
3. Chalmers, David J. (2023). Could a large language model be conscious? Boston Review, August 9, 2023, `https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/`.
4. EU High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI, `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai` (2019)
5. Hans-Johann Glock and Eva Schmidt (2021). Pluralism About Practical Reasons and Reason Explanations. Philosophical Explorations, 1-18, `https://doi.org/10.1080/13869795.2021.19085`
6. David J. Gunkel and Joanna J. Bryson (2014). Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient. Philosophy and Technology 27, 5–8. `https://doi.org/10.1007/s13347-014-0151-1.`
7. Markus Langer, Daniel Oster, Timo Speith, Lena Kästner, Kevin Baum, Holger Hermanns, Eva Schmidt and Andreas Sesing (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296, 103473, doi: 10.1016/j.artint.2021.103473.
8. Eva Schmidt (2022). Wie können wir autonomen KI-Systemen vertrauen? Die Rolle von Gründe-Erklärungen, in Britta Konz, Karl-Heinrich Ostmeyer und Marcel Scholz (eds.) Gratwanderung Künstliche Intelligenz – Interdisziplinäre Perspektiven auf das Verhältnis von Mensch und KI. Stuttgart: Kohlhammer, 11-29.
9. Eva Schmidt (forthcoming). Stakes and Understanding the Decisions of AI Systems, in Juan Durán and Giorgia Pozzi (eds.) Philosophy of Science for Machine Learning: Core Issues, New Perspective, Synthese Library.
10. John R. Searle (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences 3 (3), 417-424.