Let's Talk AI with Timo Speith

Timo Speith[1] and Barbara Steffen[2]

[1] University of Bayreuth, Department of Philosophy, Germany,
timo.speith@uni-bayreuth.de
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"It shouldn't be the AI of big tech companies, as it is currently emerging, but rather AI designed for society, and perhaps even by society, through participative approaches, community work, and citizen science."

The Interviewee - Timo Speith

My Personal AI Mission:
Making AI comprehensible and beneficial to every stakeholder.

My Takes on AI

Artificial Intelligence: The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages [11].

Trust: An attitude of a person towards the reliability/functionality of a certain entity. Trust must be distinguished from trustworthiness: the actual reliability/functionality of the entity. Ideally, trust is based on trustworthiness; however, there are many contingent factors that influence trust [6].

Explainability: Providing information about certain aspects of an entity (for example, the decision-making mechanisms of an AI) to better understand that aspect (see also [3] for a definition).

Essential Elements of Human Capabilities: Approximately those mentioned in the above definition of artificial intelligence, supplemented by adaptivity to new situations and the generation of new ideas.

The Interview

Barbara *Welcome Dr. Timo Speith to this interview. Please briefly introduce yourself and your relationship to artificial intelligence (AI).*

Timo Thank you for the invitation! My name is Timo Speith, and I am a fixed-term lecturer at the chair for philosophy, computer science, and AI at the University of Bayreuth. My background encompasses both computer science and philosophy, having pursued various studies in both fields, including a bachelor's degree and a PhD in philosophy, as well as a master's degree in computer science. This places me at the very intersection of these disciplines. In addition to my academic background, my primary research focus is closely related to computer science and AI; specifically, it is in the area of explainable AI (XAI). This represents a significant relationship to AI for me. Moreover, I am generally interested in almost every topic that lies at the intersection of computer science, AI, and philosophy.

Barbara *Can you give an example or two of specific research questions you're currently addressing with your AI research?*

Timo Certainly. Within the realm of XAI, a key question involves understanding how providing explanations of AI systems' predictions and decision-making processes can be beneficial to various parties involved with AI [9, 3, 12]. This includes elucidating the 'what' and 'why' behind an AI's predictions, such as the rationale for its outputs. Such insights are hoped to assist technicians in debugging AI systems to identify errors or misbehaviors [7, 10, 3]. Additionally, it aims to empower laypersons to assess the appropriateness of AI-based predictions, particularly in cases where the predictions may seem potentially discriminatory or influenced by irrelevant details [1, 10, 3]. Overall, this encompasses a broad area of interest, touching on both technical and ethical considerations.

"The type of explanation required varies significantly depending on the stakeholder."

Barbara *In terms of explanations, do you focus on explanations for experts who need to understand how the system works to determine its trustworthiness, or is it for users to get explanations that allow them to decide whether they personally trust the system and/or the output?*

Timo My interest spans both these aspects and extends even further to encompass all stakeholders. One of my core research theses, and a focus of the project I'm working on, is that the type of explanation required varies significantly depending on the stakeholder [9, 12]. For example, a layperson requires a different explanation than a developer, regulator, or a decision-maker, such as a hospital manager deliberating the use of a specific AI diagnostic system. The explanation needs of stakeholders change based on their role and many other aspects, highlighting the importance of tailoring explanations to suit diverse needs and perspectives.

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

Timo The emphasis on trust can be misleading, in my opinion [6]. People often trust a system for irrelevant reasons. For example, studies have shown that merely providing explanations can increase people's trust in a system, even if these explanations don't actually offer any real insight into how the system works [5]. So, the mere presence of an explanation can lead people to trust a system more. Unfortunately, trust, as subjective and fleeting as it is, is often seen as a prerequisite to even consider using systems. However, as just mentioned, trust is a very subjective attitude towards an entity, and thus not reliable. To directly answer your question, trust is indeed important for AI adoption. Yet, from a philosophical standpoint, this is unfortunate, as the focus shouldn't be on the fact that trust exists, but rather on ensuring that trust is based on the right reasons. Ideally, such a justified trust should be the foundation upon which people decide to adopt AI. However, considering the mismatch between the reasons people trust a system and what would be needed for justified trust, it's clear there's more complexity to the issue.

Barbara *And what measures are essential to ensure the ethical use of AI?*

Timo That's actually another focus of my research. I'm deeply interested in machine ethics and ethical AI. In my opinion, there isn't a one-size-fits-all solution for achieving ethical AI. It essentially depends on various characteristics you'd want an AI system to embody, which, in turn, are contingent upon the context in which the AI system is deployed. For example, the accuracy of a song recommendation algorithm might not be critically important to me as a user. While the company behind the algorithm might prioritize its accuracy for reasons of reputation, and as a user, I might be slightly inconvenienced by an unappealing song suggestion, an incorrect song recommendation is not a significant ethical concern. However, particularly in high-stakes scenarios, the situation changes dramatically [2]. Aspects such as fairness, robustness, explainability, and high accuracy become crucial in such scenarios. Justified trust also plays a significant role within the ethical framework, but even more so does the trustworthiness of the system—its ability to function as intended. In any case, addressing this question isn't straightforward due to the multifaceted nature of AI and ethics. It's challenging to single out one characteristic as the definitive criterion for an AI system's ethical use, as it greatly depends on the application context and adopted ethical view.

"The focus shouldn't be on the fact that trust exists, but rather on ensuring that trust is based on the right reasons."

Barbara *Is there already some kind of framework that describes application scenarios from low to high stakes and lists the different criteria that need to be met to be considered sufficiently ethical or sufficiently reliable and trustworthy?*

Timo As of now, I'm not aware of any such framework. I also question its feasibility. However, looking at legislation, the AI Act does attempt to adopt a

risk-based approach. It categorizes different levels of risk, each with its corresponding obligations. Nevertheless, it's important to note that law and ethics are distinct fields. An ethical framework might therefore take a different shape. Context is crucial, and it might be necessary to evaluate each system or use case individually.

Barbara *Moving on to the next question, what do you think the technical capabilities of AI will be in the future? If we look at a scale from one, which describes the artificial intelligence systems we see today like ChatGPT, to ten, which describes artificial general intelligence that surpasses human capabilities. What do you think will be possible?*

Timo I consider myself somewhat of a tech optimist, so I'd say potential technical capabilities are closer to the ten end of the scale—perhaps an eight or nine. Philosophically, it's a challenging question. The current debates often extend beyond artificial general intelligence to superintelligence. If the question were about superintelligence, I'd be skeptical about its realization.

Barbara *How do you distinguish artificial general intelligence from superintelligence?*

Timo Artificial general intelligence usually refers to a single AI system that can perform multiple tasks traditionally performed by humans at a human or superhuman level, such as playing chess, generating images, or detecting cancer. On the other hand, the term superintelligence often refers to an AI system with consciousness whose intelligence far exceeds that of humans. This should suffice as a detour; to address your original question, we're already witnessing significant advancements of AI systems. For instance, chatbots are being utilized for various purposes, with some people even suggesting their capabilities in certain areas surpassing those of humans. In the field of medical AI, there are numerous instances where AI has been recognized as more proficient than highly skilled doctors in diagnosing cancer. So, in some respects, we're already there.

"There isn't a one-size-fits-all solution for achieving ethical AI."

Barbara *And how do you see AI and its impact on the future? Today we hear all kinds of future scenarios, from dystopian nightmares to utopian dreams. What is your view?*

Timo As I've mentioned, I'm a tech optimist. Despite the negative impact that AI might have on social media and, by extension, society, I'm encouraged by the research community's efforts to address these issues. Furthermore, with legislative measures like the forthcoming AI Act in the EU, along with the Digital Services Act and other initiatives, I believe we're moving in the right direction. There's undoubtedly a lot of destructive potential, but the path forward looks promising.

Barbara *Destructive potential from artificial intelligence applications themselves or from the intentions and actions of actors?*

Timo Well, when you think about it, if you possess a tool—take a hammer, for example—you can use it to drive in a nail or to commit a crime. The core issue often lies with the individuals wielding the tool. However, AI introduces unique challenges, such as training biases, stereotypes, and the perpetuation of historical biases, which pushes the hammer analogy to its limits. However, as I've mentioned, I hold a strong belief in the research community and in legislative bodies. There are individuals deeply concerned with these issues, actively working to address them, and I believe their efforts are, to some extent, successful.

Barbara *Reflecting on the last few days of this interdisciplinary conference, what was the most interesting insight for you?*

Timo I've always considered myself an interdisciplinary individual, drawing significant insight from psychology, among other fields. Philosophers often engage in discussions about how the world ought to be, crafting normative claims that attempt to outline how reality should be or how people should perceive various aspects of life. However, reality often diverges from these philosophical ideals, a fact that becomes particularly evident through psychological studies. From a philosophical perspective, it seems logical to argue that people should desire explanations and benefit from them. Yet, empirical studies frequently reveal that individuals may not actually seek explanations. For example, it has been observed that after receiving an explanation, people's perception of a system does not necessarily improve, it can even deteriorate [8]. This is because the explanation unveils the factors considered in the decision-making process of an AI or any system, leading to a more critical view of it. Philosophically, one might argue that understanding how a system operates should enhance our perception of it, as it provides a more justified belief about the system's functionality. This discrepancy between philosophical expectations and psychological findings is always fascinating to me.

"People often trust a system for irrelevant reasons. [...] studies have shown that merely providing explanations can increase people's trust in a system, even if these explanations don't actually offer any real insight [...]."

Barbara *Is there a specific research question you would like to see addressed from an interdisciplinary perspective?*

Timo Admittedly, there are numerous questions, primarily centered around my research interests. I'm particularly fascinated by the XAI debate, delving into the nuances of which systems require certain types of explanations for specific purposes. This is so that users can achieve their objectives with these explanations, feel satisfied with the system, and use it correctly. Basically, if a system is faulty, users should be able to find out and not use it. Lately, my focus has shifted slightly towards understanding how explanations and fairness intersect.

Barbara *Fairness itself seems to be a difficult concept because it is very subjective. Are there any agreed-upon definitions, or do you combine different ones and integrate them into the system?*

Timo That's precisely what captivates me about this debate: depending on your understanding of fairness, explainability can serve different roles. In some cases, it can directly contribute to fairness; in others, it acts merely as a debugging tool for fairness, and sometimes, it may not aid in achieving fairness at all [4]. What intrigues me is the challenge of unraveling these concepts of fairness and explainability and then attempting to connect them.

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

Timo That's a challenging question, to be honest. Basically, it should be AI for the people.

Barbara *People in terms of society?*

Timo Yes, it shouldn't be the AI of big tech companies, as it is currently emerging, but rather AI designed for society, and perhaps even by society, through participative approaches, community work, and citizen science.

Barbara *Which disciplines are you already collaborating with? And are there other disciplines that should get involved in the future?*

Timo I'm involved in a project where we collaborate with legal scholars, psychologists, computer scientists, and philosophers. I've also worked with political scientists. However, I would also be interested in incorporating the viewpoint of sociology to gain a broader perspective on AI.

Barbara *The progress in AI in the last year has been crazy. And that progress is being driven primarily by companies because of the competition in the industry. Companies are rushing to make sure that the competition is not faster, that they are not losing potential users, and that they are not running the risk of shrinking their networks. How do you see the role of business professionals in this? For example, finding new ways to move from today's money-driven perspective to a society-driven perspective, and finding new incentives and benefits for companies to slow down the current pace of AI progress? Would that be interesting?*

Timo I'm not sure it's necessary. In a way, this competitive pressure also impacts governments, prompting them to increase funding for research projects that don't involve private companies. So, I believe there might even be a benefit to this kind of AI race.

Barbara *Could you elaborate on that?*

Timo Increasing pressure on governments from corporate competition can lead to more funding for research and progress in areas such as AI legislation. For example, during our discussions this week, we noted how law tends to lag behind technological advancements. However, initiatives like the European AI Act are promising and show an opposite trend. They aim to create legislation that is

broad enough to accommodate the rapid changes in the AI landscape, including future shifts in infrastructure and the types of AI being developed.

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

Timo Thank you for organizing the conference and for striving towards greater interdisciplinarity.

Barbara *Thank you, Timo, for your time and your interdisciplinary perspective from computer science and philosophy. Enjoy the last days at AISoLA!*

Timo Thank you!

References

1. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Banjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115. DOI: 10.1016/j.inffus.2019.12.012
2. Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. Philosophy & Technology, 35(1), 12. DOI: 10.1007/s13347-022-00510-w
3. Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: a definition, a model, and a knowledge catalogue. In 2021 IEEE 29th international requirements engineering conference (RE) (pp. 197-208). IEEE. DOI: 10.1109/RE51729.2021.00025
4. Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A Critical Survey on Fairness Benefits of XAI. arXiv preprint arXiv:2310.13007.
5. Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. In Extended abstracts of the 2019 CHI conference on human factors in computing systems (pp. 1-6). DOI: 10.1145/3290607.3312787
6. Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW) (pp. 169-175). IEEE.
7. Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a non-functional requirement. In 2019 IEEE 27th International Requirements Engineering Conference (RE) (pp. 363-368). IEEE. DOI: 10.1109/REW53955.2021.00031
8. Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. Computers in Human Behavior, 81, 19-30. DOI: 10.1016/j.chb.2017.11.036
9. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021a). 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. DOI: https://doi.org/10.1016/j.artint.2021.103473

10. Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021b). Explainability auditing for intelligent systems: a rationale for multi-disciplinary perspectives. In 2021 IEEE 29th international requirements engineering conference workshops (REW) (pp. 164-168). IEEE. DOI: 10.1109/REW53955.2021.00030
11. Oxford Reference (2024, March 15). Artificial Intelligence. https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095426960
12. Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2239-2250). DOI: 10.1145/3531146.3534639