Let's Talk AI with Nicole Krämer

Nicole Krämer[1,2] and Barbara Steffen[3]

[1] University Duisburg-Essen, Department of Human-Centered Computing and Cognitive Science, Germany,
[2] Research Center Trustworthy Data Science and Security, Germany,
nicole.kraemer@uni-due.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"We need to better understand how humans interact with artificial intelligence."

The Interviewee - Nicole Krämer

My Personal AI Mission:
To contribute to a better understanding of the mechanisms when humans and AI work together.

My Takes on AI

Artificial Intelligence: Systems capable of automated intelligent behavior or decisions, based on, for example, machine learning.

Trust: Specifically calibrated trust is important: The degree to which the abilities of the system match the trustworthiness the human user perceives.

Explainability: The degree to which the system is able to communicate the basic functioning of its algorithms.

Essential Elements of Human Capabilities: All human abilities are important (like when interacting with a fellow human): perception, cognition, emotion, behavior.

The Interview

Barbara *I have the pleasure to interview Professor Nicole Krämer. Please briefly introduce yourself and your relation to artificial intelligence.*

Nicole My name is Nicole Krämer. I'm a professor of Social Psychology, Media, and Communication at the University Duisburg-Essen. I'm also a member of the scientific board of the newly established Research Center "Trustworthy Data Science and Security". I've been working in the field of human technology interaction for over 20 years. In the last five years, I have specifically focused on how humans interact with artificial intelligence from a social psychological perspective.

"That it talks makes people think it's human-like, so they immediately apply all their social notions to it and feel that it functions just like a human."

Barbara *Do you have one or two examples of research questions you are currently working on?*

Nicole I'm particularly interested in understanding how humans develop trust in artificial intelligence and how we can communicate to them whether they can trust or distrust a system. I'm also keen on exploring the relationship between understanding and trust. Specifically, I'm interested in whether people need to understand how an AI works to trust it, or if understanding doesn't play a role at all. How can we communicate trust to people?

Barbara *What is your definition of understanding in this context?*

Nicole The XAI community uses explainable AI and tries to find methods to explain what's happening in an algorithm or a system that's built on machine learning [8]. This explainability is often employed to help experts better understand how the algorithm works [3, 5]. However, I'm more interested in how we can teach laypeople about a system's functionality and how it works.

Barbara *What do you mean by explanation? For example, should users take a tutorial before using ChatGPT for the first time to get a better understanding of how it works? Or does it rather refer to specific answers ChatGPT gives to my prompts, for which I could or should get an explanation as to why I got that particular answer?*

Nicole In the explainable artificial intelligence community, or what psychologists do when they are part of that community, they try to provide some sentences about what the system can or cannot do, how it was trained [6]. So for example, for ChatGPT, it would be some aspect that you need to understand, such as that this is a statistical method used to predict the next most probable word, and with large databases, it can produce sentences that look just like a human would write. Most people probably do not know much about this.

Barbara *Which brings me to my next question. What is the role of trust in the adoption of AI?*

Nicole That's a good question. It's challenging to answer because it's so broad. We believe that in many contexts, whether I can trust a system or not will determine if I decide to use it. There will be instances where I don't have a choice and might not even be aware that there is a lot of artificial intelligence in place. For example, Instagram, Facebook, TikTok, they all have artificial intelligence implemented seamlessly. So people often don't know how much artificial intelligence is really there. They can't really make a choice in some of these systems. In others, they will need to make a conscious decision of whether to use the system or not, like with clinical support systems [1].

"[The way] to achieve calibrated trust is to have meetings like this one, to talk to each other, to come up with measurements, both on the side of the technology and measurements of human trust."

Barbara *Are there any essential measures you have in mind to ensure the ethical adoption of AI?*

Nicole In terms of trust, we need to ensure, and it's ethically desirable, to employ artificial intelligence in systems only when we can guarantee that they are trustworthy. The most unethical aspect would be to deploy systems that people trust but are not genuinely trustworthy, leading to overtrust. We strive for people to trust a system to the extent that the system truly deserves that trust – which is what we call calibrated trust [7]. People should not have excessive trust in a system when it is independent of its actual capabilities.

Barbara *Do you think there are standard approaches to calibrated trust, or does it depend on the individual and their base level?*

Nicole Hopefully, it won't need to depend on the individual because then we would struggle to implement calibrated trust in a system. What we hope for is that for any given system, we can find measurements of how trustworthy the system is. That's the first problem that needs to be solved, and it's not trivial. We need to check how reliable, how trustworthy the system is by, for example, formal verification guarantees, uncertainty measurements.

Barbara *In terms of the technical capabilities of artificial intelligence, what do you think will be possible in the future?*

Nicole I have stopped making predictions because if you had asked me two weeks before ChatGPT appeared in our lives, I would have told every journalist that I don't believe that such a thing as an AI you can really talk to on a dialogue basis will be available soon. I would have said that's 50 years away, or even more, let's make it 100.

Barbara *We see a lot of speculation about possible futures now that AI has entered our lives. These range from dystopia to utopia. Where would you place yourself?*

Nicole As an empirical researcher in this area, I tend to avoid personal feelings about whether this is positive or negative for humankind. Instead, I am trained to look at data to see what kind of positive and negative effects on humans we can observe.

Barbara *Looking back on the last few days, in particular on the interdisciplinary sessions and your interdisciplinary work in general, what are some of the most interesting insights you have gained?*

Nicole Even though I've been working with computer scientists for 20 years now, I learned new things about what's possible and saw many things where I could immediately say, "Wow, that's interesting also from a psychological point of view."

Barbara *How does your interdisciplinary collaboration work? Do the computer scientists explain the systems, how they work and how trustworthy they are, so that you can look at it from a psychological perspective and design measures to make sure that actual trust and perceived trust are properly calibrated? This way ensuring that users demonstrate the right level of trust given the underlying technology and the output of the system?*

Nicole That's an interesting question and already describes our approach well. This is why we are here, to make progress on these complex questions. To be honest, I don't yet know what we will know in three years' time, but the only chance we have to make progress on these questions of how to achieve calibrated trust is to have meetings like this one, to talk to each other, to come up with measurements, both on the side of the technology and measurements of human trust, and then talk to each other to establish this connection and balance it.

Barbara *What is your goal for calibrated trust? Some kind of shared understanding or framework that helps you to properly translate between different disciplines and perspectives?*

Nicole Frameworks are always helpful. In the end, for practical purposes, we want to ensure that even laypeople have the chance to judge how reliable a system really is - unliketoday, where large companies do massive field studies by launching things like ChatGPT on the market and having people use it without having tested for any kind of trustworthiness. Especially when ChatGPT appeared, people immediately trusted the system more than it deserved because it was so good at dialogue management, until it became obvious that it hallucinates and reports wrong facts (for an early study on the degree of trust users put in ChatGPT, see [2]).

"I doubt that companies will act in terms of the greater good or can be incentivized except with money."

Barbara *What do you think of ChatGPT? For example, how it engages in discussions about empathy, feelings and emotions?*

Nicole Well, systems like ChatGPT have a lot of social cues. The fact alone that it talks makes people think it's human-like, so they immediately apply all their social notions to it and feel that it functions just like a human. This is a mechanism we have known for years and that has already been described in the "computers are social actors" paradigm [4].

Barbara *Is there a specific research question that you would like to see more interdisciplinary research focused on? And which disciplines should be involved?*

Nicole I think the questions are sufficiently complex, so I wouldn't add any more questions. In terms of disciplines that need to be involved, I have had very positive experiences when computer scientists and psychologists work together. However, people from ethics should definitely also be part of this to better reflect on these normative aspects, and people from law who can help regulate certain aspects.

"My vision of AI [is] to be helpful [rather than taking] people's data to make more money out of it."

Barbara *Do you think that there could also be incentives for companies to behave in a more desirable way, or do you think that this can only be achieved through regulation, for example by restricting certain behaviors and progress?*

Nicole My husband always says, "Well, it's all about money in that world," so I doubt that companies will act in terms of the greater good or can be incentivized except with money. But who wants to spend money to make systems more trustworthy or more honestly communicate how trustworthy the system really is? I don't think that there will be someone trying to regulate it this way, so we probably need regulation.

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

Nicole The AI vision. Well, in the end, AI can and should help people make better decisions and solve tasks faster and easier. There can be great assistance from AI, but it needs to be for the people and not patronize people or rely on people's data compromising their privacy. So, all these negative side effects that we currently have should be avoided. That would be my vision of AI, to be supportive to people, to be helpful, and not just take people's data to make more money out of it.

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

Nicole Nothing comes to my mind right now.

Barbara *Nicole, thank you very much for your time and the psychological perspective on AI. Have a great evening!*

Nicole Thank you very much. That was exhausting.

References

1. Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for clinical practitioners - Saliency- based XAI approaches. European Journal of Radiology, 162, 110787. https://doi.org/10.1016/j.ejrad.2023.110787
2. Choudhury, A. & Shamszare, H. (2023). Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. Journal of Medical Internet Research, 25:e47184. https://doi.org/10.2196/47184
3. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Comput. Surv. 55, 9, Article 194. https://doi.org/10.1145/3561048
4. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153
5. Saeed, W. & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, C. https://doi.org/10.1016/j.knosys.2023.110273
6. Szczuka, J., Horstmann, A., Mavrina, L., Artelt, A., Hammer, B., & Krämer, N. C. (2024, accepted). Let Me Explain What I Did or What I Would Have Done: An Empirical Study on the Effects of Explanations and Person-Likeness on Trust in and Understanding of Algorithms. NordiCHI 2024.
7. Wischnewski, M., Krämer, N. C. & Müller, E. (2023). Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 755, 1–16. https://doi.org/10.1145/3544548.3581197
8. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J. Tang, M.Y. Kan, D. Zhao, S. Li, H. Zan (eds), Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture