Let's Talk AI with Daniel Neider

Daniel Neider[1,2] and Barbara Steffen[3]

[1] TU Dortmund University, Department of Computer Science, Germany,
[2] Center for Trustworthy Data Science and Security, UA Ruhr, Germany,
daniel.neider@cs.tu-dortmund.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"AI will increasingly impact our future. Let's work together to make it safe and bright."

The Interviewee - Daniel Neider

My Personal AI Mission:
My mission is to advance the field of artificial intelligence (AI) by developing novel machine learning techniques and formal methods that ensure the correctness, security, and trustworthiness of AI systems. By doing so, I hope to contribute to a future where AI is widely adopted and benefits society without compromising safety, privacy, or ethical values.

My Takes on AI

Artificial Intelligence: AI is an umbrella term for machines – usually computer systems – that mimic human intelligence.

Trust: I like the definition of [7]: Trust is the willingness of a party [the trustor] to be vulnerable to the actions of another party [the trustee].

Explainability: Explainability is the challenge to empower humans to understand the decision making of AI.

Essential Elements of Human Capabilities: Empathy.

The Interview

Barbara *Welcome, Professor Daniel Neider from the Technical University of Dortmund. Thank you for joining me for this interview. Please introduce yourself and your relationship to artificial intelligence.*

Daniel First, I want to thank you for having me. I'm Daniel Neider, a professor at TU Dortmund University and the Center for Trustworthy Data Science and Security specializing in formal guarantees of machine learning. My team focuses on making artificial intelligence safer, more reliable, and more trustworthy.

Barbara *Can you name one or two AI-related research questions that you are currently working on?*

Daniel We are working on formally proving practically relevant properties of neural networks, such as robustness and fairness [4, 3, 5]. Moreover, we are investigating what crucial properties neural networks must satisfy regarding safety and reliability so that they can be used safely in the real world.

Barbara *What do you mean by fair?*

Daniel That's an excellent question, as there exist numerous definitions of fairness [1]. We do not view it as our primary research to devise these definitions. However, once formalized, we can automatically check neural networks against them to determine whether these properties are satisfied.

Barbara *Can you do this even if you don't know what the correct result is? In some cases, for example, you just want to distinguish dogs from cats, which makes it very easy for humans to evaluate the results. But in other cases, you don't know what the correct result is, which makes the evaluation process much more complicated. What do you do in cases where the correct results are not predefined?*

Daniel It's important to note that we're interested in thoroughly checking a vast number of inputs, not just the test or training data, but ideally all possible inputs. This is, of course, a massive undertaking. Since we cannot label - and test - an infinite number of inputs, we require a formal description of the network's desired behavior. The challenge with this approach is that machine learning bypasses the problem of creating such a formal specification in the first place: we're given data and then use machine learning to find a model that captures the patterns in the data. In the end, we hope this model will do something good, but it's unclear what that means [6]. A novel trick my team devised is using other neural networks as part of the specification. That allows us to check, for instance, whether a network we are interested in performs similarly to another network we know already performs very well. We call this approach neuro-symbolic verification [11].

"My question would be: does AI have to be perfect, or is it enough if it is indistinguishable or better than humans?"

Barbara *That's interesting. What role does trust play in the adoption of AI?*

Daniel I'm not entirely convinced that trust plays a significant role at the moment. It should, but I don't know whether it actually does. If technology is sufficiently helpful and provides enough value, people might even use it without worrying too much. I fall for this myself: if technology is convenient and valuable, I rarely question it or reflect on whether I should trust it.

Barbara *The benefits are so tangible that they outweigh the doubts. Do you have any essential measures in mind to ensure the ethical use of AI?*

Daniel My first question would be: how exactly do you define the ethical use of AI? That's a question arguably best answered by philosophers.

Barbara *Do ethics play a role in AI, and who should be involved in discussing such measures?*

Daniel Yes, AI should be designed with ethics and trustworthiness in mind. Unfortunately, the current approach is that companies develop AI systems, release them, and see what happens. This is arguably not an ethical approach, and we need to change how AI technology is developed. My team can provide technical tools to this end. Still, we require societal input on precisely what these ethical considerations are to implement them.

"AI will be our future, and we have to make sure this future will be safe and bright."

Barbara *Regarding the future technical capabilities of AI on a scale of 1 to 10, where 1 stands for artificial intelligence systems like ChatGPT and 10 for artificial general intelligence systems that surpass human capabilities. What do you think will be possible?*

Daniel I don't know, but a 7 or an 8 seems likely. I am convinced we will see AI systems where the average user can't discern between humans and AI. In analogy to the Turing test [10], that might be enough. I do not see much difference between an actual AGI and an AI that convincingly acts, looks, and feels intelligent.

Barbara *So the question is not so much whether we trust, but rather whether we should trust?*

Daniel My question would be: does AI have to be perfect, or is it enough if it is indistinguishable or better than humans are?

Barbara *What is your personal view of the future? Are we moving towards a dystopia or a utopia? Where would you place yourself on this scale?*

Daniel I'm uncertain about the long term. It's probably right in the middle in the short to medium term. Some people will become very wealthy, and many will become much more productive and successful. But there will also be people who will lose their jobs and have to learn entirely new and different skill sets. It seems to me like a new "industrial revolution". We as a society need to consider what measures to take to alleviate the drastic transformation we will likely see in the next 10 to 15 years.

"If technology is convenient and valuable, I rarely question it or reflect on whether I should trust it."

Barbara *Looking at ChatGPT, do you think that users should be informed about how ChatGPT works? And how to use it correctly?*

Daniel In principle, yes. But this is not specific to ChatGPT. We should require information and transparency for any sufficiently complex system that is out there and easy to use.

Barbara *Are we doing it sufficiently? For ChatGPT? And in general?*

Daniel Probably not. However, this field moves so rapidly that it would take a lot of work to keep up with all these changes. For instance, it's difficult for me to imagine how to teach this topic in schools when changes happen with a few months. Speaking of schools, how do we deal with ChatGPT when pupils can use it to do their homework? Should they do it or not? The jury is still out on that, and I'm unsure what to recommend. However, I am optimistic and lean toward embracing the opportunities, provided that there is close supervision by the teachers.

"I am convinced we will see AI systems where the average user can't discern between humans and AI."

Barbara *Reflecting on the last few days and the various interdisciplinary presentations. Do you remember an insight that was particularly interesting to you?*

Daniel I enjoyed the legal or regulatory perspective on artificial intelligence - not for any specific reasons other than to satisfy my curiosity. I'm convinced this is where AI advancement in Europe will flourish or fail, depending on whether we are smart in regulating this technology. Hence, I found the presentations on this perspective on AI fascinating.

Barbara *Do you have a research question or a topic in mind where you would like to see more interdisciplinary collaboration in the future?*

Daniel I would like to incorporate more ethical considerations into my team's research. I have some ideas of how to do that, and collaborations with people from ethics and machine learning would be very helpful. For instance, colleagues of mine have shown how to ensure that generative AI creates images with a controllable degree of nudity or violence [9], which I find fascinating!

Barbara *Do you already have a specific research question in mind, or would you like to develop it with ethics experts to see where further collaboration would be beneficial?*

Daniel Let me give you an example. At the moment, I collaborate with colleagues from TU Darmstadt on reinforcement learning to align autonomous agents better with human ethical values. It's too early for results, but I am excited about this research direction. Unfortunately, a huge obstacle is the lack of a solid understanding or notion of the desired behavior of AI systems, as we have already discussed earlier.

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

Daniel Let me tell you what my personal vision is: that AI will become as reliable as current hardware and software systems. A burgeoning research community, including my group, has evolved around this topic, and we have already made great strides toward this goal [8]. AI will be our future, and we have to make sure this future will be safe and bright.

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

Daniel No.

Barbara *Thank you very much, Daniel, for your time and insights. Have a great day!*

Daniel Thank you for this engaging interview.

Barbara *Thank you.*

References

1. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. Fairness and Machine Learning: Limitations and Opportunities. MIT Press.
2. Böing, Benedikt, Rajarshi Roy, Emmanuel Müller, and Daniel Neider. 2020. "Quality Guarantees for Autoencoders via Unsupervised Adversarial Attacks." European Conference on Machine Learning and Knowledge Discovery in Databases - European Conference (ECML/PKDD). Springer. 206-222.
3. Bollig, Benedikt, Martin Leucker, and Daniel Neider. 2022. "A Survey of Model Learning Techniques for Recurrent Neural Networks." A Journey from Process Algebra via Timed Automata to Model Learning. Springer. 81-97.
4. Khmelnitsky, Igor, Daniel Neider, Rajarshi Roy, Xuan Xie, Benoît Barbot, Benedikt Bollig, Alain Finkel, Serge Haddad, Martin Leucker, and Lina Ye. 2022. "Analysis of recurrent neural networks via property-directed verification of surrogate models." International Journal on Software Tools for Technology Transfer (Springer) 25: 341–354.
5. Khmelnitsky, Neider, et al. 2021. "Property-Directed Verification and Robustness Certification of Recurrent Neural Networks." 19th International Conference on Automated Technology for Verification and Analysis (ATVA). Springer. 364-380.
6. Leucker, Martin. 2020. "Formal Verification of Neural Networks?" Brazilian Symposium on Formal Methods. Springer. 3-7.

7. Mayer, Roger C., James H. Davis, and F. Schoorman David. 1995. "An Integrative Model of Organizational Trust." The Academy of Management Review 20: 709-734.
8. Neider, Daniel, and Taylor T. Johnson. 2023. "Track C1: Safety Verification of Deep Neural Networks (DNNs)." First International Conference on Bridging the Gap Between AI and Reality (AISoLA). Springer. 217-224.
9. Schramowski, Patrick, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models." IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 22522-22531.
10. Turing, Alan M. 1950. "Computing Machinery and Intelligence." Mind LIX (236): 433–460.
11. Xie, Xuan, Kristian Kersting, and Daniel Neider. 2022. "Neuro-Symbolic Verification of Deep Neural Networks." 31st International Joint Conference on Artificial Intelligence. Vienna: ijcai.org. 3622-3628.