Let's Talk AI with Joost-Pieter Katoen

Joost-Pieter Katoen[1,2] and Barbara Steffen[3]

[1] RWTH Aachen University, Department of Computer Science, Germany,
[2] University of Twente, Department of Computer Science, Netherlands,
katoen@cs.rwth-aachen.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"AI has great potential but needs to be taken with care. It is of utmost importance that AI is introduced and used with common sense. AI needs to be safe, reliable and sustainable."

The Interviewee - Joost-Pieter Katoen

My Personal AI Mission:
Make AI systems more dependable and safe, in particular if these systems have to act under uncertainty

---

My Takes on AI

Artificial Intelligence: a set of methods that use learning to accomplish a certain task.

Trust: the outcomes of AI methods need to be such that they are transparent and fair.

Explainability: outcomes of AI engines such as (deep) neural networks come with a justification, e.g., a certificate that one can easily check providing understandable arguments for the given outcome.

Essential Elements of Human Capabilities: social intelligence, emotional intelligence (emotions, affection, mood, etc.), creativity, and adaptivity.

The Interview

Barbara *Welcome, Professor Joost-Pieter Katoen. Thank you for taking the time for this interview. Could you please briefly introduce yourself and your relationship to artificial intelligence?*

Joost-Pieter Certainly. I'm Joost-Pieter Katoen, a professor of software modeling and verification at RWTH Aachen University in Germany. I'm also affiliated with the University of Twente in the Netherlands. As for my connection to AI, while I'm not an AI expert, I believe that when using AI components, particularly in safety-critical systems like autonomous robots, cars, and satellites, it's vital that these AI components are reliable and do not exhibit unexpected or potentially dangerous behaviors. I strongly believe that formal verification methods can be used to analyze AI systems.

Barbara *What AI challenges are you currently addressing in your research?*

Joost-Pieter Our research primarily deals with uncertainty aspects. This has recently been identified as a crucial issue in trustworthy AI [3, 2]. A scenario that best illustrates this involves a robot that needs to navigate, for instance, finding the exit of a labyrinth while avoiding potential collisions with other robots in the room. This is a multi-robot system. The robot also needs to conserve energy, so it should reach the exit in the fewest possible steps. The question is, can we synthesize or define a planner for the robot that achieves these safety goals, reaching the exit with a minimal energy budget, without colliding with other robots? My recent talk "Facing Uncertainty in AI Systems" goes much more into detail on this [6].

"[I]t's vital that these [safety critical] AI components are reliable and do not exhibit unexpected or potentially dangerous behaviors."

Barbara *What role do you think trust plays in the adoption of AI? And what measures do you think are important to ensure the ethical use of AI?*

Joost-Pieter That's a complex question. Trust is more than just correctness. For end users to trust AI software, fairness is a crucial aspect. For instance, there was a scandal in the Netherlands a couple of years ago when tax authorities used software to identify "suspicious" individuals who might be misusing the tax system for childcare [4]. It turned out that this AI component was discriminatory, particularly against people from ethnic minorities. This scandal led to some politicians resigning. I mention this because such incidents certainly do not contribute to public trust in AI components. So, for me, trust involves correctness and fairness. People need to be able to rely on what these components suggest. And honestly, I think we're not there yet.

Barbara *In terms of the future capabilities of AI, where do you think it is headed? On a scale of 1 to 10, where 1 is the artificial intelligence systems we know today, such as ChatGPT, and 10 is artificial general intelligence, such as*

autonomous systems that surpass human capabilities. Where on that scale would you place the future capabilities of AI?

Joost-Pieter Yes, I believe it's fair to say that AI holds enormous potential. We've seen this in agents that play Go, even winning against world champion players. Go is a game that is significantly more complex than chess, so this certainly demonstrates AI's potential. However, I remain somewhat skeptical. If I were to rate my optimism on a scale from 1 to 10, I'd say I'm at a 4 or 5. I hope I'm wrong, but I feel our expectations of AI might be too high. Currently, it seems we're inclined to believe that ChatGPT or similar tools and their successors can solve all kinds of problems. For example, the belief that they can automatically generate program code, like software [1] is growing. I think we need to exercise caution here. We shouldn't always trust the output of AI without question. We need a rational approach to dealing with potential outputs. It's important to seek a second opinion on the output. Is it good enough? Is it really doing what it's supposed to do? Many people hope, and tend to believe, that AI is a silver bullet that will solve all problems. I don't share this belief

"We should strive to not only focus on the technical aspects and potentials, but also the ethical and social implications."

Barbara *Okay, now that we've covered the possible future capabilities of AI from your perspective, what does the future we're headed towards look like on the spectrum from utopia to dystopia? What should we be preparing for?*

Joost-Pieter I believe we will definitely see AI integrated into various aspects of our daily lives. I also think that current developments, not only with AI but also with the so-called metaverse that major companies like Amazon and Microsoft are developing, will lead to communication playing an even larger role in our lives. It's hard to imagine life without a mobile phone today, but if communication extends beyond that to include sensors and emotions, for instance, and not just information and data, I think there will be a wide range of possible applications. I strongly believe this will have a significant impact on our social structures and the way we live as human beings. Of course, this also carries risks. I may be a bit skeptical, but I see that AI has not only a bright future. For instance, we already see drones and machine learning techniques being used in conflicts like the one in Ukraine and the Israel-Palestine situation. I fear this will only increase in the future, which is a serious cause for concern.

Barbara *Reflecting on the last few days here at AISoLA, which looks at artificial intelligence from an interdisciplinary perspective, is there a particular insight from another discipline that was particularly interesting to you?*

Joost-Pieter Indeed, the perspective from social sciences is crucial. As computer scientists, we often focus on the technical aspects, such as understanding the workings of a neural network or a partially observable MDP used in AI and planning. However, we must not overlook the enormous social implications. As I mentioned earlier, I foresee a significant impact on our social existence and potential risks. It's crucial that we consider these ethical and social aspects. We should strive to not only focus on the technical aspects and potentials, but also the ethical and social implications. What kind of impact will it ultimately have on us and society as a whole? What does it mean for a business if, for instance, you replace most personnel in an organization with AI agents like ChatGPT? What effects will this have? It's important to address these issues. What I appreciate about AISoLA is that these aspects are also considered, particularly the social aspect, which I find extremely important.

"We shouldn't always trust the output of AI without question. We need a rational approach to dealing with potential outputs."

Barbara *From your personal perspective, what AI vision would you like to see addressed?*

Joost-Pieter The vision I would like to see addressed involves two aspects. Firstly, when you ask AI specialists why certain techniques are successful, they often struggle to answer. Even at NeurIPS, the flagship conference on AI, prominent people refer to the success of deep learning as a kind of alchemy [5]. We combine certain techniques, and they work exceptionally well, but we struggle to explain why. This is one aspect that I believe is crucial. The other aspect is that we need to exercise caution when using AI in certain applications, and we need regulations in place.

Barbara *Regulations aimed at safety?*

Joost-Pieter Yes, in terms of safety, not just safety in a general sense, but the safety of human beings. We definitely need regulations in this area. It's not only the responsibility of computer scientists, but also politicians and strategists. They need to be involved and aware of these issues.

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

Joost-Pieter No, I don't believe so.

Barbara *Thank you very much for your views on AI and for taking the time to share them with us. Have a great last few days at AISoLA!*

Joost-Pieter OK, thank you.

References

1. Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc

V. Le, Charles Sutton: Program Synthesis with Large Language Models. CoRR abs/2108.07732 (2021)
2. Sanjit A. Seshia, Dorsa Sadigh, S. Shankar Sastry: Toward verified artificial intelligence. Commun. ACM 65(7): 46-55 (2022)
3. Jeannette M. Wing: Trustworthy AI Commun. ACM 64(10): 64-71 (2021)
4. https://en.wikipedia.org/wiki/Dutch_childcare_benefits_scandal
5. https://www.youtube.com/watch?v=x7psGHgatGM
6. https://www.youtube.com/watch?v=cNF1-lFfNs4