Let's Talk AI with Wolfgang Ahrendt

Wolfgang Ahrendt[1] and Barbara Steffen[2]

[1] Chalmers University of Technology, Department of Computer Science and Engineering, Sweden,
ahrendt@chalmers.se
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"My vision is that we align the AI developments with technological, scientific, interdisciplinary, and societal discussions about what is it we want and do not want. If we cannot agree, let us talk, hoping that, with help of this dialogue, we are better equipped for the task. And let us work hard on the alignment of trust and trustworthiness."

The Interviewee - Wolfgang Ahrendt

My Personal AI Mission:
To contribute to methods where we use AI in the context of program development, such that the resulting programs are trustworthy even if we do not rely on the trustworthiness of the used AI. Neuro-symbolic methods should help us with that mission.

My Takes on AI

Artificial Intelligence: A term which invites mystification, ever since it is used. Earlier, AI was a label for a family of symbolic methods. Today, the term AI is used almost exclusively for machine learning techniques and applications. It would be better if we call it machine learning, which is more descriptive than the term AI. Having said that, in this interview I always say AI, and never machine learning. Probably, it is too late.

Trust: Trust is the belief in trustworthiness. Admittedly, this is recursive definition. Still, this is a useful approach to thinking about trust. For instance, when aiming to increase trust in something, one can either try to influence the belief directly, or the trustworthiness which then feeds back to the belief in trustworthiness.

Explainability: The existence and transparency of cause-effect relations, be they exact or probabilistic.

Essential Elements of Human Capabilities: Intentions. Emotions. And the interference of emotions and rationality. Don't we often witness pretendedly rational debates which however are driven by emotions?

The Interview

Barbara *Today I have the pleasure of speaking with Professor Wolfgang Ahrendt. Could you please briefly introduce yourself and your personal relationship to artificial intelligence?*

Wolfgang Thank you for asking me to be part of this interview series. I think this is a great initiative. I am a computer scientist working at Chalmers University of Technology in Gothenburg, Sweden. I moved there 22 years ago, after doing my PhD in Karlsruhe, Germany. The focus of my research is software verification, but I started more generally in the area of automated theorem proving, which some people have labeled as AI.[1] Today the label AI has very much changed, and we refer to AI mostly in connection with neural network-based systems, which developed very rapidly in the last years. So, one could label the area I come from as “old AI”. However, I never liked these labels, not then, and not now. But that is a different discussion. I was not very involved with “new AI” until lately, when I started looking into how we could use tools like ChatGTP and CodeCopilot for programming in a good way. How can we exploit the power of large language models for programming without ever trusting them? This is my current connection to AI. I will elaborate on that in my talk later at this conference. Also, I am organizing here a track on AI-assisted programming, together with Klaus Havelund [2]. It is high time that we talk about the consequences, the possible and desired future of programming, in the light of this development.

"[...] there is a strong tendency towards wanting to believe what we are told by a machine; in particular if it comes across in an eloquent and seemingly informed way."

Barbara *AI-assisted programming is one of the main challenges that you are currently addressing in your AI research?*

Wolfgang Yes, and I am thrilled to work in the intersection between very new developments in AI and exact methods of the kind people like me have been investigating for a long while. There is a keyword used lately, "neuro-symbolic AI" [3]. It refers to combinations of, on the one hand, neural networks which are very powerful but difficult to analyze or explain, and, on the other hand, symbolic methods, like what people like me have been doing for a longer while. What I want to work on is neuro-symbolic methods for trustworthy software development.

Barbara *Are you mainly interested in the technical side or also in what we can learn from it, how we should work with it?*

Wolfgang Both. There are a lot of technical challenges which I find very interesting, the solution of which can make a real difference. So in some sense, there is a strong technical focus. At the same time, the talk I will give later today pictures a software development process which uses new AI and symbolic methods. And in the middle of the picture, there is the human developer who takes all crucial decisions, accepting or rejecting suggestions generated by AI and analysed by exact methods [1]. The focus of the work is neuro-symbolic software development, and the role the human plays within.

Barbara *Moving on to my next question about trust, which I think is also relevant to your research. What role does trust play in the adoption of AI, and what kind of measures are essential in terms of ethical AI adoption?*

Wolfgang Trust is one of the biggest problems we are having with the late AI hype. There is an enormous explosion and popularization of AI usages. Very many non-technical people actively use AI now. Actually, people are heavily interacting with AI for a longer while already, but under the hood. Social networks have been feeding us with content selected or filtered by AI. But now, many people use AI actively. One of the biggest challenges with that is a too big trust in AI systems. Whether we are non-technical or technical, there a strong tendency towards wanting to believe what we are told by a machine [4]; in particular if it comes across in an eloquent and seemingly informed way. Had it been a person writing that, it would be someone knowing what he or she is talking about. The trust we have learnt to put in humans expressing themselves like that is transferred to trust in machines who write in the same way. We see this phenomenon even in a technical context. As I said, I am interested in what all this means for programmers, who really are technical people. There are comparative studies about developers who use AI for programming and developers who do not. One such study also compares the trust the developers put into the security of the final result [5]. (The context was a security critical application.) The team which was not allowed to use AI had less trust in what they developed, compared to the team which was using AI when coding. But actually, the real security of the product was the opposite. The product developed without AI was more secure than the other. So there was an inverse relation between the trustworthiness and the trust, even in this technical context. This is one of the big challenges we have on the technical as well as the societal side with this AI boom. There is too much trust, actually. Let us not be too trustful here.

"I hope that all of us make a stronger effort to not just work on what is possible, but focus on what we want to happen and what we do not want to happen and how we can influence things in a better way."

Barbara *In terms of the technical capabilities of artificial intelligence in the future, on a scale of 1 to 10, where 1 stands for the AI systems we see today, such as ChatGPT, and 10 stands for artificial general intelligence that surpasses human capabilities. What do you think will be possible?*

Wolfgang I am skeptical. But it is also true that I, like most of us, did not think a few years back that an AI tool could do what ChatGPT is now doing. So what does it even mean that I am skeptical? But I am. I think we have a very long way to go, if it is even possible, for AI to make necessary connections between very different domains. To give you an example, take a car with autonomous functionality driving through a neighborhood with houses and gardens. Now, imagine a ball rolls onto the street from a garden. A human driver will likely connect a rolling ball with a child which may run after. But an AI based controller of a car will not anticipate the child. Why not? Because neither the controller nor the human can know this connection from the training data collected by driving around and all the situations which appear there. The human knows from totally different contexts the possible reason for the ball rolling around. Probably, someone is playing with it, most likely a child. This was not in my training set when I learned driving. I know it from a very different context, but it still helps me in the given context to make a decision. I think there are many such things, not all of them so life-threatening, that is not what I mean to say, but I think that we connect very many different things with each other. I think that machines are very far from that, and I am not sure they can ever do that. That is another debate, very speculative. Myself, I would rank it fairly low my trust into that a very general AI can act in a similar way as we can.

Barbara *If you had to pick a number from 1 to 10, what would it be?*

Wolfgang 3.

Barbara *Building on that, many different future scenarios are discussed, ranging from dystopia to utopia. Where do you stand?*

Wolfgang Dystopias and utopias are both speculative, with positive and negative connotations, respectively. I would not answer your question by saying that I like the discussion to be more grounded on where we currently are, what we are already experiencing as problems. Even if, hypothetically speaking, this technology would stay where it is (which is of course not true), we would have a lot of work to do to deal in a good way with the AI we already have. A good example of that was the discussion that took place in a session earlier today, about regulatory needs for big tech and companies in the information and communication sector. I hope that all of us make a stronger effort to not just work on what is possible, but focus on what we want to happen and what we do not want to happen and how we can influence things in a better way. – I am not sure I answered your question, actually. I can make another attempt. What was the question?

> "We have to prevent, as I said, unwanted consequences of AI. None of our disciplines can do that alone."

Barbara *Where would you place yourself on this dystopian-utopian spectrum? Can we look forward to the future or should we fear it?*

Wolfgang My personality is more of an optimistic kind. But when it comes to this topic, I think we will be better off if we have a somewhat pessimistic approach, in order to guide our actions to prevent unwanted consequences. Some of them have already materialised, like the effects of AI on general opinions in elections. I think we are better off if we are not driven by a belief that every progression in technology and ability is a good thing. Let us be a bit skeptic.

Barbara *Reflecting on the last days of this interdisciplinary conference, was there a particular insight from another discipline that you found interesting?*

Wolfgang I can immediately name one. There was a talk here by a researcher in law about liability. Early in the talk there were reactions from some of us computer scientists. The following discussion revealed a clash in terminology, but also in the conceptual approach. This is what is great about such an interdisciplinary event, that we can clear these things out and broaden each other's spectrum. It is not only about labelling, it is also about how we frame concepts and which kind of distinctions we like to be sharp on and less sharp on. This varies greatly among different disciplines. It is very inspiring to be exposed to this. More concretely, that talk was about liability in the context of AI, the different legal principles which are applied and have always been applied, like developer responsibility versus product supplier responsibility. Someone develops something, and someone else uses that thing for a product and sells the product. Where does the liability lie? Different principles exist and have been applied in different legal frameworks, on the national level or the level of the European Union, for instance. There were legal initiatives, some of which have been rejected, and so on. Researchers in law highlight in such discussions the consequences of law to insurance, for instance. Every law in this sector triggers insurance policies. In my community, we do not think of these things. I do not mean to say that we should focus on that, we have enough to do in our respective areas, and we also need coherence in a scientific tradition. But we do need this dialogue. This was an example where I learnt a lot.

> "I think, [we should] align the AI developments with technological, scientific, interdisciplinary, and societal discussions about what we want."

Barbara *Each discipline brings its own expertise, but it is important that all disciplines are guided by a basic understanding of the other disciplines to ensure that critical factors are not overlooked. Otherwise, each discipline may overlook certain aspects simply because they are not aware of them.*

Wolfgang Exactly. I think each of these areas is lacking aspects which are important, but also contribute aspects which the others are lacking. In this dialogue, each side is realizing in which terms another discipline is thinking. And that changes the framing of how I think about my own field and the overall context it is in. We have to prevent, as I said, unwanted consequences of AI. None of our disciplines can do that alone. The task is a very difficult and we may not get it right, but not doing anything would be much worse. All these different disciplines have to make their respective contributions.

Barbara *Do you have a specific AI-related topic or research question in mind that you would like to see addressed in the near future?*

Wolfgang Oh, there are so many. Myself, I decided to look into a specific one which I think that is among the things we should be doing. At this conference, I organise a track on AI-assisted programming [2], collecting people interested in what AI means for software development.

Barbara *From your personal point of view, what is the vision of AI that will lead us to a desirable future?*

Wolfgang All kinds of technologies, very much also AI, come with the promise of a better world. And yes, the world is getting better in some aspects, but may get worse in others. Someone says "We want to make the world better", but actually, they found a business model which works very well for them, right? The driving force is commercial. May be we cannot overcome that. But I think it is good if we realize that many impactful changes have commercial driving forces, they are not all about "Let's bring the world together" and all these big words. I do not want to be ultra-negative to many of the changes. But I think it is good that we are conscious about the forces which drive certain changes. I think I am again drifting away from your question, can you repeat it once more?

Barbara *What do you think the AI vision should be?*

Wolfgang Yes, the AI vision. It holds for AI as for everything else: if we find good ways to use it, great, let's do it. And what is good is a matter of discussion, of course. We may have different opinions on what is good and not good about the chatbot talking to a lonely person. Is that good or not? Maybe I find it bad until I am a lonely person. Such discussions are difficult, but necessary. And that should be our vision, I think, to align the AI developments with technological, scientific, interdisciplinary, and societal discussion about what is it we want. If we cannot agree, fine, let us talk and hope that after this dialogue we are a bit smarter than before. That is my vision of the AI.

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

Wolfgang No, I think I added quite a bit to the scope of your questions with thoughts of mine. - By reflecting out loud.

Barbara *Thank you, Wolfgang, for your time and your perspective on the various topics. I look forward to your presentation!*

Wolfgang Thank you, Barbara, for this opportunity. Much appreciated!

Barbara *Thanks to you, it was my pleasure.*

References

1. Ahrendt, W., Gurov, D., Johansson, M., Rümmer, P. (2022). TriCo - Triple Co-piloting of Implementation, Specification and Tests. In: Margaria, T., Steffen, B. (eds) Leveraging Applications of Formal Methods, Verification and Validation. Verification Principles. ISoLA 2022. Lecture Notes in Computer Science, vol 13701. Springer. https://doi.org/10.1007/978-3-031-19849-6_11

2. Ahrendt, W., Havelund, K. (2024). AI Assisted Programming. In: Steffen, B. (eds) Bridging the Gap Between AI and Reality. AISoLA 2023. Lecture Notes in Computer Science, vol 14380. Springer. https://doi.org/10.1007/978-3-031-46002-9_22
3. Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence, arXiv:2002.06177. https://doi.org/10.48550/arXiv.2002.06177
4. Mosier K.L., Skitka L.J. (1996). Human decision makers and automated decision aids: made for each other?, in: Automation and Human Performance: Theory and Applications, Erlbaum.
5. Perry N., Srivastava M., Kumar D., Boneh D. (2022). Do Users Write More Insecure Code with AI Assistants?, arXiv:2211.03622 https://doi.org/10.48550/arXiv.2211.03622