Let's Talk AI with Falk Howar

Falk Howar[1,2] and Barbara Steffen[3]

[1] TU Dortmund University, Department of Computer Science, Germany,
[2] Fraunhofer ISST, Dortmund, Germany,
falk.howar@tu-dortmund.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"We must ensure reasonable risk before deploying AI systems at a societal scale."

The Interviewee - Falk Howar

My Personal AI Mission:
I work on methods for ensuring that AI systems, autonomous vehicles in particular, are safe.

My Takes on AI

Artificial Intelligence: Intelligence exhibited by systems that are designed and constructed by humans.

Trust: The deliberate and rational reliance on another person to do something.

Explainability: The deliberate and rational reliance on another person to do something

Essential Elements of Human Capabilities: Consciousness, empathy, and the capacity for creating and manipulating symbolic systems.

The Interview

Barbara *I have the pleasure to sit with Professor Falk Howar from the TU Dortmund University. Please briefly introduce yourself and your relationship to artificial intelligence.*

Falk Certainly, it's a pleasure. As you mentioned, I'm a professor of software engineering. I'm currently affiliated with TU Dortmund University in Germany. Before joining TU Dortmund, I managed a small institute at TU Clausthal, another German university that closely collaborates with the automotive industry. Prior to that, I worked as a postdoctoral researcher at CMU and NASA Ames Research Center, where I worked on the safety verification and validation of autonomous aeronautic systems [4]. In the automotive industry, I focused on the safety of automated driving systems, which greatly influences my current research. I work on formal verification and learning techniques to generate models of systems and environments, which I then use to assure or verify the safety of these systems [10].

"Human civilization will have to find a way of dealing with the risk of these systems. [...] consequences sometimes only show up after a couple of years."

Barbara *Could you provide one or two specific examples of your AI related research questions?*

Falk Absolutely. From my perspective, one of the most significant challenges we face with AI is how to deploy systems that incorporate AI components at a societal scale, given the risks we don't fully comprehend yet. Personally, I work with automated driving systems. There is a real risk associated with introducing automated vehicles on the road. As we've recently seen in San Francisco, for instance, companies trying to implement this technology often find themselves involved in accidents, primarily because the systems don't fully comprehend or perceive their environment [5]. How does this relate to AI? Well, these systems operate in complex environments, and we have to train components because we can't program them to perceive these environments.

Barbara *In your opinion, which role does trust play in AI adoption?*

Falk Trust plays a significant role. Studies have shown two effects: under trust and over trust. Over trust can lead to over-reliance on systems. We've seen this in the automotive industry with the introduction of cruise control systems. After activating the system, drivers (in some studies) pay less attention [9], possibly due to a lack of understanding about what the system can and can't do. Then there is also the issue of automation fatigue. In hospitals, for example, you can observe this in intensive care units where numerous devices flash warning signs or emit sounds. Nurses can become somewhat desensitized to these signs [8]. We see the same with air traffic controllers or airplane pilots who receive an influx of collision warnings when approaching an airport due to surrounding traffic. One observed effect is that people simply ignore these systems. So, if you

don't trust the system, you won't use it. Conversely, if you trust the system too much, you may not question its impact. This may be what happened with social media. It emerged in our society and is now causing significant issues, particularly among younger people and girls. There is a well-known public health crisis among teenage girls in the US, for example [1].

Barbara *So, in your opinion, what measures are essential for ethical AI adoption?*

Falk I believe the concept of calibrated trust, developed by psychologists about 20 years ago, is interesting [6]. I'm not a psychologist, so this is a layman's explanation. The idea is that when using a system or collaborating with a person, you need to calibrate your trust to an appropriate level. This principle applies at both an individual and societal scale. For autonomous vehicles, for example, we started deploying them on the streets without fully understanding how they work or the associated risks. Now, we operate them in test fleets worldwide to better understand the domain and calibrate our understanding of the capabilities and risks. This allows us to make an informed decision about whether we should use them, permit them, and assess the remaining risk. This kind of calibration process needs to work at a societal level and at an individual level when working with these systems. We then need to design systems that allow for this calibration.

"I think we're overestimating the potential dangers or effects that AI could have in the short term [and] that we underestimate what these systems will be able do in the long run."

Barbara *To which extent does this calibration depend on specific use cases and scenarios or is it rather independent of it?*

Falk You could argue that there's a difference between high stakes and low stakes situations. If you're using an AI system to select your playlist while you're cooking, you might not care too much if it doesn't perfectly match your taste in music. You can simply choose a different playlist. However, when there's a risk of harm or damage, it's essential to perform a risk assessment and have processes in place that ensure the safety or reasonable risk of these systems.

Barbara *On a scale of 1 to 10, where 1 is artificial intelligence systems we know today, like Chat-GPT, and 10 is something like artificial general intelligence systems that actually surpass human capabilities, what do you think will be possible?*

Falk I'm going to sidestep the question slightly. I believe *the terms "intelligence" and "artificial general intelligence" are often used too broadly and are underspecified. We don't fully understand what we mean when we say that humans are intelligent and definitions of intelligence shift over time [7].* Whenever a computer can perform a task, that task is no longer considered a measure of intelligence. Sometimes we say intelligence involves creativity, reasoning, or the

ability to step back and consider the bigger picture. I can imagine building a machine that can do all these things. Maybe the interesting question then will become, can it have the capability to sidestep its programming and change its own programming in the way humans can?

Computers have long surpassed us in specific tasks. They can process vast amounts of data and perform computations that we can't. Currently, there's a lot of hype about the end of the world brought about by AI. It would perhaps only be the end of mankind, hopefully, and the world could continue. But I think we're overestimating the potential dangers or effects that AI could have in the short term because we've just seen these amazing examples of what large language models can do on specific tasks and how natural it feels to interact with them. On the other hand, I think that we underestimate what these systems will be able do in the long run.

You asked me to rate this on a scale from 1 to 10, but you also mentioned that they surpass human capability. That means we would have created a system that surpasses our own intelligence. Wouldn't the AI then also be capable of building a system that's more intelligent than it? Then the question is, is there any limit on intelligence or could this go on forever? Would it become a god? Would it create a universe? We don't know. But I think *once we've taken the initial step of creating something more capable than us, and it has the same capacity, this should, in principle, continue infinitely.*

Barbara *Okay, but what if you had to position yourself on a scale of 1 to 10. Where would you position yourself? More towards we're not going to see much more progress, or more towards we're not going to see an end to it, which would suggest an 8, 9, 10?*

Falk Okay, this is by analogy and I'm not sure if this is a working analogy: if you go back to the time when the internal combustion engine was developed. The first motor cars we had, they looked very much like horse carriages. They had open seats and replaced one component in the horse carriage with the combustion engine. Even the steering wheels were quite awkward. They couldn't go fast. It was clunky. Then look at how, with the same principle at heart, over 100 years, technology has advanced so much that vehicles now have incredible capabilities, a high degree of automation, and computers in them, and so on. Then imagine that we're currently at the point where someone invented the combustion engine equivalent in AI. I think we're going to see things that we can't even imagine currently.

"Whenever a computer can perform a task, that task is no longer considered a measure of intelligence."

Barbara *Using your example of the combustion engine, could it mean that we should focus less on the engine and more on its performance like how long it took us to get from A to B, its cost, its level of comfort, and the number of people it*

could carry? Do we need different metrics, such as comparing our productivity today to a possible level of productivity in the future?

Falk Could you elaborate?

Barbara *For example, outcome, performance, or something like that. So we could say, for example, humans today produce this kind of outcome or have this kind of performance. But because of the advancement of the tools, we suddenly observe a completely new level of what we are actually able to do, or what the AI-tools will be able to do by themselves. That comparison would be more in terms of results rather than discussing solutions or functionalities.*

Falk I attended a very interesting talk today, where someone showed how they used an AI system to aid software development, automating tasks that involve not only writing programs but also tasks heavily based on natural language: document parsing, chatting with people, writing requirements, and designing the architecture [3]. I think AI can do a lot of this boilerplate stuff. The authors observed a tenfold increase in performance because you can eliminate a lot of the boring tasks and focus on the interesting tasks, or the tasks where humans currently perform better than AI. We'll probably see an increase in this trend. The hard question that I wasn't able to answer for myself, before this conference, or even during the conference, is if there's a limit to it. Will it be the boilerplate for some time, and then we develop models with a new quality of capabilities, or will people design clever systems based on the current paradigms that (in the analogy of the combustion engine) somehow exhibit new capabilities.

"Maybe the interesting question then will become, can it have the capability to [...] change its own programming in the way humans can?"

Barbara *Looking at possible future scenarios, from dystopia to utopia, where would you position yourself?*

Falk I believe we have this in our own hands. And I've said this before in discussions here, it's very easy to paint a very dystopian future, and it's very easy to paint a very utopian future. I think either could happen, but human civilization will have to find a way of dealing with the risk of these systems, and deploy them gradually, because we don't fully understand their consequences, and consequences sometimes only show up after a couple of years of using those systems.

I think we have to take into account that maybe sometimes when we see, that there's an unreasonable risk that we didn't know about, we just shut operations down. We will have to find economic models, and insurance policies. There were many colleagues from other disciplines at AISoLA, lawyers for example, who had some very interesting insights on what can work, and what cannot work.

To answer your question: I'm going to say I'm faithful that we will find a way of living with AI, and it will have negative and positive outcomes, but I am

hopeful that it will be a net positive. I think what will determine this is in the end a question of power. Will it, e.g., be possible for authoritarian governments to deploy these tools and techniques to control their populations in an unethical way, or will - what we all dreamt of - the freedom of information in the internet help people to fight such governments? Will we find a way to regulate the big companies, as we did with pharma and tobacco companies in the past, to put warning labels on their products, to limit how they can advertise, and how they can deploy their products.

Barbara *Considering these trade-offs and looking back on the past few days, what insights did you find most interesting?*

Falk On a personal level, the talks given by lawyers were particularly enlightening. I learned a lot about how law works, and how reasoning works in law, e.g., for different ways of organizing liability.

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

Falk I think *we really need to understand how we address societal risk that is the consequence of deploying AI systems. This is interdisciplinary because you have to construct laws that create a framework for operation. It has to be economically viable. You need psychologists and doctors to judge potential health consequences. And you will need engineers and designers that build these systems and can adapt how we build these systems.*

Barbara *What should be the AI vision for the future from your personal perspective?*

Falk I don't think output or how we increase human productivity should be the primary goal. There's a lot of talk about that AI will take this job and that job and automation took all the blue-collar jobs in the past 40 years and now AI is coming for all the white-collar jobs and will make office workers unemployed or knowledge workers lose their jobs. There's is a chance of this happening to some degree. Through automation it actually did happen for blue-collar jobs. Of course, we have to think about alternative models of income and wealth distribution and somehow find a way of making everybody profit from these advancements instead of only the people who own the AI profit from deploying AI.

But this alone won't solve a big societal crisis of people not being able to define what they do and the worth of their life through work, which happens today to a huge extent. There is some very interesting ancient Greek philosophy about living a good life. *Aristotle writes about what it means to live a good life [2]. And it is not only about happiness, but to a large extent also about a meaningful life and how you contribute to your community, that you strive for excellence and virtue. And I think we will have to revisit these old ideas of living good lives and the vision for AI should be that it enables us to live good lifes.*

Barbara *At AISoLA, we met with computer scientists, psychologists, philosophers, and legal experts. Are there any other disciplines you would like to add?*

Falk I think other sciences would also be interesting. There was some discussion here about how AI could aid scientific discovery. We, e.g., talked a little bit about protein folding, where AI already had an impact. It would be interesting to learn from other disciplines how AI can actually speed up their scientific discovery to understand better if AI will surpass our capabilities soon in scientific discovery and if it does, if there are ways of us still having a chance of understanding what it discovers. From a more societal perspective, it would also be great to have not only psychologists, but people from healthcare, maybe doctors who perform surgeries in hospitals or nurses who work with elderly people to get their perspective on AI.

Barbara *We are coming to the end of this interview. Is there anything else you would like to add?*

Falk No, I think I've spoken quite extensively.

Barbara *Then thank you very much for your insights and time, Falk. Enjoy the rest of AISoLA.*

Falk Thanks. Thanks for having me. Thanks for this great conference. It's been a blast.

References

1. Abi-Jaoude, E., Naylor, K.T., Pignatiello, A.: Smartphones, social media use and youth mental health. CMAJ **192**(6), E136–E141 (2020). https://doi.org/10.1503/cmaj.190434, `https://www.cmaj.ca/content/192/6/E136`
2. Aristotle: Aristotle: Nicomachean Ethics. Cambridge Texts in the History of Philosophy, Cambridge University Press (2000)
3. Belzner, L., Gabor, T., Wirsing, M.: Large language model assisted software engineering: Prospects, challenges, and a case study. In: Steffen, B. (ed.) Bridging the Gap Between AI and Reality - First International Conference, AISoLA 2023, Crete, Greece, October 23-28, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14380, pp. 355–374. Springer (2023). https://doi.org/10.1007/978-3-031-46002-9_23, `https://doi.org/10.1007/978-3-031-46002-9\_23`
4. Giannakopoulou, D., Howar, F., Isberner, M., Lauderdale, T., Rakamaric, Z., Raman, V.: Taming test inputs for separation assurance. In: Crnkovic, I., Chechik, M., Grünbacher, P. (eds.) ACM/IEEE International Conference on Automated Software Engineering, ASE '14, Vasteras, Sweden - September 15 - 19, 2014. pp. 373–384. ACM (2014). https://doi.org/10.1145/2642937.2642940, `https://doi.org/10.1145/2642937.2642940`
5. Kloukiniotis, A., Papandreou, A., Lalos, A., Kapsalas, P., Nguyen, D.V., Moustakas, K.: Countering adversarial attacks on autonomous vehicles using denoising techniques: A review. IEEE Open Journal of Intelligent Transportation Systems **3**, 61–80 (2022). https://doi.org/10.1109/OJITS.2022.3142612

6. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors **46**(1), 50–80 (2004). https://doi.org/10.1518/hfes.46.1.50_30392, https://doi.org/10.1518/hfes.46.1.50_30392, pMID: 15151155
7. Legg, S., Hutter, M.: A collection of definitions of intelligence. In: Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006. p. 17–24. IOS Press, NLD (2007)
8. Lewandowska, K., Weisbrot, M., Cieloszyk, A., Mędrzycka-Dąbrowska, W., Krupa, S., Ozga, D.: Impact of alarm fatigue on the work of nurses in an intensive care environment—a systematic review. International Journal of Environmental Research and Public Health **17**(22) (2020). https://doi.org/10.3390/ijerph17228409, https://www.mdpi.com/1660-4601/17/22/8409
9. Sanbonmatsu, D.M., Crabtree, K.W., McDonnell, A.S., Cooper, J.M., Strayer, D.L.: Automated driving experiences, attention, and intentions following extensive on-road usage of a level 2 automation vehicle. Journal of Safety Research (2024). https://doi.org/https://doi.org/10.1016/j.jsr.2024.05.002, https://www.sciencedirect.com/science/article/pii/S0022437524000550
10. Schallau, T., Naujokat, S., Kullmann, F., Howar, F.: Tree-based scenario classification. In: Benz, N., Gopinath, D., Shi, N. (eds.) NASA Formal Methods. pp. 259–278. Springer Nature Switzerland, Cham (2024)