Let's Talk AI with Jakob Rehof

Jakob Rehof[1,2] and Barbara Steffen[3]

[1] TU Dortmund University, Department of Computer Science, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Germany,
jakob.rehof@tu-dortmund.de
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"We are in the middle of an AI revolution and we need to figure out how to use this technology creatively and at the same time mitigate the risks that are associated with it."

The Interviewee - Jakob Rehof

My Personal AI Mission:
To help further the creative development and use of trustworthy AI technology.

My Takes on AI

Artificial Intelligence: Computer systems that can to a significant degree exhibit behavior characteristic of biological or human intelligence.

Trust: Believing that someone or something will behave as expected.

Explainability: The ability of a system to reconstruct the reasons why the system behaves the way it does.

Essential Elements of Human Capabilities: Background understanding of the world, semantics, reflectivity.

The Interview

Barbara *I have the pleasure to interview Professor Jakob Rehof of the Technical University Dortmund. Please briefly introduce yourself and your relation to artificial intelligence.*

Jakob Thank you for having me. I'm a professor of computer science at the Technical University of Dortmund. Much of my research has focused on algorithms and complexity, verification, and mathematical logic with applications to programming technology. To mention one example, in recent years, I have been particularly interested in program synthesis [15], the automatic generation of code, and increasingly connecting that with applications in classical engineering, for instance by applying these techniques to the automatic generation of simulation models or system designs in logistics, engineering and production technology [6, 8, 5, 9, 2]. In this context, we work on methods for automatically generating whole families of simulation models embodying rich sets of design variants and then possibly testing, measuring and exploring those models [11]. For instance, you can imagine trying to construct an object like a robotic arm and attempting to assemble various possible solutions to such a design problem. Say the robotic arm should be able to move in a certain way at a certain speed. You can now co-generate a lot of things in that context. You can generate the CAD design in several possible solution variants together with accompanying simulation models which you can then measure and explore to automatically identify the best designs with respect to a set of parameters (KPIs). The goal here can be referred to as design space exploration [7].

Now, you asked about my relation to AI because that's where it comes in. If we have the ability to generate code that may represent some system designs and we're able to generate multiple solution variants for a certain design problem, and we're able to measure various parameters on those generated designs, then you can think of combining that with learning mechanisms. You can create a generate-test-and-learn loop. You generate a possibly large number of solutions to a problem, generate some simulation models, use them to perform measurements, and then use the data that comes out of those measurements to feed them into a learning mechanism. Then you can close the loop by trying to learn how to optimize the design problem that you started out with. That is one important area of current interest to me in connection to AI in my own research.

I also have interests of a more general nature. For example, I'm associated with the Gaia-X project [17], building up a think tank in the context of Gaia-X, called the Gaia-X Institute, which is supposed to think ahead about topics related to regulation of information technology. Particularly in the context of regulations from the European Commission, as you may be aware, there are a number of regulatory acts coming out now. There's the Data Governance Act, the Data Act, and the AI Act is under discussion. And so there again, you see there's a connection. I'm interested in the area of regulation, and I'm interested in problems such as how we may help implement such regulatory acts, for example, by

studying the question of automated compliance: How can we develop technologies that make it easier to verify or certify that a system is compliant with the regulation? So that also pertains to the AI topic in as much as the AI area is an important object of regulation.

Barbara *Thank you for the overview. What role does trust play in AI adoption?*

Jakob Trust is a key factor in AI adoption, especially if we operate in a regulated space. In an autocratic environment, trust might not be as much of an issue. But in our part of the world, it is a key issue. And it's not just about increasing trust. It is as much about calibrating trust, that is, aligning the level of trust with the actual trustworthiness of the systems. And I think that's a crucial issue for AI systems.

Barbara *So essentially it is about ensuring that the end user doesn't have too much trust, but also not too little. That means that psychologists and AI experts, for example, have to determine the right level of trust in terms of how trustworthy this application is, and then figure out how to make sure that the end user places this right level of trust in the application.*

Jakob Yes, that's certainly one aspect of it. It's not just about one direction, increasing trust or decreasing trust. You may overtrust a system, but you may also undertrust a system, as you suggest. A system might actually be more reliable or trustworthy than you think. And if that leads you to not using the system, if it's a useful system, then that's not a desirable state of affairs. So it goes in both directions. Now, who is supposed to decide, for example, how trustworthy a system is? That may be a question for experts in some cases, but it may also be a question of human psychology, where in some sense the end user is the de facto judge on how trustful you as a user typically will be towards a system. And so it's a complicated matter that, depending on the situation and the system and so forth, will play out differently.

"There is a basic understanding in many parts of the industry that something in the way of regulation is needed."

So, for example, if you take something like autonomous driving, it's going to be very important what, by expert knowledge, can actually be known about certain features of such a system. For example, what is the probability that certain accidents might happen? But then weighing the psychological impact of that probability, that's not necessarily something that can or should be left for the technical expert to decide. You may have to understand how most normal human beings actually react to something like the probability X of something bad happening in a certain domain. It doesn't help us if, for example, an expert will tell you, "You can go ahead and use this system. It's no problem because the probability of something bad happening is zero point something", if most normal people don't actually react to that system and that probability that way. It's a well-known fact that, for example, most people don't necessarily deal very

rationally with things like probability [4]. So that needs to be taken into account. That's possibly also a matter for empirical psychology, for example [12].

Barbara *Can you think of relevant measures to ensure the ethical use of AI?*

Jakob Yes, I believe that regulation is a key factor. It's an area that we're not really used to in general when it comes to software-intensive systems. Of course, we have regulations in various specialized vertical domains, like, for example, health care systems or autonomous driving systems. What is rather new is that in Europe, the legislator is undertaking regulating more general software based systems, for example, systems pertaining to data exchange and the usage of cloud infrastructures in that context, or AI systems. That's rather novel and I think that's actually a pioneering effort. I think that is probably one of the main instruments for achieving ethical AI and its deployment in an ethically responsible way [13].

Barbara *Now, if we look at the technical capabilities that AI might have in the future, on a scale of 1 to 10, where 1 describes the artificial intelligence systems we already see, like ChatGPT, which are very dedicated to specific functionalities in a specific context, and 10 describes something like artificial general intelligence, and refers to surpassing human capabilities in all areas, what do you think will be possible?*

Jakob I find that very hard to say, partly because the notion of artificial general intelligence does not seem to me to be quite sharply defined. To take something like ChatGPT, it already seems to point to the possibility of something like artificial general intelligence by some definition of that concept where natural language processing is important. I think it could go very far. Whether I should quantify it as 10 or 7, I don't know, but it's probably more than 5, would be my take. I was very surprised when these technologies started to show their potential. For me personally, the first time I really got aware of that was because of Google's AlphaZero chess engine and shortly afterwards I became totally impressed with Google Translate. I'm trying to pick up Italian, and I use Google Translate quite a lot for that, and I am absolutely fascinated with its capabilities. I thought earlier that this level of performance in machine translation would be impossible. And so you then move to the generalization of that in the form of something like ChatGPT, and of course it's easy still to poke holes in ChatGPT to sort of bring it out of balance if you persist long enough and hard enough.

"Could you imagine regarding an AI system as a legal persona? What would that mean? [...] should there be a kind of rights for robots?"

But if you imagine the progress that we have seen so far still going on for some time, like every time you bump up the version number on ChatGPT, you seem to get an exponential improvement in performance and quality, then that could lead to a situation where it becomes very difficult to expose the AI as a machine,

at least in certain application contexts. So I think this could well go very far. Exactly how far it will go, I can't tell you. I don't think anybody really can.

Barbara *Take your technical understanding of AI and think of all the different future scenarios currently being discussed, from dystopia to utopia. What future do you think AI will bring?*

Jakob Well, obviously somewhere in the middle on that spectrum, because I mean, as for the dystopian perspective, I do believe that there is enough potential in the technology itself that it does make some sense to ask that question and to be concerned with it. That's why I said that I think regulation is really important so as to avoid the dystopian effects. And as I already mentioned, I think EU is performing pioneering acts in that area. But it also means that in areas and places in the world that work very differently from, say, Europe or other parts of the world that we maybe are close to in the mode of operation, you already see dystopian effects like facial recognition technology being used for surveillance of general populations, social scoring, etc., etc. These are things that under a European regulatory regime would be forbidden. It would be a criminal act to implement and operate systems in that way. And so there are dystopian aspects. They can be prevented, but it's basically going to be a political and cultural question of whether you succeed in doing that.

"One fascinating question is, would it be possible to think of an AI system as being in some sense responsible for its actions?"

There's a lot of fantastic and positive potential, which is the opposite of dystopian. If I just take something like software technology as an example, which is one field of special interest to me, we have for a very long time not really seen big jumps in automatic programming technology, for example. Software development technology has not really evolved in fundamental ways for quite many years. Something like ChatGPT applied to writing code, I think, is a great prospect. And I think it's basically good for both software research and the professional area of software development. Because it will allow us, at least for some classes of systems, to, if I may put it this way, take out a lot of the dumb work of software development, thereby making it much more interesting for humans to be engaged in that professional field.

Barbara *Take the EU AI Act, for example. Today, it is mainly the big tech companies that are at the forefront of new artificial intelligence applications. Could this regulation lead to a disadvantage for European users? As a group, they could lose access to companies, research results, advances and AI tools developed in other countries or available to users in other countries.*

Jakob Right. That's a reasonable question. Another variant of that question, which is also reasonable, would be to ask whether it might bring about a disadvantage, not for the end user, but for the industry. Does it become harder to generate digital innovation and related business models in the European space under such regulation? Such concerns have already been raised in many con-

texts and it's reasonable to discuss them. I do believe that the legislator is quite aware of these considerations. I can't tell you whether we immediately will find the right place to draw the line in each case, but it is something that is on the mind of the people who make this kind of regulation. Also, I think regulation is clearly needed for anything like our cultural and socio-economic and political kind of realm that we can identify with. And I think others will have to follow. I know, for example, there are important players in the US administration who are looking to European regulation with great interest. And so, I think in that sense also it's a pioneering act because others will have to follow in one form or the other. You also find, for example, that some of the big tech companies actually express interest in reducing uncertainty, which may involve being told by a legislator what is not viable and how they should behave vis-a-vis these huge problems of misuse that could be arise. That may not mean that they're always in agreement with a particular legislative system as to how exactly it should be done. But there is a basic understanding in many parts of the industry that something in the way of regulation is needed. And so, that's already a start.

Barbara *Reflecting on the past few days, what new insights have been particularly interesting to you?*

Jakob First of all, I want to congratulate the organizers on a very interesting and important conference. And I think it was very well received by everyone participating, is my impression. There were a lot of interdisciplinary discussions that I found very interesting. Let me mention just one example. In the context of a discussion on AI with a very interdisciplinary group of people, which came together here at this conference, there was a discussion about the notion of responsibility in connection with AI-based systems. Responsibility from a philosophical point of view, from an ethical point of view, and from a legal point of view. One fascinating question is, would it be possible to think of an AI system as being in some sense responsible for its actions? Can it be creative? Questions like that are interesting. And then you may put a legal spotlight on it. Could you imagine regarding an AI system as a legal persona? What would that mean? I mean, there have already been discussions around this, such as, should there be a kind of rights for robots? I don't happen to think myself that would be a good or even a meaningful idea. But it's interesting to reflect on the reasons we might have for choosing one or the other stance on such a question.

"I think regulation is really important so as to avoid the dystopian effects."

Barbara *Do you have a specific research question or topic in mind that you would like to see addressed from a more interdisciplinary perspective? And if so, which disciplines should be part of this research?*

Jakob It's hard to focus on one particular topic because there are quite many of them. Let me just mention a couple of things. The ethical and regulatory questions are important. The interface between technology and law is coming

more into focus for AI systems. Other than that, we already see, you know, psychology, sociology being quite active, actually, in the discussion. And so I'm actually impressed with how fast these other fields outside of computer science have mobilized towards contributing to thinking about these questions, and that is good. On the more technical side there are problems of verification, testing, validation, and certification of AI systems that need to be considered in order to achieve trustworthy AI. So we will not be running short on things to do.

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

Jakob I don't think it makes sense to have the AI vision. I just I think there are so many different aspects and you have to take a very differentiated view on it. My own mission statement in this context would be, in one sentence: To help further the creative development and use of trustworthy AI technology.

Barbara *Is there an overarching goal you would like to see addressed or achieved?*

Jakob I don't see one single overarching goal. I see many different areas where you can imagine great advances happening, for example, in medicine, from diagnostics to new medication. Also other parts of science will be positively impacted, based on AI technology. At the University of Dortmund, the physics department had a recent breakthrough in applying machine learning to interpreting data coming from astrophysical measurements, making it possible to discern patterns that come out in huge data sets from measurements of cosmic radiation. And so science in general may see great advances based on this technology. Then there are all the areas of life that can be improved, in quite different ways. Think of early warning systems and understanding climate change, where analyzing huge data sets and learning from past data can be helpful. So I tend to think of it as not one thing, but it's many different things. What ties those things together is a certain coherence to the underlying technology of machine learning based on statistical methods. But that is also not a quite simple matter, since there are different kinds of of AI technologies within that spectrum. For example, neural network based technology [3, 1] is a different sort of learning strategy than, say, reinforcement learning [10]. These things can be combined, of course, but they have different characteristics.

I also believe we may see at some point a more integrated approach where machine learning and statistical methods are combined with more classical technologies based on mathematical logic. We already see that happening in the context of trustworthy AI and explainable AI, where learning systems which might do dangerous things need to be controlled or even verified using logical methods. Think for example of a system controlling a fleet of drones moving around in a populated environment. One would like to have hard logic based guarantees that certain bad things simply cannot happen, right. So the technology is not fixed. And, incidentally, in recent years we have seen absolutely spectacular progress in the area of automated proof with proof assistants and proof checking [16], which is relevant for formal verification [14]. More generally, there is a need to combine

the machine learning dimension with two further dimensions, that of integrating with prior knowledge and that of integrating with high quality data. That direction of so-called "triangular AI" is an important direction for the research program which is pursued in the new Lamarr Institute for Machine Learning and Artificial Intelligence at the universities of Bonn and Dortmund.

Barbara *Do you have anything else you would like to add?*

Jakob I think we covered a lot of interesting ground and I just want to thank you for the opportunity to to talk about it here. Thank you.

Barbara *Thank you, Jakob, for your time and perspective on this topic. Have a great day!*

Jakob Yes, you too. Thank you.

Barbara *Thank you.*

References

1. Christopher M. Bishop with Hugh Bishop: Deep Learning. Foundations and Concepts. Springer 2023.
2. Constantin Chaumet, Jakob Rehof, Thomas Schuster: A knowledge-driven framework for synthesizing designs from modular components. To appear in 34th CIRP Design Conference, Procedia CIRP 2024.
3. Ian Goodfellow, Yoshua Bengio and Aaron Courville: Deep Learning. MIT Press, 2016.
4. Kahneman, Daniel: Thinking, Fast and Slow. Farrar, Straus and Giroux 2011. ISBN 978-0374275631.
5. F Kallat, J Pfrommer, J Bessai, J Rehof, A Meyer: Automatic building of a repository for component-based synthesis of warehouse simulation models. Procedia CIRP 104, 1440-1445, Elsevier 2021.
6. A Mages, C Mieth, J Hetzler, F Kallat, J Rehof, C Riest, T Schäfer: Automatic component-based synthesis of user-configured manufacturing simulation models. 2022 Winter Simulation Conference (WSC), 1841-1852, IEEE 2022.
7. L. Nardi, A. Souza, D. Koeplinger and K. Olukotun: HyperMapper: a Practical Design Space Exploration Framework. 2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Rennes, France, 2019, pp. 425-426, doi: 10.1109/MASCOTS.2019.00053.
8. T Schäfer, J Bessai, C Chaumet, J Rehof, C Riest: Design space exploration for sampling-based motion planning programs with combinatory logic synthesis. International Workshop on the Algorithmic Foundations of Robotics, 36-51, Springer 2022.
9. T Schäfer, JA Bergmann, RG Carballo, J Rehof, P Wiederkehr: A synthesis-based tool path planning approach for machining operations. Procedia CIRP 104, 918-923, Elsevier 2021.
10. Richard S. Sutton and Andrew C. Barto: Reinforcement Learning. An Introduction. Second edition, MIT Press 2020.

11. S Wenzel, J Stolipin, J Rehof, J Winkels: Trends in automatic composition of structures for simulation models in production and logistics. 2019 Winter Simulation Conference (WSC), 2190-2200, IEEE 2019.
12. M Wischnewski, N Krämer, E Müller: Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1-16, 2023.
13. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
14. https://en.wikipedia.org/wiki/Formal_verification
15. https://en.wikipedia.org/wiki/Program_synthesis
16. https://en.wikipedia.org/wiki/Proof_assistant
17. https://gaia-x.eu/