# Let's Talk AI with Holger Hermanns

Holger Hermanns[1] and Barbara Steffen[2]

[1] Saarland University, Department of Computer Science, Germany,
hermanns@cs.uni-saarland.de
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

***"The explosion of opportunities for software-driven innovations comes with an implosion of human opportunities and capabilities to understand and control these innovations."***

## The Interviewee - Holger Hermanns

**My Personal AI Mission:**
To help preventing the disenfranchisement of the individuals through cascades of software-based automated decisions.

## My Takes on AI

**Artificial Intelligence:** "Artificial Intelligence" has become a misnomer for something that is "Artificial Imitation" at a large scale.

**Trust:** Since trust appears to be a chiefly subjective notion, it seems worthwhile to instead focus on trustworthiness as an indicator of justifiable trust.

**Explainability:** Explainable intelligent systems need context-aware and situation-specific approaches to explainability. The resulting requirements might not be suitable for promoting trust, but trustworthiness.

**Essential Elements of Human Capabilities:** Nothing specific, except in expert contexts as needed for human oversight of high-risk AI systems enforced in the upcoming (finalised) EU AI Act.

## The Interview

**Barbara** *Hello, Professor Holger Hermanns from Saarland University. Thank you very much for taking the time for this interview. Could you please introduce yourself and your relationship to artificial intelligence?*

**Holger** Certainly, I'm a professor of computer science at Saarland University. My background is in the theory of computing with a focus on verification. Over the years, I have branched out into other areas and become more applied. Since around 2016, I have been investigating what we call perspicuity, which refers to the complexity of modern systems, including AI systems, and the challenges this poses for transparency and comprehension. This has led to two main research activities. One is a so-called Sonderforschungsbereich, a special research initiative involving around 50 to 60 researchers. It is funded by the Deutsche Forschungsgemeinschaft with a budget of roughly 3 million euros per year. Here, we are exploring how to design systems that are inherently explainable, transparent, and comprehensible. This research brings together experts from human-computer interaction, verification, and various areas of AI. However, the scope is broader than just AI. Today, we are facing many systems that aren't classified as AI systems, yet they are still too complex to fully understand or control. A prime example is the diesel emission scandal, which was essentially a software scandal, yet without an AI component to it. The automotive software was misused massively, against the interest of the society and car owners [2]. And the systems were so complex that their workings remained hidden. Even now, some seven years after the uncovering of the scandal, there are court cases that are partly about deciphering that software. Software complexity is a driving force in my research. The other aspect of my work that seems worth mentioning is interdisciplinary research on explainability and AI, conducted with colleagues from psychology, philosophy, ethics, and law [5]. Some of them were actually here this week.

"We're working with the premise that AI is meant to be beneficial and as such is worthwhile to be promoted, but also that mechanisms are needed to prevent it from having adverse effects [...]."

**Barbara** *Can you give an example of one or two specific research questions that you are currently working on with respect to AI?*

**Holger** We're working with the premise that AI is meant to be beneficial and as such is worthwhile to be promoted, but also that mechanisms are needed to prevent it from having adverse effects on our society or from infringing on fundamental human rights. There are instances where this balance is clearly not maintained. Current legislation, especially the upcoming European AI Act, enforces the concept of human oversight for high-risk AI systems, for instance when it comes to AI systems used to decide about access to higher education, such as for screening applicants applying to a university. These systems must be free of discrimination, and since that is nowadays almost impossible to achieve with current AI technology, there is the idea that if a qualified individual oversees the AI system in such high-risk situations, then responsibility can be reassigned to this human oversight. The trick then is that with the presence of an oversight person and the reassigned responsibility, the known legal system – such as insurances covering damages or providing compensations in case of wrong decisions – can be reinstalled, despite the blatant problems of the underlying AI technologies. This concept brings up many challenges with respect to the technical design of the system. It is essentially a form of human-computer interaction where the human is a trained expert. Questions arise about how the technical system should communicate with the oversight, how the system can be influenced, when the oversight is actually held responsible, and how they know of their own responsibility. This is interesting, especially as the laws are currently being shaped. In our interdisciplinary research, lawyers are discussing with us how the law is designed and formulated, and we aim at distilling requirements regarding the technical aspects - which ultimately boil down to computational aspects [1, 6]. What properties can we guarantee, and how can we ensure them? This is one of the most pressing issues at the moment, since the legislative European institutions, as per the AI Act, anticipate guidelines for this to be developed within the next few years. We have just finished a paper on the "AI act for the working programmer" to provide assistance in navigating the 450-pages document.

**Barbara** *What role does trust play in the adoption of AI?*

**Holger** It's practically important, but difficult to grasp. Some psychological studies suggest that people tend to trust computing systems more than they trust humans, and that explanations do not necessarily facilitate trust [4]. The more mysterious the system, the more trust they seem to place in it, to say it simply. If they understand how it works, their trust diminishes. There's also the phenomenon of unjustified trust, where people trust an app simply because a celebrity endorses it, even though there's no rational reason for this trust. So, while trust is likely important for adoption, justified trust may not be as crucial.

**Barbara** *Could this be addressed with more awareness and education?*

**Holger** I firmly believe in education. I think our society needs to be better educated about the systems they interact with. Whether this pertains to trust, perhaps. The more competent individuals are, the more justified their trust or mistrust may be. But this could also be a subjective perception. Perhaps I'm overestimating our capabilities.

**Barbara** *Could it be that people perceive systems as trustworthy because we see them as aggregates of the knowledge of all the people who helped design and develop them? Whereas when we deal with just one specific person, we recognize that that person may know more than we do, but assume that his or her knowledge is most likely limited compared to the system's knowledge?*

**Holger** That could very well be the case. However, the knowledge in these systems is largely syntactic, based on combinations of words, rather than semantic, understanding the meaning behind the words. This could be a misconception among the public, but it’s an important distinction.

**Barbara** *It’s often described as gathered and aggregated knowledge, isn’t it?*

**Holger** Yes, but the misunderstanding is that this isn’t factual knowledge. It’s knowledge about combinations of words or phrases. The meaning behind these words is what’s missing.

**Barbara** *And what kind of measures do you think would be helpful or essential to ensure ethical use of AI in the future?*

**Holger** About seven or eight years ago, I came across this issue. Then, I took the initiative for a lecture series, nowadays called "Ethics for Nerds". It got a few distinguished awards. The aim of that lecture series is to ensure that the computer scientists we train or collaborate with behave in a morally responsible way. I believe education is a key factor in promoting the ethical use of AI. It’s a complex issue, partly because the effects are so indirect. Unlike a knife, which is obviously dangerous, the potential misuse of AI for unethical purposes is not as straightforward to pinpoint. Despite this, I appreciate initiatives like “AI for good”, which uses AI in non-conventional ways to tackle societal issues. For instance, my colleague Ingmar Weber uses satellite imagery to study poverty and its changes over time, particularly after disasters. However, ensuring that AI technology remains in good hands is a significant challenge.

"However, the knowledge in these systems is largely syntactic, based on combinations of words, rather than semantic, understanding the meaning behind the words. This could be a misconception among the public, but it’s an important distinction."

**Barbara** *In terms of the technical capabilities of AI in the future, on a scale of 1 to 10, where 1 refers to artificial intelligence systems like ChatGPT and 10 refers to something like artificial general intelligence that surpasses human capabilities, what do you think will be possible?*

**Holger** Firstly, I think "intelligence" is a misnomer for what we’re seeing. There’s no intelligence in artificial intelligence. It’s artificial imitation at a large scale. So, anything like general intelligence is nonsense. Therefore, my best guess is that the limit is at 2 or 3 out of 10. So, I lean towards the pessimistic side regarding the capabilities.

**Barbara** *So, in essence, not much more progress than we are already seeing today. How do you define intelligence? What is missing to talk about intelligence instead of imitation?*

**Holger** What’s missing is understanding. These systems don’t understand. They can be creative in combining things in ways that haven’t been done before, which can be surprising while the structures are as to be expected. But the meaning, the semantics, is missing. That is the main point. If ChatGPT tells you something where a “2” appears, it does not understand that "2" is a number, for example. It’s just connecting words based on sophisticated statistics and a little surprise element, the latter for the purpose of avoiding generated texts get boring.

**Barbara** *When you look at that difference, do you think that we need to change something in the educational system, in terms of how we learn or work, to make sure that we maintain our advantage over AI in terms of intelligence versus imitation? Or to ensure that we don’t lose that ability that makes us potentially unique?*

**Holger** There are repetitive tasks where AI will advance. It can be a relief for certain simple tasks. Now, your question is whether we need to change the educational system to maintain our advantage. I think we will keep our advantage. We don’t necessarily have to change because of that. But, for instance, it will be much more challenging for students to pass exams if they were so far mainly based on repetitive tasks, especially if those are given out as homework. If they instead are presented as part of an exam and there’s no way to cheat, then maybe it’s still possible to maintain a major share of simple repetitive tasks among what is examined. Still, higher cognitive processes are what distinguishes us from the capability of carrying out mechanizable tasks. And that should be what is actually being taught and evaluated in education. I therefore believe that the way we assess whether students have gained sufficient knowledge may need to change. And the other question is should we also teach different things? I frankly think we are teaching the right things, but our exams consist of a sizeable portion of repetitive tasks, at least for many of our courses. Yet, a standard computer science lecture on university level usually includes intellectually challenging tasks. Likely, the examination must focus more on these aspects.

**Barbara** *So, it is about a balance between checking that students can repeat the definitions and concepts to establish a common language, and assessing students’ understanding by checking whether they can apply these insights to specific scenarios.*

**Holger** Yes. If students are asked to show that they can reproduce definitions, they should also be asked to prove that it was them and not a system that was doing so, right? So, we agree, essentially.

**Barbara** *Interesting. Now, in light of this new AI reality, a lot of different future scenarios are being discussed, from dystopia to utopia. You hinted at it, but given what you said, where do you personally fall on that dystopian/utopian spectrum?*

**Holger** I am on the dystopian side. I think AI, especially machine-learned systems, are extremely good at optimizing for the average case. If we have tasks where it’s not an issue that the non-average cases are treated sub-optimally, then I think these systems are great. That’s where they should be used, but only there. As soon as we have populations where it’s not enough to optimize for the average case, then I think we need strict rules to prevent people from being disadvantaged. This may seem a bit like German scepticism, of course, but I would see myself as justifiably dystopian in this regard. Actually, the AI Act seems to be taking the right steps in this respect, by defining high-risk AI systems as those that are to be subjected to regulations and to human oversight [8, 3].

**Barbara** *Given recent developments and the strong push from big tech companies, do you think it will be possible to integrate AI only in safe or non-critical contexts? Or do you fear that we will see some kind of push to integrate AI in environments that we can't really control? Towards dystopia.*

**Holger** There is a push by big tech to lure us into new functionalities that are fancy and get our attention. But we pay for that by giving out our data and losing our anonymity. Luckily, here in Europe, we do have politicians that are alert and seem to understand what is at stake. I think on the European level the right moves are being made. Sometimes this also happens on the national levels. Even if some of the regulations are a bit fuzzy, let them be fuzzy. As much as they are fuzzy, it is difficult for big tech to sneak out easily.

"We should strongly fight for the right for inspection of software that influences us and the things we own."

**Barbara** *Looking back on the discussions and conversations of the past few days, was there an insight from another discipline that you found particularly interesting?*

**Holger** I particularly enjoyed the discussion of the legal techs. I was mostly in this part of the program that was interdisciplinary from the start. There was psychology there, but especially law. That I found quite illuminating. Also, the discussion that happened in the corridors, how these techs were designed. Why is it that the AI liability directive is as it is and so forth. That I found very instructive. That is not directly influencing the work that I'm doing, but it gives me a context.

**Barbara** *Is there a particular research question or area you would like to see addressed more interdisciplinarily?*

**Holger** I mentioned human oversight already. I think this will become an interdisciplinary topic. For psychology people who are basically interested in the organization of work, there is a new job profile emerging, which is human oversight. What are the psychological capacities? What are the stress situations? I think there's a lot that needs to be framed there. We are working on this indeed. The legals define the context and we support with software tools. Other than that, I think what is chiefly underdiscussed is the problem of intellectual property of software. Beyond AI in software, your smartphone or your car are working because of the software embedded therein, right?

"I do think that open source is a good way to enable finding all kinds of issues and reporting them. I do think code secrecy and obfuscation is no good strategy to prevent attacks [...]."

And: You own the smartphone; you own the car. But you don't own the software. And you are not even allowed to look into the software. It is the intellectual property of the manufacturer. Maybe you have an electric bike, then you will have a charger for your battery, but you don't have any information what the software running the charger is effectively doing to your battery. It could well be, for instance, that once the two years of warranty are over, the battery charger stops charging, or charges less effectively than truly possible. And that is all because we are not allowed to inspect the software. And I think that is a gross mistake. We should strongly fight for the right for inspection of software that influences us and the things we own. The electric bike, the battery charger, the espresso machine, whatever.

**Barbara** *If people knew exactly what the software was doing, that would introduce a new safety risk, wouldn't it?*

**Holger** That may well be the case, for instance users may then want to customize it to their needs. And that is where the research part comes in. As an example, it would be good to have some sort of open-source software for battery charging that is configurable. If someone then wants to change the charging behaviour so that it only charges to 70% instead of 100% because of the desire to extend battery life, then so be it. Now if configurability is without limits, then an erroneous reconfiguration could lead to a fire accident, since the charger may then overcharge and overheat the battery. Now, to prevent that, it would be nice to have, with the open software, a verification technology that the user can submit the reconfigured code to, for the purpose of providing a proven guarantee stating that the relevant safety limits are adhered to. My ERC grant POWVER [7] has put a focus on these kinds of questions. Still, there are very interesting technical research questions associated with this that are still wide open.

**Barbara** *In your opinion, would the push for open-source AI increase or decrease the potential for attack?*

**Holger** I do think that open source is a good way to enable finding all kinds of issues and reporting them. I do think code secrecy and obfuscation is no good strategy to prevent attacks, while open-source software potentially is. And if you ask me about attacks, I am most interested in attacks that are already in the system, like with the diesel emission scandal. There was no separate attacker. It was the original equipment manufacturer who decided to build in, into the software, elements that were against the interest of the consumers and of society at large. So that seems technically a bit lame because there is no loophole that was attacked. Still, as mentioned, it was extremely difficult to detect and pinpoint the problem, while at the same time having caused premature deaths of thousands of European citizens [9].

> "The AI Act has some chances to become a blueprint for other jurisdictions outside Europe, and if that happens, then much of the dystopian effects currently dominating the discussion might get under control."

In Germany, we are having the “Kraftfahrt-Bundesamt” as the legal entity that is entitled and supposed to investigate these aspects of the automotive industry, but their experts don't know much about software analysis and verification. They can do exhaust emission measurements, but that is by far not enough expertise in face of the massive cases of fraud we have seen.

**Barbara** *What is your personal vision of AI?*

**Holger** I like the fact that the European AI Act aims at regulating the use of AI [8]. With about 450 pages, the document is a burdensome read, but the parts that relate to the daily work of everyday software and data engineers are much less. As mentioned, we have just finished a document aiming at helping the everyday programmer in navigating the Act [3], by identifying the relevant parts of the Act and including a discussion what, according to its stipulations, actually falls under the term “AI”. The AI Act has some chances to become a blueprint for other jurisdictions outside Europe, and if that happens, then much of the dystopian effects currently dominating the discussion might get under control.

At the same time, there will be a price to pay by the AI software industry, namely more regulatory burden, the need for better documentation of processes and products, and a trend towards standardization of effective testing and validation methods. And this will, I think, implicitly lead to an improvement of the quality of processes that are used to design modern software. Other than that, regarding the prospect of the very modern machine learning advances, I don't have grand emotions – I mean, I don't believe in "Wow, we will have super intelligence and that's the future," and so forth. The wave of stunning results achieved by generative AI is currently creating an amazing and impressive hype. Yet, there have been so many hypes before in computer science. Always the same synopsis: "Ah, wow, that hype is larger than any other hype before." And again, we are now in a situation where the hype is much larger than any hype before. And for sure there will be yet another hype and yet another hype and yet another hype larger than any hype before.

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

**Holger** No, except for saying thank you for the interview.

**Barbara** *Thank you, Holger, for your time and perspective on AI and its future. Have a great day!*

**Holger** It was a pleasure.

## References

1. Sebastian Biewer, Kevin Baum, Sarah Sterz, Holger Hermanns, Sven Hetmank, Markus Langer, Anne Lauber-Rönsberg, Franz Lehr. Software doping analysis for human oversight. In FMSD (2024), `https://doi.org/10.1007/s10703-024-00445-2`.

2. Sebastian Biewer, Pedro R. D'Argenio, Holger Hermanns. Doping Tests for Cyber-physical Systems. ACM Trans. Model. Comput. Simul. 31 (3), 16:1-16:27 (2021). https://doi.org/10.1145/3449354.
3. Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Sarah Sterz, Hanwei Zhang. AI Act for the Working Programmer. Submitted. (June 2024).
4. Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, Sarah Sterz. On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. RE 2021 Workshops, IEEE (2021). https://doi.org/10.1109/REW53955.2021.00031.
5. Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, Kevin Baum. What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296:103473 (2021). https://doi.org/10.1016/j.artint.2021.103473.
6. Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Markus Langer. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In ACM FAccT 2024, ACM (June 2024). https://doi.org/10.1145/3630106.3659051.
7. Power to the People. Verified. An ERC Advanced Grant. https://www.powver.org.
8. Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Commission (May 2024), https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf.
9. Toxic particle linked to diesel kills 6,000 a year in Germany. Reuters (March 2018), https://www.reuters.com/article/us-germany-emissions-health/toxic-particle-linked-to-diesel-kills-6000-a-year-in-germany-agency-idUSKCN1GK1UF/.