Let's Talk AI with Kevin Baum

Kevin Baum[1,2] and Barbara Steffen[3]

[1] German Research Center for Artificial Intelligence (DFKI), Department of Neuro-Mechanistic Modeling, Germany
[2] Center for European Research in Trusted Artificial Intelligence (CERTAIN), Germany,
`kevin.baum@dfki.de`
[3] METAFrame Technologies GmbH,
`barbara.steffen@metaframe.de`

"Building trustworthy AI comes with numerous challenges, ranging from robustness and fairness to explainability for effective human oversight and responsible decision-making. Interdisciplinary collaboration is key for tackling these challenges – fortunately, as the AI community grows, finding shared understanding and common ground between relevant fields becomes easier, because more and more researchers with interdisciplinary backgrounds are entering the field. This paves the way for responsible AI development."

The Interviewee - Kevin Baum

My Personal AI Mission:
As a philosopher and computer scientist, I am driven to advance responsible AI development by promoting interdisciplinary dialogue and integrating ethical considerations into core research practices. My mission is to create an environment where appropriate trust assessments in AI becomes the norm, ensuring technology serves humanity in a just and responsible manner.

My Takes on AI

Artificial Intelligence: Defining 'artificial intelligence' precisely is difficult. Perhaps it's best understood when looking at concrete examples and applications. AI as a discipline or field of study encompasses the development of intelligent agents, systems capable of reasoning, learning, and autonomous action (in a technical sense, meaning they operate without direct human control). I also consider the ethical and societal aspects of these developments in order to ensure responsible and beneficial applications of AI for humanity to be part of the field. However, it's crucial to remember that 'AI' often serves as an umbrella term for diverse software and cyberphysical systems with varying capabilities.

Trust: I'd put it that way: Trust is a relational disposition that involves vulnerability, a willingness to put oneself at risk based on positive expectations of another's agent's behavior. Some argue that (current) AI systems fail to be appropriate objects of trust, i.e., that they are the wrong kind of agents, agents that cannot be trustworthy. If so, trust in AI would be generally misguided. I disagree. While I admit that AI cannot be trustworthy in the sense human agents can, I think we can trust in AI systems much in the same way as we can trust in organizations or institutions, which involves attributing properties like benevolence, integrity, and ability to them – this is possible in theory and in practice in a meaningful and proper sense.

Explainability: When people say they are researching explainability (or "explainable AI"), they generally mean that they are working on methods to explain how and why a system, whose decision-making processes are otherwise opaque to humans, arrives at its decisions. The point is not to find out whether the output is correct or appropriate (that would be a question of justification), although explainability can help with this in many cases. Explainability should also be distinguished from various other perspicuity properties like, for example, transparency, which is about making aspects such as the system's formal properties, the data used to train it, and its role in some decision procedure available to a third party.

Essential Elements of Human Capabilities: Phew, I don't think I can offer a serious explication of this without straying too far afield. Perhaps the most important are context sensitivity and everyday understanding, including sensitivity to exceptional and marginal cases of rules (including informal non-monotonic reasoning), empathy, and sentience.

The Interview

Barbara *Thank you, Kevin Baum, for taking the time to do this interview. Can you briefly introduce yourself and your personal relationship to artificial intelligence?*

Kevin As a philosopher and computer scientist at the German Research Center for Artificial Intelligence (DFKI), I deeply engage with both the technical and ethical dimensions of AI. My roles include being the deputy head of the Neuro-Mechanistic Modeling department and leading the Center for European Research in Trusted AI (CERTAIN), where my efforts are focused on advancing the development of responsible and understandable AI systems [12]. I extensively teach AI and computer ethics and am involved in several interdisciplinary research projects, particularly on the explainability and trustworthiness of AI. Whether through research or education, AI is central to my work.

> "The fast pace of AI development calls for legal frameworks that can quickly adapt to new technologies and their societal implications."

Barbara *That's fascinating. Can you give examples of specific challenges that you're currently addressing in your AI research?*

Kevin There are several challenges we are currently tackling. For instance, in our project, 'Explainable Intelligent Systems' (EIS), supported by the Volkswagen Stiftung [13], we are primarily focused on understanding how explainable AI (XAI) methods can achieve the expectations and objectives they are often associated with, such as improving trustworthiness, robustness, and fairness in AI systems, and enabling humans to make responsible decisions when acting upon the outputs of such systems. EIS is an interdisciplinary effort that aims to bring together XAI research and computer science with insights from law, psychology, philosophy, and other fields [4, 1, 9]. My contribution centers on exploring these areas from both a technical and an ethical perspective, focusing on the development of XAI methods and examining the philosophical underpinnings of what we expect from AI in terms of ethical and societal impact.

Barbara *When you look at AI from an interdisciplinary perspective, how do you usually start? For example, does everyone start from different angles, or do you start with a research question that defines the focus of the collaboration?*

Kevin In tackling interdisciplinary AI research, our approach has evolved significantly over time. Initially, we often tried to identify a specific research question appealing across all disciplines involved. However, we quickly realized the challenge this posed due to varying interpretations of key terms and concepts like accountability, understandability, trustworthiness, and many more among different fields. Now, my starting point is to first establish some kind of mutual understanding and a shared vocabulary among team members, who bring diverse research backgrounds to the table. This approach has proven more efficient. From this shared platform, we then formulate a research question that guides our collaborative efforts. The process encourages members to explore the question from their disciplinary perspectives and later reconvene to integrate their findings. This method fosters interdisciplinary collaboration, revealing overlaps and intersections that enrich our collective understanding – and maximizes our research output.

Barbara *And how difficult is it to find common ground and understanding on these topics?*

Kevin Finding common ground and understanding in interdisciplinary research has its challenges, but it significantly improves over time, with collaboration and ongoing dialogue. Our experience, both within our projects and observed at this conference, underscores this evolution. Since initiating our interdisciplinary endeavors around 2016-2017, we've noticed a marked increase in the interdisciplinary community's size and engagement. This growth has facilitated easier collaboration across different fields, as more researchers are now engaging in interdisciplinary work. And within our already established research environment, the initial hurdles of establishing a common language and shared objectives have become less daunting over time, thanks to the cumulative experience and the expanding network of researchers committed to this approach.

> "I advocate for the development of structured educational frameworks that emphasize these core topics [like ethics] and their relationships with each other."

Barbara *How do you see the role of trust in AI adoption?*

Kevin That's a big question. Trust in AI is a multifaceted issue that extends far beyond user acceptance or, more generally, AI adoption. It involves a complex network of relationships between humans, machines, and institutions [3, 8]. Our recent research delves into these dynamics, examining how trust and trustworthiness assessments are impacted not only by direct interaction between individuals and AI systems but also how these systems are perceived within the broader societal and institutional context. Factors such as shared experiences, certificates and seals, including the certification processes, and the overall reliability and trustworthiness of several involved institutions play critical roles in shaping trust in AI [7]. Understanding these interconnections is crucial for advancing AI adoption in a way that aligns with users' expectations and societal norms. Essentially, navigating the intricacies of trust requires a comprehensive approach that considers the entire ecosystem, the whole society in which AI operates.

Barbara *And what measures do you think are needed for ethical AI adoption?*

Kevin First, we need more ethical expertise on side of those who develop AI systems. However, in the realm of AI ethics, a rapidly expanding field, we're encountering a paradoxical situation [11]. The pace at which AI technology develops often outstrips the depth of ethical considerations we're able to apply to each new advancement. For instance, the rush to address fairness in different AI applications – from scoring and recommender systems to generative AI – sometimes lacks thorough analysis of the underlying ethical frameworks and their practical implications. This isn't solely an issue within the AI ethics community but is rather exacerbated by the rapid advancement of AI technologies themselves. But how, in light of this pace, can we teach the necessary skills then?

Reflecting on my experience since 2015, when I was motivated by Professor Hermanns to initiate 'Ethics for Nerds' – a course aimed at instilling a foundational ethical understanding in computer science students – it's evident that the challenge has shifted. Initially, the scarcity of established teaching content was a barrier; now, the sheer volume of material necessitates a more structured educational approach. This structured approach is not only essential for preparing future computer scientists but also critical for effective communication with the broader public, including citizens, policymakers, and regulators.

Thus, the measures needed for ethical AI adoption extend beyond slow deliberation to include the development of comprehensive educational frameworks. These frameworks should facilitate deep engagement with ethical principles, tailored to keep pace with technological advancements and accessible to a wide audience. This approach will ensure that as AI continues to evolve, it does so within a context of informed, ethical consideration that benefits society as a whole.

Barbara *You just said that there is a lot of content on ethics today. Is it unstructured and scattered all over the place, or can you see certain patterns that allow you to prioritize and focus on a small set of, say, three topics and their interrelationships? Which need to be considered to create a more holistic and actionable understanding?*

Kevin There are certain patterns emerging, notably around fairness, robustness, and the role of properties like transparency and explainability when it comes to the imperative for responsible decision-making or accountability in AI development, including questions of effective human oversight. Beyond these, we see recurrent high-level themes such as technological solutionism, questions regarding human autonomy, privacy concerns, and the ethical challenges posed by AI's dual-use potential, which prompts significant reflection on the ethical responsibilities of those working in the field [2]. Despite the emergence of these patterns, I think the field suffers from a lack of structured, comprehensive educational content that can guide both current and future practitioners in navigating these complex ethical landscapes. A more organized approach to AI ethics education would not only help in delineating clear priorities for the field but also enhance communication with a broader audience, including policymakers, regulators, and the general public. This structured approach would ideally focus on integrating technical and ethical considerations, thereby facilitating a more holistic understanding of AI's societal impacts. Therefore, in response to the vast and somewhat scattered nature of content in AI ethics, I advocate for the development of structured educational frameworks that emphasize these core topics and their relationships with each other. Such frameworks should aim to equip individuals with the tools needed to address both present and future ethical challenges in AI, ensuring that the field's rapid development is matched by equally robust ethical considerations.

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

Kevin It's challenging to fit this into a one-dimensional scale. But if you forced me to do so, I would lean towards a 7 or 8. This reflects my anticipation of substantial progress in AI's ability to handle complex, generalized problem-solving through the integration of specialized and increasingly multimodal systems, rather than the emergence of a singular, superintelligent AI entity. We're likely to see advancements that significantly surpass current limitations, addressing challenges previously deemed insurmountable by human or individual AI capabilities. While these systems might qualify as artificial general intelligence, I do not believe that we may encounter strong AI systems with self-awareness or consciousness akin to humans in the foreseeable future. While we may witness the creation of AI with capabilities that seem to mimic creativity or multifaceted intelligence, such as composing poetry, generating whole movies, or performing music with exceptional skill, these should not be confused with genuine consciousness, empathy, or emotional understanding.

> "My concerns gravitate towards dystopian scenarios, not due to fears of super intelligence or autonomous AI dominance, but because of more immediate issues like the concentration of power within a few corporations."

The distinction is crucial, not just from a technological standpoint but from an ethical and societal perspective. The integration and application of these advanced AI systems will necessitate careful consideration of their impact on society, employment, privacy, and security. Moreover, the potential for AI to contribute positively to humanity, such as in medical breakthroughs or solving complex environmental challenges, should be balanced against the risks and ethical dilemmas posed by their capabilities. But there seems no reason to believe that, major breakthroughs aside, such systems will qualify as moral patients [6]. Thus, while we edge closer to the upper limits of the scale in terms of technical proficiency and application, the journey demands a concerted focus on ethical governance, public engagement, and interdisciplinary collaboration to ensure that advancements align with societal values and needs as well as the constant caution not to unjustifiably anthropomorphize the systems that are to come.

Barbara *This leads to my next question. So, what is your opinion on the utopian, dystopian spectrum? What do you think is coming and what do we need to be prepared for?*

Kevin Hard question! I find myself navigating the spectrum between dystopian and utopian outcomes with a critical eye. My concerns gravitate towards dystopian scenarios, not due to fears of superintelligence or autonomous AI dominance but because of more immediate issues like the concentration of power within a few corporations. This concentration raises significant risks, including the manipulation of public opinion and the erosion of democratic processes, exacerbated by the lack of transparency in AI research and development. The opaque nature of AI systems, from data usage to the complexity of their architectures, poses a challenge to understanding and regulating these technologies effectively. I am also very concerned about the possible, even foreseeable misuse of these capabilities by state actors. In the near to medium term, I foresee these challenges manifesting in increased surveillance capitalism, with potentially destabilizing effects on societies and democratic institutions.

However, looking beyond these immediate concerns, I believe in the transformative potential of AI to address some of humanity's most pressing issues, such as climate change and healthcare. The key to realizing this potential lies in avoiding the realization of the current dystopian risks, requiring concerted efforts in ethical AI development, transparent research practices, and equitable governance of AI technologies. Ultimately, while my current stance leans towards a cautious approach due to the visible risks, I remain optimistic about the long-term prospects of AI. Achieving a utopian future with AI will demand a proactive stance on addressing ethical and societal challenges head-on, ensuring that AI development is aligned with human values and societal well-being.

Barbara *Looking back on the last few days of this interdisciplinary conference, is there one insight that was particularly interesting or striking to you?*

Kevin Although as a computer scientist and philosopher, I would like to say it was from these two fields, the potential on the side of the law stood out to me. It became evident that the law plays a pivotal role in shaping the future of AI in society. The journey towards a society where we can be sure that AI contributes positively requires more than just ethical guidelines and good will on side of the developers and researchers; it demands hard regulation informed by a deep understanding of the nuances in AI application and its impact [5]. This regulation must navigate the "strategic ambiguities" inherent in AI ethics, bridging the gap between theoretical ethical considerations and practical, enforceable standards.

In this regard, the conference illuminated the inextricable link between interdisciplinary collaboration and the development of effective regulation to address the multifaceted challenges posed by AI technologies – not only to refine and interpret ethical principles but also to ensure that these principles are operationalized in a way that upholds human oversight and societal well-being.

Insofar, the conference reinforced my conviction that progress in AI ethics and regulation cannot occur in silos. Instead, it requires a concerted effort from a broad spectrum of disciplines, all converging towards the creation of legal frameworks that are both robust and adaptive. This interdisciplinary approach is not just beneficial but essential for realizing the full potential of AI in enhancing societal good while mitigating its risks.

Barbara *Would you like to see more interdisciplinary research in this area? For example, to what extent do you think that law needs to be informed by other disciplines to be aware of what is possible and what are the potential impacts on individuals and society as a whole? This could ensure that regulation introduces rules that fit the overall context.*

Kevin In response to the need for more interdisciplinary research, especially in bridging the gap between law, technology, and ethics in AI, I see a path forward that involves both conceptual and structural initiatives. First, the establishment of regular interdisciplinary conferences, similar to what we've experienced here at AISoLA, is essential. These gatherings should not only maintain a high-level discourse but also delve into specific challenges, such as defining and implementing effective human oversight within AI systems. This requires a concerted effort from diverse fields – computer science for developing transparent and explainable AI technologies, psychology to address human factors like automation bias, and sociology to understand the broader societal impacts [10].

> "The pace at which AI technology develops often outstrips the depth of ethical considerations we're able to apply to each new advancement."

Moreover, the fast pace of AI development calls for legal frameworks that can quickly adapt to new technologies and their societal implications. This adaptability hinges on creating an organizational infrastructure capable of continuous evaluation and monitoring of laws, ensuring they are informed by the latest technological advancements and ethical insights. Such projects would not only foster a responsive legal environment but also encourage a deeper, practice-oriented research collaboration across disciplines. Hence, moving beyond the current state requires not just occasional interdisciplinary interactions but a sustained, structured effort to integrate insights from various domains. This approach will ensure that AI development is guided by a comprehensive understanding of ethical, societal, and legal considerations, ultimately leading to regulations that are both effective and reflective of our collective values and goals.

Barbara *And what do you think the AI vision should look like?*

Kevin I advocate to consider a diverse array of visions that collectively aim towards leveraging AI for societal benefit instead of one unique overall vision. However, here is one specific and rather concrete vision: The establishment of a robust non-profit infrastructure dedicated to nurturing AI systems designed to address specific societal challenges – from enhancing urban mobility and educational opportunities to mitigating misinformation by transparent algorithmic curation. Such an infrastructure would enable sustained support for AI projects beyond the typical funding cycles, ensuring their long-term impact on society. This vision, again, underscores the necessity of interdisciplinary collaboration, bridging the gap between technological innovation and societal needs. By fostering a close relationship between researchers, practitioners, and the broader community, we can ensure that AI developments are not only technologically advanced but also ethically grounded and socially beneficial. As we look to the future, the goal should not merely be to advance AI technology in isolation but to integrate these advancements within a framework that prioritizes human well-being and societal progress. It’s about creating an AI ecosystem that is as much about empowering individuals and communities as it is about algorithms and data. This balanced approach to AI development and implementation is what I believe will lead to a better future for all.

Barbara *Would you like to add anything else?*

Kevin Oh, I could go on for hours, but I think what’s most important to me has been said.

Barbara *Thank you, Kevin, for your time and your views on AI, interdisciplinary collaboration, and future developments.*

Kevin No need to thank me! Thanks for having me!

References

1. Baum, K. et al. (2022) ’From responsibility to Reason-Giving explainable artificial intelligence,’ Philosophy & Technology, 35(1). https://doi.org/10.1007/s13347-022-00510-w.
2. Baum, K., Bryson, J., Dignum, F., Dignum, V., Grobelnik, M., Hoos, H., ... & Vinuesa, R. (2023). ’From fear to action: AI governance and opportunities for all,’ Frontiers in Computer Science 5. https://doi.org/10.3389/fcomp.2023.1210421.
3. Henrique, B.M. and Santos, E. (2024) ’Trust in artificial intelligence: Literature review and main path analysis,’ Computers in Human Behavior. Artificial Humans, p. 100043. https://doi.org/10.1016/j.chbah.2024.100043.
4. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021) ’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. https://doi.org/10.1016/j.artint.2021.103473.
5. Lucaj, L., Van Der Smagt, P. and Benbouzid, D. (2023) ’AI regulation is (not) all you need,’ FAccT ’23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3593013.3594079.
6. Moosavi, P. (2023) ’Will intelligent machines become moral patients?,’ Philosophy and Phenomenological Research [Preprint]. https://doi.org/10.1111/phpr.13019.
7. Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2022). A micro and macro perspective on trustworthiness: theoretical underpinnings of the Trustworthiness Assessment Model (TrAM) [Preprint].https://doi.org/10.31234/osf.io/qhwvx.
8. Shen, M.W. (2022) ’Trust in AI: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient,’ arXiv (Cornell University) [Preprint]. https://doi.org/10.48550/arxiv.2202.05302.

9. Sterz, S. et al. (2021) 'Towards Perspicuity Requirements,' 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). `https://doi.org/10.1109/rew53955.2021.00029.`
10. Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P. and Langer, M. (2024). 'On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives,' FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. `https://doi.org/10.1145/3630106.3659051.`
11. Tang, X. et al. (2020) 'The pace of artificial intelligence innovations: speed, talent, and Trial-and-Error,' arXiv (Cornell University) [Preprint]. `https://doi.org/10.48550/arxiv.2009.01812.`
12. `https://certain.dfki.de/`
13. `https://explainable-intelligent.systems/`