Let's Talk AI with Ellen Enkel

Ellen Enkel[1] and Barbara Steffen[2]

[1] Universität Duisburg-Essen, Department of Engineering, Germany,
ellen.enkel@uni-due.de
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"AI is an interdisciplinary topic equally important for academy and practice with the potential to completely change our way of working and living!"

The Interviewee - Ellen Enkel

My Personal AI Mission:
As researcher in innovation and technology management, it is my mission to focus on the positive aspects of new innovation like AI-systems and their potential for social equally and wellbeing.

My Takes on AI

Artificial Intelligence: AI is an interdisciplinary topic equally important for academy and practice with the potential to completely change our way of working and living!

Trust: I call a system (s) trustworthy in a scenario (r), when it satisfies the user's (u) requirements for the scenario. A system is trustworthy when it is trustworthy in all scenarios within its operational design domain.

Explainability: I stick to Wikipedia, which explains AI systems with the ability for humans to retain intellectual control or refer to the methods to achieve this.

Essential Elements of Human Capabilities: Do not have a special definition.

The Interview

Barbara *Today I have the pleasure of interviewing Professor Ellen Enkel from the University of Duisburg-Essen. I would like to start by asking you to introduce yourself and your relationship to artificial intelligence.*

Ellen Thank you, Barbara. As you mentioned, I'm Ellen Enkel. My research specialty is innovation management, and I currently hold the chair of business administration and mobility. At present, I'm focusing on innovation, particularly technology innovation, in the mobility industry. One of the significant aspects in the mobility industry is semi-automated or fully autonomous driving, which necessitates artificial intelligence applications. This is not only for various driving assistance systems but also for other expected features in future cars, such as providing guidance on when and where to drive. All of this will be guided or controlled by artificial intelligence.

"I believe that a lot of AI systems could compensate for our different upbringing, our different exposure to technology, our different education."

I've also worked in other industries, such as the health industry, where artificial intelligence already plays a significant role [4, 3]. This includes analyzing health data to predict future health status or recommend therapies. You might be familiar with Watson, the IBM system, but I've worked on many similar cases.

Barbara *Do you have examples of one or two specific research questions that you are currently working on in artificial intelligence?*

Ellen Yes, I'm currently working with colleagues to clarify the relationship between the trustworthiness of a system and the development of human trust. We're looking at how trust in artificial intelligence develops [7]. For instance, does one need to have technological knowledge, a certain level of education, or even a specific gender or age to trust the system more or less? We're also trying to define what we call perceived trust or perceived trustworthiness. Is the system manipulating the user into believing it's more trustworthy than it actually is?

Barbara *Do you see different concepts and responsibilities when it comes to trust? For example, that the trust an expert should have in the system is a different kind or level of trust than the end user should have. And how would you approach these two different concepts?*

Ellen That's an interesting question. We know about overtrusting and undertrusting systems [2, 1]. From a psychological or sociological perspective, for example, if you have no knowledge of how the system works, you might overtrust it because you can't relate to anything the system does. This can be compared to the data we willingly give to Google and other applications because they provide us with useful information.

On the other hand, we see that, for example, elderly people generally undertrust any system, whether it's equipped with artificial intelligence or not. They tend to trust humans more than algorithms or machines. This could be related to their upbringing, age, or exposure to technology. We're considering both groups [6].

We're also discussing if this relates to a kind of mental model. A mental model, as we describe it, is the human's expectation of how the system will react in a certain environment. For example, if you're in a self-driving car and it makes a left turn when you were expecting a right turn, you might be surprised and afraid because the system is doing something unexpected, causing you to lose trust.

On the other hand, we can also assume that the AI system has a mental model. As an adaptive system, the AI can learn from the user's behavior and anticipate the user's reactions and the level of autonomy the system can exhibit. If the system knows that unexpected driving behavior frightens the user, it can explain its behavior in advance. For instance, it could verbally explain that it's taking a different route due to a traffic jam or an accident, which can increase the user's trust. So, we're dealing with a mental system of behavior expectations, both from the system to the human and vice versa.

Barbara *That's interesting. I know that there are often already very different expectations in the interaction between just two people. How would you deal with that subjectivity? You can't anticipate what the other person is expecting at that moment.*

Ellen That's another interesting question. We're trying to identify criteria or factors to assess trustworthiness and trust in different situations, such as before, during, and after interaction with the system. A common measure of trust is eye tracking. For example, if the eyes are focused forward, the user is calm [5]. You can also measure factors like heart rate or certain elements in the blood. You can tell when a user is calm and when something happens that makes the user afraid or less confident than before. However, the problem is that we don't know for sure why this happens and how to deal with it. Ideally, we could address it by increasing the number of explanations, simplifying the wording, or limiting the system's control so that the human remains in control. But our current challenge in research is to link certain behaviors of not fully trusting the system to certain factors that generate trust to counteract the loss of trust in the human.

"Everything that is very good and can be very useful can also be manipulated in a criminal way."

Barbara *One challenge with AI applications is that many of them are quite generic. So, I might understand to some extent how this AI application works, but the accuracy of its answers varies from topic to topic and from time to time. Some answers are great, and some answers have a lot of meaning that doesn't fit exactly. How would you deal with that? That it's not just a question of how much I can trust the system, but it depends a lot on what I'm using it for, and*

the correctness of its answers. These are all aspects I can only understand if I have enough knowledge in those areas to be able to evaluate it.

It is the same with human experts. If I were to go to a doctor, I probably wouldn't ask for financial advice. So, we generally don't use specialized experts for their general intelligence. Instead, I know that if I want to learn about innovation, I should come to you. I go to a doctor for a diagnosis, etc. And of course, we all have more knowledge and opinions, but we would probably be a little more skeptical of advice that goes beyond the expert's area of expertise. But with AI tools, we don't really understand that distinction and their area of expertise yet. What are the areas of expertise of the different AI tools? Where can we trust them? And where does their expertise blur? Do you address this in terms of expectations and trust?

Ellen I think you're mostly referring to transparency. This includes data transparency [8]. The EU guidelines strongly recommend enhancing the transparency of data as one of the factors for trustworthy AI. This means showing where the system learned how to behave, and what data was collected to make it proficient in a certain area. This allows the user to understand if the system is experienced enough to provide solid advice or answers [9]. However, we have users with different levels of education and exposure to technology. We also have different ideas about how open and how tolerant we should be of human and system errors. There are very individual factors related to our upbringing, education, and daily exposure that influence how we perceive the system's responses and how we evaluate the system's area of expertise.

"The more human-like the interaction [with AI], the easier it is to gain the human's trust."

I wouldn't say that we should develop every system for every kind of user. I'm very concerned that right now we are mostly developing systems for experienced and advanced users as the systems are developed by developers. And they are professionals in the field. Thus, for any developer, it is difficult to switch positions and see the solution from a user's perspective with no technical background. For example, it's very difficult for an automotive engineer to put himself in the position of a user with little technical knowledge who is easily overwhelmed with understanding the technical functions. If you read the manuals of any kind of technical system, you will see that they often haven't thought about a normal user. They always talk to other experts, maybe an expert in a different field, but still an expert, with whom they can converse in expert language.

It is difficult to configure the system so that is easy to understand and use by all kinds of users, including users who aren't experts, who are from a different social status or different income class and so forth. If we can manage this, I believe that a lot of AI systems could compensate for our different upbringing, our different exposure to technology, our different education. They could support social equality if you do it right. But currently I don't see that as a core interest

of the developers. Typically, people with lower incomes and less exposure to technology aren't able nor willing to pay for expensive systems. So that might be something that we should work on in the future.

Barbara *Yes, that's interesting. In a lot of the interviews with leading experts from AI companies, most of which also have a free version, you hear the claim that the companies want to provide a personal assistant for everyone. The intention is to narrow the gap that exists today. But of course, making these AI tools available is not enough. There's also the question of whether users have an early adopter mindset. Have they heard about AI, its lever, and how to interact with it in a useful and sufficiently skeptical way? This again depends a lot on how much they interact with these AI systems and whether they seek advice from experts, etc.*

Ellen Exactly. Let me add something, only something minor. We spoke before about over- and under-trust, along this line, maybe it's a good thing when the systems are first used by experienced people with a little bit more knowledge on the system and a little bit more experience with technology. So maybe it helps to improve the system so that in the next step, people with less experience that are generally overtrusting the system because they don't have any technical knowledge about how the system works can start trusting the system because it's also proven by experts that it is trustworthy or reliable.

Barbara *Yes. How do you feel about the development that we are now interacting with these tools in natural language? Doesn't it make it harder to maintain a certain distance and skepticism now that it's so easy and intuitive to interact with these tools?*

Ellen You know, I don't have a strong opinion on it being good or bad. I see the advantages and disadvantages. The more human-like the interaction, the easier it is to gain the human's trust. There are a lot of scenarios, for example with elderly people, where I see advantages in using natural language. On the other hand, I'm aware that it can be used for manipulation. So where is the threshold where we as users don't understand that we're talking to a system and not another human being? Think about getting a call and you really don't know if it's a human or a deep fake from an AI-based system. Everything that is very good and can be very useful can also be manipulated in a criminal way. And that is something that I think everyone is going to be afraid of.

Barbara *Do you have any key measures in mind to ensure the ethical use of AI?*

Ellen In the EU guidelines, ethical guidelines are an important aspect of assessing the trustworthiness of an AI system. I know that ethical behavior can be very different for different stakeholders, for different religions, and so on. So, what we perceive ethical can be considered unethical in other parts of the world or in other stakeholder groups. So, it's very, very difficult to generalize ethical behavior. I would say examples of ethical measures are ensuring that your personal data is secure and that no one is trying to manipulate you. These are things

that I would call uniquely ethical behavior, whereas other things, manipulating people to establish a personal relationship, can be perceived as unethical. So, if you give the AI, let's say, a female voice instead of a male voice, that can be disturbing to some people from certain religions and so on. So, I'm very shy to point out five factors that are clearly ethical or unethical. But I think there are some things that are generalizable. For example, security, privacy, diversity, gender equality, and so forth, these should be common, but they are not necessarily.

Barbara *Looking to the future, on a scale of one to ten, where one refers to today's AI tools like ChatGPT and ten to artificial general intelligence. What do you think are the possible future capabilities of AI?*

Ellen Okay. Generally, you know, I think that in certain areas, for example, if you're thinking about chatbots and so forth, like understanding language and giving appropriate responses, we're already quite far. If we think about ChatGPT for example, as you pointed out before, an uneducated or less experienced person doesn't see that there are still a lot of flaws in it. So, if I'm thinking about an expert level assessment, I think that the integration of these systems on a very wide scale has nothing to do with the technological development of the system, because we're very good at that. I don't see us stopping at a certain point. And I see that there is the development of self-taught systems and so on. At a certain point, the human doesn't need to interfere anymore because the systems evolve themselves. I see that a major problem in integrating the system is how much we are going to allow the system to interfere and what areas of our lives should be safe from these systems taking over control. An AI-based system should support our decision making and therefore make our lives easier, instead of manipulating us in a certain direction like voting for a certain political party. AI will be integrated into every industry, used by nearly every company in the world, and will dramatically change every aspect of our lives, just like digitalization did.

"As an innovation manager, I never see the technical limit of an innovation. It's always the human who limits the innovation."

On the other hand, I think we are lacking behind in preparing people to work with AI systems. I think no one in our school education or in our professional education or upbringing is teaching us how to interact with these systems, where to be cautious, when to use it, how to use it, and so on. I think the main obstacle at the moment is the human being, because we haven't prepared the world for all the possibilities of AI systems. And at the moment the development in AI is accelerating quite a bit because it's being discussed publicly in the press and everyone knows or uses ChatGPT.

Barbara *If you were to just evaluate the technical possibilities without considering self-imposed limitations due to safety or security concerns. Do you think something like artificial general intelligence or super intelligence is technically possible?*

Ellen As an innovation manager, I never see the technical limit of an innovation. It's always the human who limits the innovation. Think about all the weapons and bombs, you know, we can do endless mass destruction, but fortunately at least some people think that we shouldn't kill the world's population. So, I think it will be the same with AI. AI is useful, and it can be implemented in many areas of our lives, and it can make life easier for all kinds of groups. On the other hand, I think that when we reach a certain threshold or limit, hopefully the government or smart people will step in and ask question whether we should really take that next step, because from an ethical, legal, or safety point of view, we shouldn't go any further. I firmly believe, that humans should stay in control and technology should help ease our lives.

Barbara *Do you think there's a difference between being in control and feeling in control? For example, we already observed in past innovations that people tend to outsource more and more competencies to technologies. So, if I use AI systems to help me diagnose patients or make a decision as a judge, etc., and I start to notice that 90% or 95% of the time the suggestions are very good. Over time, I've become less skeptical of the AI system and started to trust it more and more. I find that AI makes my job easier. It takes less time, effort and thought to get reasonably good results. So, while I still want to be the expert, stay in control, and be considered, treated, and paid like an expert, I may unconsciously transfer more and more power to the AI system. Who minds if the AI makes their lives easier as long as they retain the benefits of their current roles?*

Ellen If you look at Watson, which is a scenario that you just described, the problem for IBM was that doctors didn't accept Watson because it compensated and evaluated their work. It made their diagnosis, their time spent with the patient much more transparent. So, one step before taking over control or trusting the system, the doctors didn't even want to make their interaction with the patients transparent to the health insurance, like how much time they spend with patients and what therapy they prescribe. They didn't want to be evaluated, compared, or lose their independence in decision making.

"I think the main obstacle at the moment is the human being, because we haven't prepared the world for all the possibilities of AI systems."

But as I described before, with overtrusting and undertrusting. As humans, when we see that a system like Google Maps is valuable to us, we become more and more dependent on that system. We rely on these systems. So, this is the scenario you described. Is there something that we trust so much that we lose the skills or abilities to do it ourselves? Yes, that will happen. And it doesn't necessarily have to be a bad thing because it gives us more time to do other things.

Barbara *What are relevant areas for interdisciplinary collaboration in the context of AI?*

Ellen Everything. You know, I don't think AI can really be developed or understood from a disciplinary point of view. The system developer doesn't necessary think about the influence and impact of the system on a social group, or think about the fact that the user experience is highly dependent on their prior education and experience. And from a psychological or sociological point of view, we don't think about the system itself. We think about how the person receives the system, what is done to the person by using the system, and so on. So, I think everything in the area of AI is an interdisciplinary field. I don't see a single part where it could be purely disciplinary.

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

Ellen I really like the movie Terminator, but I wouldn't like to live in a world totally controlled by AI. I think there are certain areas where it's fine for me to give up control, for example, cleaning the house. In other areas, like educating my children, I wouldn't like the AI to have any influence at all. I want to be in total control. So, my future vision for AI would be to be able to decide on an individual basis, not on a nationality or social group basis, which areas of our lives should be heavily influenced by AI-based systems and which areas of our lives should be less influenced by AI-based systems. And I think the decision should be made at the individual level. It should be transparent where an AI system is in place and being used and where the human with his experience is. I think it is best to let the individual decide how much AI and in what areas they want to integrate it into their lives.

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

Ellen It was a very, very nice interview with you. Thank you very much.

Barbara *Thank you very much, Ellen, for your time and insights, especially from the perspective of innovation and innovation management. Have a great day.*

Ellen Thank you. You too.

References

1. Aroyo, A. M., De Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., ... & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. Paladyn, Journal of Behavioral Robotics, 12(1), 423-436.
2. De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. International journal of social robotics, 12(2), 459-478.
3. Enkel, E. (2017): To get consumers to trust AI, show them its benefits, Harvard Business Review Blog, 17 April 2017. `https://hbr.org/2017/04/to-get-consumers-to-trust-ai-show-them-its-benefits.`
4. Hengstler, M.; Enkel, E. & Duelli, S. (2016): Applied artificial intelligence and trust – the case of autonomous vehicles and medical assistance devices. Technological

Forecasting and Social Change, 105: 105–120. https://doi.org/10.1016/j.techfore.2015.12.014
5. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5, 1096257.
6. Lai-Chong Law, E., van As, N., & Følstad, A. (2023). Effects of Prior Experience, Gender, and Age on Trust in a Banking Chatbot with(out) Breakdown and Repair. In Proceedings of 19th International Conference of Technical Committee 13 (Human- Computer Interaction) of IFIP (International Federation for Information Processing). https://doi.org/10.1007/978-3-031-42283-6_16
7. Liebherr, M.; Enkel, E. Law, E. L., Mousavi, M. R.; Sammartino, M.: Dynamic Interaction of Trust and Trustworthiness in AI-Enabled Systems. Special issue on Trust and Trustworthiness in Autonomous Systems International Journal of Software Tools for Technology Transfer (JSTTT), 2024 (forthcoming).
8. Winfield et al., IEEE P7001: A Proposed Standard on Transparency. Volume 8 - 2021 | https://doi.org/10.3389/frobt.2021.665729.
9. R. V. Zicari et al., "Z-Inspection®: A Process to Assess Trustworthy AI," in IEEE Transactions on Technology and Society, vol. 2, no. 2, pp. 83-97, June 2021, doi: 10.1109/TTS.2021.3066209.