Let's Talk AI with Edward A. Lee

Edward A. Lee[1] and Barbara Steffen[2]

[1] UC Berkeley, Department of Electrical Engineering and Computer Sciences, USA,
eal@berkeley.edu
[2] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"The capabilities of the large language models that merged in public in 2022 surprised nearly everybody in the field. I believe this new technology will prove as transformative as any humans have ever devised, with highly unpredictable effects on our culture. By means of token prediction, the machines have acquired the ability to summarize large bodies of knowledge, to reason logically about numbers and mathematics, and to create novel work influenced by prior human work. Many observers have been quick to criticize, pointing out errors in reasoning and fabrications presented as facts, ignoring the remarkable reasoning abilities that emerged from token prediction and the relationship between fabrication and creativity. This new technology offers insights into how human cognition could have emerged and how it works, particularly since the deficiencies identified by the critics are remarkably human-like (we also make errors in logical reasoning and fabricate facts)."

The Interviewee - Edward A. Lee

My Personal AI Mission:
I believe that the recent revolution in AI can teach us a great deal about human cognition. The fact that so many capabilities that we consider fundamental, such as the ability to use logic for deductive reasoning, appear to emerge from token prediction is remarkable. It suggests that the human brain may be fundamentally a prediction engine, and things that we consider fundamental, such as logic, may be just mechanisms that enhance prediction.

My Takes on AI

Artificial Intelligence: Machine behavior that resembles human behavior and that would generally be considered signs of intelligence.

Trust: Willingness to permit another entity or individual to take actions that could cause harm with confidence that the actions taken will not cause harm.

Explainability: Something is explainable if it is possible to provide a human-understandable sequence of rational deductions that lead to that something.

Essential Elements of Human Capabilities: Creativity, intuition, feelings, introspection, and reasoning.

The Interview

Barbara *Welcome, Professor Edward Lee. Thank you for joining me for this interview at the AISoLA conference. Could you start by briefly introducing yourself and your relation to artificial intelligence?*

Edward A. Sure, Barbara. I am a professor at Berkeley, where I've been teaching for over 30 years. I am an engineer and a computer scientist, specializing in cyber-physical systems, which involve both computing and physical processes. I am particularly interested in AIs that are embedded in robots. Most of my career has been highly technical, but I have written two general-audience books that step back and examine the big picture of technology from a more philosophical and societal perspective. These are, The Co-evolution: The Entwined Futures and Humans and Machines (2020) [3] and Plato and the Nerd: The Creative Partnership of Humans and Technology (2017) [2]. I also had the pleasure of co-edited a volume on Digital Humanism [8] with many excellent essays by top experts in the field. All of these are available open access.

"[...] one of the things that amazes me about large language models is that the ability to reason logically and think about numbers emerged from token prediction. I think there's a lot of potential [here] to gain insight into how humans have developed our ability to reason and think logically."

Barbara *That sounds intriguing. Could you elaborate on the specific challenges of AI that your research addresses?*

Edward A. My research might seem a bit eclectic. I have two complementary lines of work. My technical work is tangentially related to AI. However, about eight years ago, I started focusing on issues around technology and society and the philosophy of technology. I've written a couple of books on that topic [2, 3], which was a transformative learning experience that pushed me to learn about other disciplines.

Barbara *Interesting. How do you perceive the role of trust in AI adoption, and what measures do you believe are essential to ensure ethical AI use in the future?*

Edward A. I'm skeptical that there are measures that will guarantee ethical use of AI in the future. We will inevitably see a variety of uses, as we've always seen with any powerful technology. I think humans are the more concerning part of the equation for me [4]. Humans have a rather grim track record of using technology against one another. I'm fairly certain that AI won't be an exception, and humans will find creative ways to use it against each other.

"[...] if the question is whether we have AIs that exceed human capabilities, we certainly do."

Barbara *What are your thoughts on the idea of making large language models available open-source? Do you find that risky?*

Edward A. That's a thought-provoking question. I'm a strong believer in open source as it enables the exploration of technology for a wide range of applications, both good and bad. I believe many potential positive applications are enabled by open sourcing these AIs, which might not otherwise be possible due to lack of commercial viability. Moreover, I have argued before that it is an illusion that we humans have much control over the trajectory of the technology [3, 4]. Keeping the mechanisms hidden is probably a fool's errand. So, despite the risks, I'm very much in favor of making these AIs open source.

"I'm skeptical that there are measures that will guarantee ethical use of AI in the future."

Barbara *Regarding the future capabilities of AI, on a scale from 1 to 10, where 1 represents current dedicated AI systems solving specific problems, like ChatGPT or DALL-E, and 10 represents artificial general intelligence systems surpassing human capabilities. What do you think will be possible and what should we prepare for?*

Edward A. Honestly, I believe we already have the whole range, from 1 to 10. I'm not fond of the term artificial general intelligence, but if the question is whether we have AIs that exceed human capabilities, we certainly do. For instance, if you interact enough with ChatGPT, its breadth of knowledge is something no human can match. But in some ways, this is not new with technology. Every useful technology is beneficial because it exceeds human capabilities in some way. We've always used technology as an intellectual and physical prosthesis, and I think AI will be no different.

"Currently, the key difference between ChatGPT and human cognition is that the AIs are not embodied."

Barbara *Do you think it becomes even more concerning if we start integrating AI into robots which then start to move in the real world where we also operate?*

Edward A. I believe that will probably be the next significant phase in the development of these large neural network- based models. The term people use for this is embodied robotics. Currently, the key difference between ChatGPT and human cognition is that the AIs are not embodied [6]. They don't have a body to interact with the physical world. However, that's going to change rather quickly. Many people are working on applying this technology in robotics. I find it both scary and exciting. I think we are likely going to see robots that are extremely adept at manipulating things, which has been a significant challenge in robotics.

Barbara *Looking into the future and the potential impact AI will have, where do you see yourself on the utopian- dystopian spectrum that is often discussed in public?*

Edward A. My view is that things are going to change, and there's no question about that. This technology will affect our culture in very unexpected ways. It will change the role of humans and how we interact with our world. It's hard to predict how. We need to make every effort to ensure that we function synergistically with this technology. I've been involved in an initiative called the Digital Humanism Initiative [8, 9], which focuses on how we can keep the interests of humans at the forefront of the evolution of technology and the changes in human culture that come with it. It's a tremendously challenging problem.

"[As numerous AI-generated] papers are used to train the next generation of AIs, a feedback loop emerges that can result in AIs whose knowledge base is largely fabricated."

Barbara *Could you give examples of specific challenges you're currently addressing in this group?*

Edward A. One particularly striking challenge is regulating AI. It's a tremendous challenge. It's hard to even define the terminology needed to create legal constructs to work with this technology. I believe we need to put some effort into figuring out how to do this because any powerful technology requires societal control and regulation. This is no exception, and we don't know how to do it currently.

Barbara *How do you view the challenge that technology-push often brings us into settings in which we are confronted with new challenges for which we do not have suitable regulation yet. It is the nature of this kind of progress that regulation always lags behind as technology first needs to impact society before we can find ways to regulate it.*

Edward A. I can't give you a definitive answer to that question as I'm not a public policy person or a legal scholar. I respect people who are tackling those problems. I see my role as helping them understand the technology better so they can be more realistic about how it's going to function in society and what the possible risks and benefits are.

Barbara *Are there specific challenges or research questions you think we should tackle together in an interdisciplinary fashion? If so, which disciplines would be suitable in your opinion?*

Edward A. There are many opportunities. I'm personally excited about interacting with people in psychology. I think there's a lot to learn about human cognition by observing how AIs have evolved and are changing [5]. For instance, one of the things that amazes me about large language models is that the ability to reason logically and think about numbers emerged from token prediction [1]. I think there's a lot of potential to gain insight into how humans have developed the ability to reason and think logically by observing how AIs have developed similar abilities from language models.

Barbara *From your perspective, what is your vision for AI that we as a society or people should tackle in the future?*

Edward A. That's an extremely broad question. I would like to see the term AI changed to IA, which stands for intelligence augmentation. I would like us to work with machines in a synergistic way, using them as cognitive enhancers to improve our abilities. I hope we can use them to improve our research in medicine, address climate change, and make our society fairer. The AIs learn human biases and prejudices, but they also expose them. We can use this to better understand our culture and maybe find ways to mitigate these problems. I'm optimistic about the positive uses of AI. However, I also acknowledge the potential for negative uses. As a society, we will have to be proactive about curbing these uses. We may have to be reactive in some cases. When bad things happen, let's adjust and try to correct the course as much as possible. One challenge that I would like to highlight concerns what happens when more of the data used to train AIs is generated by the AIs themselves. Today, the AIs are trained mostly with human-generated data. But it seems inevitable that that will change. Even this interview has been edited by an AI and will become training data for the next generation. More seriously, recent studies show an increasing number of sham academic papers, which are written largely by AIs, being published [7]. As these papers are used to train the next generation of AIs, a feedback loop emerges that can result in AIs whose knowledge base is largely fabricated. If, while this happens, society gives more trust and responsibility to the machines, we could end up in a very bad place.

"Humans have a rather grim track record of using technology against one another. I'm fairly certain that AI won't be an exception, and humans will find creative ways to use it against each other."

Barbara *Is there anything else you would like to add to this interview?*

Edward A. Perhaps just a comment that the group of people brought together by this conference (AISoLA, 2023) is exactly what we should be doing more of. We have a mix of computer scientists, psychologists, philosophers, and historians. I believe these cross-discipline interactions are essential. The emergence of these neural network-based AIs is somewhat new for computer science because their behavior is harder to explain and understand than most of what computer science has dealt with. These other disciplines are more accustomed to dealing with complex systems. They have methodologies that are new to computer scientists that we could learn from. I believe conferences like this really help with that.

"One particularly striking challenge is regulating AI. [...] It's hard to even define the terminology needed to create legal constructs to work with this technology."

Barbara *Do you think it would be useful to derive a few questions that seem critical to AI and AI development and then ask different disciplines to work on those questions collaboratively?*

Edward A. There's a lot of potential there. I still see many gaps. I hear thoughtful ideas that I can immediately recognize won't work because that's not how the AIs work. I'm sure they hear ideas from me that they know won't work because I don't understand societal systems the way they do. The only way we can close those gaps is by getting people to talk to each other.

Barbara *Thank you very much for your time, Edward, and for your insights. I wish you a great time at AISoLA and hope you enjoy the conference and its interdisciplinary discussions.*

Edward A. Thank you very much. I appreciate it.

References

1. Bubeck, S., et al. (2023). "Sparks of Artificial General Intelligence: Early experiments with GPT- 4," arXiv:2303.12712v1, cs.CL https://doi.org/10.48550/arXiv.2303.12712.
2. Lee, E. A. (2017). Plato and the Nerd. The Creative Partnership of Humans and Technology, MIT Press.
3. Lee, E. A. (2020). The Coevolution: The Entwined Futures of Humans and Machines. Cambridge, MA, MIT Press.
4. E. A. Lee (2021). "Are We in Control?", in H. Werthner et al. (eds.), Introduction to Digital Humanism, Springer, 2021, https://doi.org/10.1007/978-3-031-45304-5_11
5. E. A. Lee (2022). "What Can Deep Neural Networks Teach Us About Embodied Bounded Rationality," Frontiers in Psychology, v. 25, https://doi.org/10.3389/fpsyg.2022.761808.
6. E. A. Lee (2024). "Deep Neural Networks, Explanations, and Rationality," in B. Steffen (Ed.): AISoLA, LNCS 14380, pp. 11–21, 2024. https://doi.org/10.1007/978-3-031-46002-9_1, 2023
7. McKie, R. (2024). "The situation has become appalling: fake scientific papers push research credibility to crisis point," The Guardian, Feb. 3. https://www.theguardian.com/science/2024/feb/03/the-situation-has-become-appalling-fake-scientific-papers-push-research-credibility-to-crisis-point
8. Werthner, H., W. Prem, E. A. Lee, and C. Ghezzi, Eds., (2021). Perspectives on Digital Humanism, Springer.
9. Werthner, H., C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, and A. Stanger, Eds. (2024). Introduction to Digital Humanism: A Textbook, Springer