AI 4 Science Discovery Network+

AI4SD Interview with Professor Carlos Zednik
30/11/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

21/07/2022

Humans-of-AI4SD:Interview-24

AI4SD Interview with Professor Carlos Zednik
Humans-of-AI4SD:Interview-24
21/07/2022
DOI: 10.5258/SOTON/AI3SD0223
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Interview Details

Title	AI4SD Interview with Professor Carlos Zednik
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	CZ: Professor Carlos Zednik - Eindhoven University of Technology
Interview Location	Online Interview
Dates	30/11/2021

2 Biography

Figure 1: Professor Carlos Zednik

Carlos Zednik: 'AI is a technology that cannot be stopped'

Professor Carlos Zednik's research is focused on the explanation of natural and artificial cognitive systems. He is PI of the DFG-funded project on Generalisability and Simplicity of Mechanistic Explanations in Neuroscience. Before arriving in Eindhoven he was based at the Philosophy-Neuroscience-Cognition program at the University of Magdeburg, and prior to that, at the Institute of Cognitive Science at the University of Osnabrück. He received his PhD from the Indiana University Cognitive Science Program, after receiving a Master's degree in Philosophy of Mind from the University of Warwick and a Bachelor's degree in Computer Science and Philosophy from Cornell University.

In this Humans of AI3SD interview he discusses the importance of explainable AI, the need for transparency, the problem of cross-disciplinary crosstalk and offers his advice for early career researchers.

3 Interview

MP: What's been your path to where you are today?

CZ: I'm an assistant professor at the Technical University in Eindhoven, in the Netherlands, but the path to where I am hasn't been linear. For my undergraduate degrees I studied philosophy and computer science, then I did a Masters at the University of Warwick in philosophy of mind, then I went back to the US to do my PhD in cognitive science, which blended philosophy, psychology and neuroscience. After my PhD, I worked as a postdoc in Germany, where I slowly drifted away from psychology and neuroscience towards the science of natural and artificial intelligence. I started looking at the general questions of philosophy of AI, like "Can machines think?" But then I began looking at questions of transparency and explainability, which is where my research is focused now.

MP: What made you shift towards AI?

CZ: It was partly to do with the hype that was growing around AI about six or seven years ago. I was at the end of one research project and decided to work on something new, which came very naturally to me. I understood the work, what it's doing, why they're doing it, and also some of the challenges around it. People seemed to like my writing on AI, as a philosopher who understands the technology because of my background in computer science.

MP: You spoke about the hype around AI. Do you think it's overhyped?

CZ: I don't think so. Some of the details are overhyped, for example, everybody equates AI with machine learning, and machine learning with deep learning. That's too narrow a focus, and not everything can be solved with machine learning or deep learning. You need AI methods more generally. So in that specific sense, there's maybe too much hype around deep learning but, in general, the hype around AI is understandable. It's a technology that cannot be stopped.

MP:What kind of research are you working on at the moment?

CZ: For the last two to three years, I have been working on a project of explainable AI. It looks at the black box problem, meaning, the problem that current machine learning systems are fairly opaque, we don't really know why they do what they do or how they work. It's not like a car engine that we can open up and see what's broken. We might not know why they are malfunctioning or even why they're functioning so well. This is also true for experts, not only laypeople. Explainable AI tries to tackle this problem and to develop tools to allow you to better understand how these systems work.

We talk about explainable AI, making it transparent, but we often don't know what that means. Traditionally, it's the philosopher's job to look at certain concepts that are important in a certain domain and try to understand exactly and transparently what they mean. I try to help engineers to come up with good accounts of what it means to make a system transparent. So, for example, if you're rejected for a loan application, then maybe you should have a right to know why you were rejected. As long as we don't know what it means to answer that question of why you were rejected, how to explain the decision, then people cannot realise their right to know why certain decisions are made about them. So I'm trying to step in and say: "This might be an appropriate explanation for this kind of situation."

I try to stress the different roles that explanations can play in different contexts. One area that is often overlooked is engineering. Engineers try to develop systems for drug discovery, for example. These explanation tools help them develop better systems because they can be used to analyse the system's performance. Although some people might just be happy having the end product, the drug discovery, and not care why, we still might need explainable AI techniques to promote good and efficient engineering.

MP: In having more explainable AI, do you have to trade off something else?

CZ: That's a big debate right now. It's known as the accuracy interpretability or performance interpretability trade-off. With certain systems, if you make them more easily understandable you reduce complexity, and it might not be able to solve the problem as well as it used to. But that's not always true. There are people who say "We can approximate the same behaviour to an arbitrary degree of precision with a much simpler system," and that's great. What I tend to focus on, however, are tools that are applied post-hoc: we have a very complex system and then we use some kind of tool to provide us with insight into the complexity of the system, and explain how that complexity is used to solve a particular problem.

MP: What challenges do you come up against in your work?

CZ: Explainable AI is a really new discipline, so the research around it is somewhat disorganised. Different research groups are inventing different methods that do similar things, but they will talk about these methods in slightly different ways. There's no theoretical framework that guides the whole research problem or programme of explainable AI, which can lead to crosstalk or people reinventing the wheel in various ways. That's a challenge for the discipline but it's also where a philosopher like me comes in use. I can ask "Can we develop a unified framework? Can we help people not talk past one another, and have a concerted effort to go in one direction or another?"

MP: What kind of ethical issues are being discussed around AI and explainable AI?

CZ: I'm not primarily an ethicist, but I have worked with ethicists in trying to understand which notions of transparency are important: ethical responsibility, for example. I understand that ethics whitewashing is a problem, which we shouldn't condone. If a big tech company, for example, hires ethicists to write a moral code of conduct for the company, which they use to justify morally reprehensible behaviour, then they will say, "We were following our own code of conduct." That, of course, is problematic. I see that there's a need for some kind of external regulation or standardisation.

Any technology that has the kind of impact that AI has should be regulated at least for things like safety and reliability. Companies profit from clear rules because they know what technology they can develop and what the best practice is in developing it. But I also understand the need for innovation and novelty. The general challenge is how to balance regulation with room to innovate. Regulation, however, is slow and, unfortunately, academic involvement from ethicists has sometimes been lacking. There's no easy fix for that problem. The best we can do is educate, and at Eindhoven, we ensure that all our engineers take ethics classes so they are aware of value-sensitive design and sustainability as core principles. These things need to be instilled from the bottom up, so you understand why ethics is important and ethical behaviour is necessary.

MP: What has surprised you in your research?

CZ: One of the things I have been surprised by is the need or desire to involve philosophers in these debates. There is a genuine interest in understanding what philosophers have to say, and this is not limited to ethicists. With the domain of explainable AI, there is a need for epistemologists who ask, for example, how can we acquire knowledge through machine learning systems that we potentially won't understand. There is also some influence from philosophers of science and philosophers of engineering who look at different users or situations in which certain technologies are deployed to understand what needs to be done to make them usable, efficient, reliable, and safe. These are the abstract problems in which philosophers specialise. It's a happy surprise that there's a perceived value in philosophers.

MP: What is the knottiest problem you've encountered in your research?

CZ: One problem I'm working on currently with a PhD student is around causal inference in machine learning systems. So there are many different reasons you might want to explain a machine learning system's behaviour, such as a deep neural network. You might want to predict what it's doing, or you might want to justify its behaviour. But there's another thing you might want to do, which is to intervene in the system in order to improve its behaviour. The main tool that they have at their disposal is retraining: taking new data, or modifying the dataset, or tweaking parameters in the learning process, and then doing the whole thing again. But doing it from scratch is really inefficient.

If we want to fix a problem in a car engine, we don't want to throw out the entire engine and start from scratch. Instead, we look for the source of the problem, we look for one limiting factor and we try to intervene with that particular element to make the desired improvement. What I'm working on with the PhD student is trying to understand how explainable AI can be used to identify the intervention points, and that's really difficult to do. We don't know how to isolate the individual kind of contributing parameters in this complex system. Of course, we can tweak things at random, and that will change the behaviour, but we want a targeted intervention.

So targeted intervention would be desirable, for example, with combating bias in systems. If we know a system is biased, how do we make it unbiased? Right now, all we have is retraining the whole system, which takes a long time and is subject to random variations. But if we go in there and can make just a few tweaks, and now it performs better than before, that's fantastic. We don't have to invest all the time and energy to retrain the whole thing.

MP: What advice would you give to early career researchers?

You need to be aware of the challenges of the academic profession and the limited opportunities for success. There are far fewer professor positions than there are PhD positions. Often what happens, if you're a bright student with good mentors doing interesting research, is that it's fairly easy to get a PhD position. But then you graduate with a PhD and you see that the job market is very challenging, and full of really good candidates. So there is a need to have realistic expectations that even if you have a PhD, you can't assume you'll have a long term career in academia. In many ways I was lucky — sure, I do good work, and people like it, but luck is, unfortunately, part of the game.

So it's always important to have a plan B in mind while you're doing your research, to make preparations around what you will do if the academic career doesn't work out. Try to see

which areas in industry are interested in the kind of work you do, and make contact with people in those industries.

Another piece of advice in my area is that, if you're a philosopher working in a domain that is close to science or engineering, make sure you are in touch with scientists and engineers! Don't just look at it from your philosophical armchair; get your hands dirty, join a lab, collaborate on a software development project. That way, you get a lot of cross-pollination from different angles. People with different disciplinary backgrounds think about problems in different ways, and that can help you think through a problem. You can always profit from learning about a different perspective.

Disclaimer: The opinions raised by Carlos Zednik are his own and do not necessarily reflect the opinions of his employers.