Transcript

21 March 2025,

Interviewer 1:30

OK, alright. So xxx, you've I've assumed since you've signed the….the consent form you've read a bit of the background. I'm just gonna' give you a bit more about my research and … and what it's all about. So, as you know, AI is the big buzzword lately, but it's basically software that does human tasks in various different aspects of society. So, we have AI that can play, Go, game of Go. We have AI that can help a doctor do CT scan analysis. We have AI that say in the legal field helps with recidivism predictions. We have AI that can drive a car. We have AI that can recommend movies to you on Netflix. All of that. So, basically, about 10 years ago. DARPA in America decided. OK, these are black boxes, as in, they're opaque. They're not transparent. Because… in order to do these human tasks, the algorithms that these engineers came up with were stochastic. So back when I went to school, we were very deterministic software programmes. If this, then that.

We could go line by line code by code and audit these… these particular algorithms, most of them that are stochastic are black boxes, so there's a bunch of inputs, the box inside trains it and then out comes the output, and the engineers optimise it by modifying the inputs and looking at the outputs. So, the issue here is they don't explain themselves, so when they're performing human tasks they're acting like agents in the world. I think we need to know what's going on inside them. So this field of explainable AI came into being, and engineers being engineers, built explainable AI for themselves to make their algorithms work rather than looking at, you know, what the general public wants. So what I'm doing is my research is looking at social science knowledge about the … the topic of explanation itself and applying that knowledge to these AI systems, explainable AI systems, to tell engineers, hey, guys, before you even build and explain, will AI model. You don't even know what an explanation is, and you need to find out who you're giving an explanation to as well as what you're explaining before you even start building it. So, as you know, we humans, give and ask for explanations all the time. So it's a universal thing we do, but it's not uniform for example, explanations vary from situation to situation, they vary from person to person. So, let's talk about a varying situation. We have the case of a fire in a house and a fire with a tanker, an oil tanker or a gas tanker. The

explanation for a fire in a house is a lit candle and oxygen. The explanation for a fire with this tanker is high pressure oxygen tank with you know a spark. So different explanations for different situations. If we talk about the same situation but different people, let's talk about a cancer diagnosis. The doctor explaining why to the patient, a layperson is simply you were smoking for X number of years and your lungs have now got cancer. A doctor talking to another doctor would be, oh, these chemicals in this area of the lung impacts these cells, et cetera, et cetera, et cetera. So different explanations even for the same situation. So we are constantly asking why questions right? We're asking why did our friends get divorced? Why did the economy tank? Why did that airplane crash? So because explanation is varied, how do you build an explainable AI model that's universal? You don't, right? You have to build the right one for the context. So what I'm doing is gathering in depth information in this interview on the subject of explanation from the perspective of humans. So, please remember this isn't a test for you. I'm only gathering information about your views, perspectives, thoughts, questions, opinions, opinions about a real-life scenario which I'm about to describe to you. There are no right or wrong answers so, let's take a deep breath, OK?

Stakeholder19_Judge 6:32
Yeah.

Interviewer 6:33
OK. All right. Are you ready for the scenario?
Stakeholder19_Judge 6:37
Sure.

Interviewer 6:38
OK, so we have a real-life case within the AI application of automated vehicles called AVs. These are what you call driverless cars or self driving cars now. It involves the occurrences of actual car crashes involving one particular AV brand called Tesla and its advanced driver assistance system called Autopilot. So, Tesla's Autopilot controls the steering, braking and acceleration functions of the AV without any assistance from the human driver. Furthermore, note that Autopilot could at anytime disengage and hand over the controls to the human driver. OK.
All right. So according to USA NHTSA, which is the National Highway Traffic Safety

Administration, their office of Defects Investigation said between January 2018 and January 2022, so four years, Tesla AVs, with Autopilot engaged were involved in 16 (one-six) crashes where they struck highly visible stationary in road, so on the road or roadside first responder vehicles that were attending to preexisting collision scenes. So these are police, ambulance, fire trucks, road maintenance vehicles, lights flashing, people with highly visible vests on attending to another collision. OK?

Stakeholder19_Judge 8:22
Yeah.

Interviewer 8:23
These Teslas drove up, drove up and crashed into them. And, on average, in these crashes, Autopilot aborted vehicle less than one second prior to the impact.
Stakeholder19_Judge 8:39
Say that again on average.

Interviewer 8:41
So on average, in these 16 crashes….

Stakeholder19_Judge 8:46
Yeah.
Interviewer 8:47
Autopilot aborted vehicle control so it disengaged….
Stakeholder19_Judge 8:52
Yep.

Interviewer 8:53
Less than one second prior to first impact. Remember, there is a human driver behind the wheel, but that human driver, for whatever reason, was not paying attention in these 16 crashes did not take over control.
Stakeholder19_Judge 9:16
OK

Interviewer 9:16

But Autopilot disengaged Less than….

takeholder19_Judge 9:16
But you said within the the scenario is one second before impact.

nterviewer 9:17
Yeah, less than one second before.

Stakeholder19_Judge 9:22
Yeah.

Interviewer 9:23
OK.
Any other questions on the scenario?
Stakeholder19_Judge 9:29
No

Interviewer 9:30
OK. I would give you a link on the chat, but I don't want to go through that.

Stakeholder19_Judge 9:37
Sorry.

Interviewer 9:38
Yeah, it's OK 'cause, the link would show you pictures of these crashes. You know, news reports, all of that. So….

Stakeholder19_Judge 9:44
OK, I think by the way this scenario, this is not a fabricated one. This is an actual?

Interviewer 9:45
Yes.

Stakeholder19_Judge 9:53

I mean the …the ….the facts behind it are accurate. That's what I'm saying.

Interviewer 10:11
Yes, correct. This is not an anecdote. I chose this scenario based on the USA NHTSA report. And it's a real-life case study. It's not anecdotal. It's a series of events that occurred over a four-year period. It's an odd thing like NHTSA has investigated a lot of other crashes involving not just Tesla, but other automated vehicles. And, you know, but this one, in particular, struck out for me. Yeah. Stuck out. Sorry for me so.

Stakeholder19_Judge 10:21
Yeah, yeah.
Yeah. OK.
Interviewer 10:37
OK, so based on the scenario that I've just described, you are seeking explanatory information about these crashes from autopilot. OK, the system that controls steering, braking and acceleration functions of the AV. So let's assume everything else is fine. The hardware, everything else is functioning perfectly on the AV. Let's assume that the driver wasn't paying attention even though he or she should have. Let's assume the visibility was fine, everything. You, from your life … lived experience, professional experience, everything, …. The way you see the world, … When you ask why questions about this like why did these car crashes happen, what kind of information do you have in the back of your head about the decisions that the AV made or didn't make? The tasks that the AV performed or didn't perform?

Stakeholder19_Judge 11:40
My initial reaction or my initial thought, is an issue of trust. And whether or not I trust the autopilot. I trust the algorithms behind it. I trust the design. So my thought process would lead me first of all to say, OK, accepting this is a factual result, then how can I, on a personal basis, trust this autopilot mechanism, if you will? And what I would, I'd be comfortable sitting behind a wheel, and so-called driving by just really being a passenger in this vehicle? I … I don't know that I would put any weight on the driver not paying attention because in my mind one second would not be enough time for that driver to react. So you'd have to go back to the basics. And… and how are these programmes created if you will? And they're looking for the public trust.

But then you have this result so, I have to understand the makeup of the programme itself and trust it. Because frankly, I'm at an age that I don't know enough about the technology. Or whatever thoughts go behind it. I.. I just. I wouldn't be comfortable. So how do I, how do I get myself understand the entire mechanism, so that then I can have a ...a comfort zone that I would not mind getting behind a wheel and say OK, press this button autopilot and in the meantime read a book in a sense?

Interviewer 13:49
OK, so you want an explanation. What kind of an explanation do you want so your trust will go up? So when it comes to the functions that this thing performs, remember it's an agent in the world, right?

Stakeholder19_Judge 14:04
Yeah.

Interviewer 14:04
So if an agent, like a human agent, went about all over America and did this 16 times, would you not ask why questions and what kind of questions would you ask about its decisions?

Stakeholder19_Judge 14:21
About what?
Interviewer 14:23
The decisions that this entity made, what questions do you have about its decisions? Whether you want to ask the question of it or of its designers? What questions do you have about its decisions?

Stakeholder19_Judge 14:26
Yeah. Well, I go ... I go back to what I said a moment ago, I would have to truly understand and fully understand. Not truly, but, ... but fully and truly understand how this whole thing is designed and how it is that the designers create this ... this technology and are comfortable enough to say, OK, this is good for the consumer. This is safe for the general public. What information do they input? How do they create these programmes? To me, it's not to me it's ... it's technology, but I don't understand a lot the algorithms behind it. So I... I quite honestly would have to

understand that and those are the questions that I would ask the designers. It's not the technology. You don't ask the machine, you ask the people that design the machine.

Interviewer 15:45
OK. So the questions for the yeah, sorry, go on.

Stakeholder19_Judge 15:46
And if and if there's a,... if no... if there is a... a way of improving the design or if there is an inherent design flaw. Or if there is a weakness in the system that manifests itself by disengaging one ... one second before reaching a first responder vehicle with visibility with everything, what is it that they have to put in ... to their algorithms to prevent that from happening? I honestly, I don't know that I understand her enough to be able to ask the questions.

Interviewer 16:31
Actually, you'd be surprised you understand more than you realise.
So let's... let's parse let's let's.

Stakeholder19_Judge 16:38
Yeah. You, you, ...you, you... you say that ... is it Unnati by the way?
Interviewer 16:44
Yes, yes.

Stakeholder19_Judge 16:45
OK. You ...you say that Unnati, but I am. I'm feeling like honest to God like a fish out of water.

Interviewer 16:52
Well, guess what? The reason we have explainable AI in this field is because even the designers themselves don't know how these entities came up with the outputs. They came up with based on what they input into it. Remember this. OK, so let's... let's parse your questions because they're extremely valuable questions, and they're right on point. You may not believe this. So it's a black box, even to the designers. OK.

So you're asking, what did you put into it to make it come up with the outputs before you …you decided it was safe enough to go out in the road? That's your question, right?

Stakeholder19_Judge 17:36
Yeah.

Interviewer 17:37
OK. So that's the first question. Excellent. In technical terms, just so you know, these AI engineers would say, oh, what was your training data? OK, So what they do is they come up with these algorithms, they put the… the situation, they put a whole bunch of inputs into it, which is called training data us, you know, and then out it comes with, OK, do this. Don't do that, whatever. And then they say it's accurate. So they go put more training data in or they tweak the outputs and say this is correct. This isn't correct. Go back. And that's what training data is. So you're asking about training data. Excellent question.

Stakeholder19_Judge 17:48
Oh. OK.

Interviewer 18:18
And they're also asking these questions. Here's a question for you. I want to delve deeper into how they trained this system. OK? Before they said it was good enough before they declared it good enough.

Stakeholder19_Judge 18:20
Yeah.
Yeah.

Interviewer 18:34
So let's talk about which functions are you talking about, because if you look at these cars, they have different algorithms, OK? And they have an algorithm for the vision function, and the perception function. So actually seeing the surroundings and processing it and identifying what each thing is like. It's a traffic light, it's an obstacle. It's a human being. It's another car, it's, you know.

Stakeholder19_Judge 19:02
Yeah.

Interviewer 19:03
These cars are trained to do that. That's one algorithm. Another algorithm is motion planning. OK, to go from point A to point B, I have to do this, I have to turn here, I have to do that. The other one is motion control, so I have to turn. I have to turn at this angle. I have to go at this speed. I have to put the brake on. Right? So there's the steering, the braking, the acceleration. So when you ask these questions, which of all of these functions are you asking questions about?

Stakeholder19_Judge 19:30
Yeah.

Interviewer 19:40
Because it's making decisions on all of these things. While it's doing this.
Stakeholder19_Judge 19:43
Yeah, no,... it... it probably, I mean in terms of identity or identifying to me whether it is a.. a ...a fire truck ahead or a child, or an older person, or some object that could be dangerous. That, I mean it, it concerns me, but not as much as the control itself. And I think the motion control is probably what I would focus on most. Or at the outset. So how do I know that this is going to put on the brake or reduce the ...reduce the speed or do whatever 'cause you might be in a line of cars I don't know on a highway on a superhighway and all these vehicles are monitoring themselves and that's the control motion as I understand it. That would concern the heck out of me for sure.
Interviewer 20:46
It's ...OK. So you're more concerned about the control functions than you are about the visual identification.

Stakeholder19_Judge 21:00
And because the visual identification, in my mind, so long as it identifies a something that will cause a danger that would result in a ...in a... in a disastrous or dangerous situation. To me, it doesn't matter whether it's a traffic light, a person, or otherwise. There there's got to be a sensor of sort. We know that. I accept that there are cameras and you know all sorts of things. My ... my worst fear, is once it identifies something, what about the control?

Interviewer 21:41
Mm hmm mm hmm.

Stakeholder19_Judge 21:42
And to me, you've got to err on. To me personally, I would rather err on the side of caution than not. Because if the vehicle identifies a traffic light and I have, maybe it's red, maybe it's not. I don't know. I would rather the decision be made to control it by putting on the brakes and, you know, controlling the motion.

Interviewer 22:08
So the question in these 16 cases, what do you have about the motion control questions?

Stakeholder19_Judge 22:15
Well, the motion control rather than … rather than putting on the brakes, I don't understand how it is that the whole autopilot disengaged? So is there, I'll call it a manufacturing defect, because of lack of better knowledge, but is there a defect in the motion control where a confused putting on the brake to disengaging the whole autopilot? Because maybe….

Interviewer 22:50
Oh. Oh, that … that's…. that's. Yeah, that's a distinction. You know, that's a distinction I hadn't thought of and no one else I've interviewed has thought of. That's amazing.
Stakeholder19_Judge 22:53
You know… sounds stupid. It sounds stupid and I apologise.

Interviewer 23:02
No, no, not at all. It's amazing. You're absolutely right. It could have done that. It could have been confused between slowing down or, you know, hitting the brakes rather than disengaging.

Stakeholder19_Judge 23:13
Or… or steering to the left or right, you know, depending on where…

Interviewer 23:18
Yeah.

Stakeholder19_Judge 23:18
He can control of directional motion or otherwise. Because I'm not going to say within a second it could do it, but it might have.
Interviewer 23:29
Yeah. Why did it do that instead of disengaging is your question. Or did it get confused between disengaging and doing one of those?
Stakeholder19_Judge 23:33
Yeah. Yeah, exactly. That would be my....
Interviewer 23:38
OK.
Stakeholder19_Judge 23:41
Kinda... my focus is say you know what the heck happened here? Why did it do that and not what it was programmed to do? And basically it's control.

Interviewer 23:53
Are you sure you didn't study engineering? Are you sure you didn't study engineering?
Stakeholder19_Judge 23:57
Ah sweetheart, I wish.
Interviewer 24:00
You sound... seriously, you sound like an engineer. You're tearing this down. This is amazing. This is great. OK, keep going. What other thoughts? Yeah, that's it. Oh, OK.

Stakeholder19_Judge 24:09
That's it. That's it. That's it. I quit.

Interviewer 24:13
OK. OK, you know what? I'm gonna give you a little bit of a break because the main question is this one for the questionnaire. And then I have a couple of secondary questions and we'll talk a little bit about those. Why I'm asking them and after you give me the answers and then I'll double back to this main question again. OK.

{Secondary Questions and General Discussion}

End Transcription for analysis …. General discussion continued until 56:20 when Interviewer stopped recording and transcription