Interviewer: 0:00

And we're recording here. Okay, so … I just want to make sure that you know. What I'm going to do is ask you two questions, okay, two, one main question and some secondary questions after that. The secondary questions are about the stakeholder group that you represent in society. So, driver, pedestrian, passenger, vehicle, police, etc. Okay. So the main question I'm going to ask you is about a real life case study, okay, involving automated vehicles. Okay, so automated vehicles are actually driven by software, artificial intelligence, software specifically, and they actually perform the human tasks of steering, braking and accelerating the vehicle itself, without the human being doing anything. Okay, so this is about the actions taken or not, decisions made or not, by this artificial intelligence software. Okay, so I'm going to describe a scenario involving a real-life case study, and what I'm looking for is your views, your perspectives, your thoughts, your questions, your opinions. It's not a test for you. Okay, there are no right or wrong answers, and at any point, if I'm using terminology that you want an explanation for, stop me and ask. If you don't ask, I'm assuming you know what I'm talking about. Okay,

Stakeholder10_Passenger:

okay

Interviewer: So remember, it's just about your views, okay, and the less you know about AI, the better. Okay?

Stakeholder10_Passenger:

Okay.

Interviewer:

Okay. So, the description of this case study, okay, now this case study scenario involves a real life case within the artificial or AI application of automated vehicles, which I'm going to call AVs, okay?

Stakeholder10_Passenger:

okay,

Interviewer: All right. So, it involves the occurrences of actual car crashes involving one particular AV brand, which is Tesla. You've probably seen them driving around, right?

Stakeholder10_Passenger:

I haven't,

Interviewer: Okay, so it has an advanced driver assistance system called Autopilot.

Stakeholder10_Passenger:

Okay,

Interviewer: so Tesla's autopilot system controls the steering, the braking, acceleration functions of the AV without any assistance from the human driver.

Stakeholder10_Passenger:

Correct.

Interviewer:

Okay. Also note that Autopilot could, at any time, disengage and hand over controls to the human driver, so it could just stop.

Stakeholder10_Passenger:

Oh,

Interviewer:

okay, it could. It could just like the vehicles moving, and it would just stop controlling acceleration, braking and steering like do nothing.

Stakeholder10_Passenger:

Could I just ask… these AVs that you're … that you're speaking about, is there actually a person at the wheel?

Interviewer:

Yes,

Stakeholder10_Passenger:

well, there is,

Interviewer:

there is a human being sitting behind the wheel to take over at any time.

Stakeholder10_Passenger: Okay,

Interviewer:

There has to be right now.

Stakeholder10_Passenger:

Oh, yes, okay.

Interviewer:

But they can press a button and it goes on, and they just go like, you know, so they don't control the steering, braking or acceleration, the machine does,

Stakeholder10_Passenger:

okay, okay.

Interviewer:

So according to USA National Highway Traffic Safety Administration, NHTSA, it's called, they have an office of defects investigation. And between January 2018 and January 2022, so like a four year period, Tesla AVs, with Autopilot engaged, were involved in 16 crashes, where they struck highly visible stationary in road or roadside first responder vehicles. So these are police, ambulance, fire trucks, road maintenance vehicles with lights flashing. These first responder vehicles were attending to preexisting collisions. So there was a crash up ahead. They were there … lights flashing, people with vests, you know, highly visible vests on and Tesla Autopilot,… Autopilot engaged, crashed into … And on average, in these 16 crashes, Autopilot aborted the vehicle controlled less than one second prior to first impact. So it just did this. And less than one second the human had to, you know,

Stakeholder10_Passenger: 4:46

it was too late,

Interviewer:

yeah, okay, any questions about this scenario?

Stakeholder10_Passenger:

Okay, is there a passenger in this car?

Interviewer:

Um, not known. There may have been passengers. They may not have been okay, okay, so for this particular PhD research project, okay, what I'm looking at is, and I described it in the consent form and background information my PhD research project is about the subject of explanation. Okay, so we ask why questions all the time, right? Why did our friends get divorced? Why is there a recession? Why did that airplane crash? Right? We're constantly walking around asking and giving explanations, okay, for all sorts of things in the world, and

depending upon the context of the situation and the person, the explanation is going to have to be different, right? And so these AI systems, most of them are opaque, which means they're not transparent. We don't know, including the designers, how they take the input and come up with the output. And these AI systems don't explain themselves adequately to the users or to their designers or whatever. Okay, so the technology that's out there, okay, is doing this. All right? And it's opaque, and it's performing human tasks. So it's not like the old fashioned software when I grew up, you grew up, etc, where you could go line by line, code by code in the software, okay?

Stakeholder10_Passenger:

Opaque means?

Interviewer: 6:27

Not transparent. You can't see into it. It's a black box, right? Okay, so what I'm doing is I'm coming up with a framework to say, Okay, there's this subsection that engineers have come up with of engineering… AI engineering, called Explainable AI, which is trying to break into these black boxes to understand more about how it came up with the outputs. Okay, so I'm helping that by saying, hey, engineers, you don't even know what an explanation is. How can you come up with an explanation? You're just coming up with an explanation for yourselves. But that doesn't necessarily mean the world agrees with what you're giving them that may not be an adequate explanation. So I'm doing research into what's an adequate explanation, and how do you come up with what an adequate explanation is, before you break into the black box and try and get the answer. You may be getting an answer that someone else may not want, right? So, so that's what I'm doing. I'm gathering inputs from all sorts of different people, different walks of life, different perspectives, different lived experiences. About one case study to show, hey,

different people want different explanations for this same particular case study. Okay, so policemen may want a different explanation to from you or me or xxx or xxx or you know Maria. A lawyer may want a different explanation. A judge may want a different explanation. Okay, so based on these 16 crashes, what I've just described, what kind of information you know when you ask the why did this crash happen question, what kind of information are you looking for about the steering, the braking and acceleration functions that this cars perform? So you could finish the sentence. What I would ask for is x y and z. You know, you could finish sentences like what I would want to know about, you know, X Y or Z, or I want information about X Y or Z.

Stakeholder10_Passenger: 8:40

Okay, I guess the first thing that comes to mind is, if the this may not have anything to do with what you're asking, if the law requires a guy behind the steering wheel, I would think that part of his job is just to kind of sort of stay alert through the whole thing. So he sees the responders, first responders, ahead of the vehicle. He knows the car is not slowing down or not getting in the proper lane or whatever. Why do they leave it to the last second to… to take over? So I guess

Interviewer: 9:20

the question is about the driver?

Stakeholder10_Passenger: 9:20

the driver,…

Interviewer: 9:21

the human guy,

Stakeholder10_Passenger: 9:22

correct. Is the system …the automated system is set up such that the human guy can over, like, disconnect the automated thing whenever they feel they need to, because something's not right, they feel something's not right.

Interviewer: 9:23

The answer to that question is yes, the human driver can take over at any time and do it instantly.

Stakeholder10_Passenger: 9:24

Oh, okay, okay, then so it's …it makes me wonder. Then. Are all those 16 drivers basically stupid or asleep or just simply not attentive to their job?

Interviewer: 9:24

The answer is yes. …we don't know the specifics in these 16 cases, but the answer to that question is yes,

Stakeholder10_Passenger: 9:26

OK, because I'm just seeing if, if they were attentive and stuff and these 16 things happened. Could it have been that the car and the driver didn't see the accident ahead of them because they had to turn around a corner, and all of a sudden, it was right in front of them. I'm not sure. Well, obviously nobody.

Interviewer: 10:26

Well, they could approach. They from the report, they were approaching it from a distance.

Stakeholder10_Passenger: 10:27

Okay, so that's off the list. Okay, then,

Interviewer: 11:12

But I don't know if they sensed it. We don't know. Okay, so you're asking these questions of the human being, what about a car itself?

Stakeholder10_Passenger: 11:20

No, I understand, but if the human couldn't foresee it, then how could the car have foreseen that? And therefore there was nothing wrong with the car's engineering or acceleration, deceleration, whatever.

Interviewer: 11:35

Well, okay, let's, let's put a plug in that for a second. These are great questions, and I want to narrow it down to … doesn't matter what the human did or didn't do,

Stakeholder10_Passenger: 11:45

yes,

Interviewer: doesn't matter what the human saw or didn't see.

Stakeholder10_Passenger: 12:00

Okay,

Interviewer: 12:05

for this scenario, let's talk about what the car did or didn't see, did or didn't take action. And when we say, see the It has sensors. Okay, let's assume also the sensors and all the hardware and all other systems are working perfectly. Okay, let's focus on the decisions of the car. Let's assume the car got all this information about the scene itself.

Stakeholder10_Passenger: 12:15

Okay,

Interviewer: 12:20

go from there and ask questions about the decisions it made and the actions it took. Can we narrow our thoughts on that?

Stakeholder10_Passenger: 12:43

I….I mean, when you're a driver, you with the machine, the person who's programmed the machine. When you're …when you're programming it, you know how long it takes, at whatever speed, to stop a car. How that not just the distance, but the time as well, how much it has to be brake, how fast it has to be braked in order to… to brake the thing. So if, … if the if the car is running into troubles like this. I … honestly Interviewer. I I'm not too familiar enough with the car to know too much. I'm gonna be one of your worst cases to help you.

Interviewer: 13:00

No, that's not true. Hang on, okay. Hang on. No, no, no, let me disabuse you of that notion.

Stakeholder10_Passenger: 13:30

Okay,

Interviewer: 13:33

the less you know about the design of the car, the better.

Stakeholder10_Passenger: 13:34

Okay,

Interviewer: 13:40

put yourself in the shoes of anyone in the world who's faced with a driver, a human driver, who does this 16 times, goes all over America. And does this 16 times. Do you not wonder why that happened? And if you're wondering why that happened, what kind of information are you looking for from the driver? Because this thing is a driver, okay, whether it's

human or whether it's a machine, it's a driver. It knows how to drive, okay, it knows the rules of the road. But still, despite that, there's this pattern of 16 crashes. What do you want to know?

Stakeholder10_Passenger: 14:36

Well, I.. I guess, who who did the who did the programming of the software? What were the measurements of.. of it? Maybe somebody programmed it incorrectly, so that the deceleration, or whatever, of the of the car wasn't input correctly, maybe..

Interviewer: 15:05

Assume that worked in other cases. Assume it's been driving around for a couple of years and it knows how to stop, decelerate, accelerate. Knows the distance between events and objects.

Stakeholder10_Passenger 15:22

Well, see, I have a computer at home, and I have a printer, and sometimes they talk together, and sometimes they don't, and you sit there and you just struggle with technology. And so I'm thinking, yeah, sure, the car was programmed to drive perfectly. But hiccups happen with technology, and I don't know that anything can be 100% accurate all the time. As you said, the car is a driver, just like a human is a driver, and we all make mistakes. So would I get …would I get overly excited about this? I kind of say … each instance, of course, is very important. You've probably killed somebody at the side of the road, on and on. This is all very, very important. But if you looked at the 16 versus the hundreds of … of problem of cases that where automated cars have been driving around. I mean, maybe there are only 17 cases of them driving around in those four years, so maybe 16 out of 17 is holy crap. What is happening here? The obviously, the

technology wasn't put into the car properly. But if it was 16 out of 16 million, well, that's really pretty good, being that things do always go wrong, could go wrong,

Interviewer: 17:07

so you don't need an explanation.

Stakeholder10_Passenger: 17:08

I'm saying maybe I don't need an explanation.

Interviewer: 17:10

Okay, okay, very valid point.

Stakeholder10_Passenger: 17:15

I mean, Tesla's, I'm going off rails here, but Teslas blow up. They get on fire and on and on. But I don't see the production of Tesla stopping and … and Teslas. I never hear Mr.

Interviewer: 17:20

Musk

Stakeholder10_Passenger: 17:34

Musk. Thank you, Elon … I kept thinking …. ever apologizing for anything or or why things are going wrong or anything it doesn't …I'm maybe not tuned into the proper news channels, but I'm thinking that I that … I'm thinking that maybe the story isn't as bad as it could be, as it could be interpreted.

Interviewer: 18:03

Let me, let me disabuse you. I'm not trying to alarm you. I'm not trying to make a judgment. I want these machines to work better, okay, being a driver in the GTA. I'm sick and tired of driving. I want automated vehicles, but I want to be able to trust them, and right now, I don't believe they're ready for me to just sit and ignore it driving, you know what I'm saying?

And there have been cases, there have been more than this, like crashes have occurred with Autopilot engaged, okay, in some instances, people have died, etc, no one died from these 16 crashes. The reason I chose these 16 is number one, it's documented, and it's official by a government agency, and it's evidence of a pattern of hitting very highly visible objects and large you know? And so it allowed me to probe into what's going on in this black box, right? How can we make it better? How can we make sure it doesn't crash here? Or even the statistics that you posed … to me, one crash is too many. I'm an engineer with really, you know, really want no crashes. Okay. I really would ideally love it okay, if the safety records of car crashes just went down completely, you know. So this is all about what kind of an explanation. So I understand you don't necessarily want an explanation, which is fine, because you're looking at the numbers of these crashes, etc.

Stakeholder10_Passenger 19:50

I mean, I guess the other thing I could say is, … like somebody out there made … made the programming software so ..

Interviewer: 19:58

but the people who made the software don't know what it does. It's not like what you're the way you're describing it. This technology is very different from previous technologies. All the technologies I've worked on, xxx has worked on, other engineers have worked on, we could audit it. We could look into all the nitty gritty parts in that this software, you can't do that. It literally is a black box. It's a layer upon layer upon layer of processing the data that gets inputted. And the data that gets inputted is cut up into little pixels, and you know what I mean, little bits and bytes of zeros and ones, and it's processed, and then out comes a result, and don't know what

the inside is doing. That's … That's why this AI system is called opaque and black box. So you know, how can you improve it if you don't know what it's doing right?

Stakeholder10_Passenger: 20:10

Exactly.

Interviewer: 20:10

And part of knowing what it's doing is what an explanation is all about it's like, okay, so what kind of information do you have… Do I want, so that I can improve the performance of this thing? That's what this project is about. And… and it's perfectly valid to say, statistically, Tesla is doing really well, millions of miles driven without accidents and crashes and that and human beings make mistakes. These things makes mistake perfectly valid, perfectly valid answer. Okay, I'm gonna stop a little bit. Yeah. I'm like, …. go ahead. If no, no, go on. You had a train of thought.

Stakeholder10_Passenger: 20:15

Well, you were saying I'd like to talk to lawyers. I like to talk to policemen and stuff, yeah, but I'm sorry, but what more would they wish to know?

Interviewer: 20:20

I don't know.

Stakeholder10_Passenger: 20:21

Then, if it is a black box, how can anyone?

Interviewer: Well, we humans, we well, because the inputs and outputs, they do control and in that they change the inputs and look what the outputs are, and they go back and say, This is a correct output or not. That's how they improve these black boxes. They're now able to go in and probe and go into some of the layers and say, okay, these particular factors played a higher,

more important role than those factors in determining the output. So they're probing the black box right now, like, that's what this Explainable AI initiative is, where they're literally probing the insides, different parts of the insides, and trying to get a sense of what played a higher role in the output. You know which aspect of the input? You know which factor, which parameter or characteristic played a higher role than this one? Sometimes it comes back to, Oh, we didn't give it enough inputs, or we didn't give it enough of the right type of inputs when we were training this thing. So this is the initiative for Explainable AI. But in order to figure out whether the output is we need to also understand why did it do this? You know, if it's a human being, we're also black boxes, yet we can explain ourselves. Do you know what I'm saying? Like, if it's a human that went around and crashed into 16 collision scenes, would you want …not sant to ask the human … What the hell?

Stakeholder10_Passenger: 23:48

yes.

Interviewer: 23:48

Okay, what would you ask and what kind of an explanation would satisfy you from that human being?

Stakeholder10_Passenger 23:50

I look at it differently, though a person versus a box.

Interviewer 23:56

Tell me what you would do. Okay, look at it as if it's a person doing this,

Stakeholder10_Passenger 24:02

Well, just why? What were you possibly thinking and why weren't you alert? And what were you doing when this happened? Were you on your phone, and were you not aware that flashing lights means watch out and get out of the way. And were you drunk, and were you alert, and were you distracted? And what was the conditions of the road, the weather? And

Interviewer 24:40

when you said, What were you thinking? What kind of thoughts did you have in mind? Stakeholder10_Passenger: Well, were you concentrating on… on the job at hand, the driving, or were you thinking about what your day brought, or what you did last night, or what you're worried about and … we think of so many things. I think as women, even our minds are constantly buzzing everywhere so … but see to me, a car isn't like that.

Interviewer: 25:22

Why not?

Stakeholder10_Passenger 25:22

Well again, if we don't even know what's going on with these pixels and stuff, you said, maybe then …

Interviewer: 25:24

But we don't know what we're doing with our pixels look. We could look at the scene around us, and we could use the analogy, everything that's hitting us. It's not one thing. It's individual pixels hitting our eyes, going into our brain processing, and we're making a determination, oh, that's a TV, that's a chair, that's, you know, that's Stakeholder10_Passenger and her glasses. And we don't know how we process it, how we put the chunks of information that comes out of you know? Information comes out of us in chunks as well. And every question you

asked is a valid question for the machine, also, because the machine is acting like a human in the world. Remember that it may not have had a fight with its spouse the night before. It may not be worried about its daughter or the job, but it's processing information just like we process information, right? It's taking all these inputs, and it's doing things with that input. Are your computers doing it? Your printer is doing it, right?

Stakeholder10_Passenger 26:52

I guess something just occurred to me also, is maybe the age of the car, the age of the software,

Interviewer 27:00

okay? Valid point, yeah. Is it out of date? Yeah? Is it current? Is it up to date?

Stakeholder10_Passenger: But even whatever you buy now, it wears and tears and repair and stuff. So should the software or whatever? I mean…

Interviewer 27:20

it's constantly being upgraded, just so, you know,

Stakeholder10_Passenger: 27:21

okay, yeah, but like, could possibly in those 16 cars, the black box said, press hit the brake, and maybe the brake wasn't actually,

Interviewer: 27:22

no, assume, all of the hardware was working.

Stakeholder10_Passenger: 27:23

Yeah,

{secondary questions discussion}