Transcript

3 March 2025,

Interviewer 0:11

OK, so just a bit of background before we get to the questionnaire. The purpose of this interview is basically to gather in depth information on the subject of explanation from the perspective of humans and use that information to develop explainable AI models. So remember, this isn't a test for you, … It's only. It's only about gathering information about your views, perspectives, thoughts, opinions, viewpoints, questions, … about a real life scenario that I'm going to describe to you and there are no right or wrong answers. So, everything you say is very, very important and valuable. OK, so let's talk about the real-life case study. So… it involves a real case within the AI application of automated vehicles, or AVs. … It involves the occurrences of actual car crashes involving one particular AV brand, … Tesla and its advanced driver assistance system called Autopilot. So, Tesla's Autopilot system controls the steering, the braking, acceleration functions of the AV without any assistance from the human driver. Furthermore, Autopilot could at any time disengage and hand over the controls to the human driver. So, according to us as NHTSA, which is the National Highway Traffic Safety Administration, their Office of Defects Investigation opened an engineering analysis between …. about crashes… between January 2018 and January 2022. So 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. And they were attending to pre-existing collision scenes and these Teslas crashed into them, and on average, in these crashes, Autopilot aborted vehicle control less than a second prior to first impact, so it stopped steering, braking, accelerating. It just went at the speed…. It wasn't just…. you know, let go. So, do you have any questions about this scenario?

Stakeholder7_Engineer 3:24

I don't think so. No, not right now.

Interviewer 3:33

OK, OK. If you want a description I could put it in… the whole thing… in chat or if you're OK, we can proceed.

Stakeholder7_Engineer 3:41
In the chat would be good as well, yeah.

Interviewer 3:43
OK, let me just copy and paste it. OK, copy. There is also a website of pictures of the crashes. If you want to see it, do you want that link or you're OK?

Stakeholder7_Engineer 4:08
Would that help and then?
Interviewer 4:10
Not really. It's just some people have …. have, you know, have. It's neutral. Some people find it helps, some people don't, so.
Stakeholder7_Engineer 4:22
That's fine then.

Interviewer 4:24
OK, the… the description does not change. The scenario does not change, so… it just shows you crashes from all over America. OK, so based on this scenario that I've described, you're gonna' be asked one main question and a few secondary questions. The secondary questions have more to do with your ….

Stakeholder7_Engineer 4:29
Right.

Interviewer 4:44
Stakeholder designation because I'm interviewing a cross section of society. So, the main question is based on this scenario that we've just described. Um…
You are seeking explanatory information about these car crashes from Autopilot, the AI system that was doing the steering, braking and acceleration function so the motion control. So, these are, you know,… we ask why questions all the time in our …. you know why? You know, why did our friends get a divorce? Why did an aeroplane crash? Why did the economy go into recession? So, this is along the lines of explanation of about this. Why did this car crash so…What information specifically are you looking for about these 16 car crashes?

Stakeholder7_Engineer 5:43
I guess this thing that's initially sticks out to me is that Autopilot was disengaged only about one second before the impact, which is not much time at all, so I would want to know: Why the Autopilot didn't notice this potential risk earlier?

Interviewer 6:12
OK. Can you elaborate a bit more about like what... what are you thinking more specifically about? About the risk or recognition of the risk?

Stakeholder7_Engineer 6:26
I mean, I understand that when it's in self driving mode, you're still supposed to be on the wheel and paying attention as if you're driving, but ...I think in reality, often drivers maybe aren't fully paying attention. And so I... I want to know.
Were ... were these drivers fully paying attention? Did they notice the emergency vehicle up ahead and.

Interviewer 7:00
Yeah, the ... the data so far basically...um...does not go into details. It's assumed that they were distracted, so. for whatever reason, we don't. If they were sleeping or watching a movie or you know, so the.

Stakeholder7_Engineer 7:18
Yeah.

Interviewer 7:20
Obviously, if they had been paying attention, they would have seen it and taken control, but I'd like to go back to the AI itself. The thing, the entity if you will, that was controlling the car and the questions about the driving functions. Do you have any questions for it?

Stakeholder7_Engineer 7:44
I guess I would want to know. When did the AI realised that there was a situation, quote unquote with abnormal vehicle behaviour up ahead and ...and when?
Why did it decide to simply disengage rather than engaging in emergency braking, for instance?

Interviewer 8:18
OK, OK.

Stakeholder7_Engineer 8:18
Because it doesn't seem like the self driving feature gave the human driver, much of a chance to counteract the … the scenario.

Interviewer 8:35
Right, right. Any questions about other driving functions other than the braking?

Stakeholder7_Engineer 8:47
I guess emergency steering would also be an option.

Interviewer 8:52
Mm hmm.

Stakeholder7_Engineer 8:56
And why that wasn't considered as an option?

Interviewer 9:01
Mm hmm mm hmm. OK. If your objective were to ensure that crashes like this don't occur in the future, what specific things like features? Functions, methodologies would you be looking for in the algorithm itself that was designed to do this?

Stakeholder7_Engineer 9:27
What features would I be looking for in the algorithm?

Interviewer 9:31
Mm hmm.

Stakeholder7_Engineer 9:33
I mean, I'm sure self driving car algorithms are incredibly complex. And they take into account a variety of sensors and inputs. So, I'm not… I'm ….I'm not quite sure off the top of my head exactly what I would want to be considered.

Interviewer 10:05
Mm hmm. Mm hmm. Well, you are aware that a lot of these AI systems that are functioning are opaque, so they aren't the traditional software programmes. You know where they are deterministic. These are stochastic. So even the designers themselves….

Stakeholder7_Engineer 10:21
Yep.
Yeah.

Interviewer 10:28
Do not have insight into how the outputs and the decisions are made.
So. Therefore, you know, I don't know if even in these particular scenarios or these all these AVs that are out on the street….The designers, I don't believe that at this point in time they have insight into the decisions, how they're coming up with the decisions, how they're taking the, … you know, deciding which actions to take or not take, which decisions to make or not make and…It's very opaque and not transparent at all so um…. Do you have any other type of information you want? Like if this was a human being that went about all over America and crashed 16 times?

Stakeholder7_Engineer 11:23
Right.

Interviewer 11:26
Into emergency vehicles would… what kind of questions would you ask of, say, a human agent as opposed to this AI agent, right. They're doing human tasks?

Stakeholder7_Engineer 11:35
Alright, alright. I guess my question would be why didn't you learn after the first couple of this type of scenario? Like once, perhaps I can understand, but 16 times.

Interviewer 11:48
OK.

Stakeholder7_Engineer 11:55
Clearly there needs to be some learning going on that isn't happening.

Interviewer 12:00
OK.

Stakeholder7_Engineer 12:02
As you're, I mean, once or twice you can a few times. Maybe you can blame on the driver not being sufficiently attentive, but given that it happened in a number of times and probably it happened other times that NHTSA is not aware of. … It seems like there's a flaw specific flaw in the algorithm regarding this type of situation, where there's an emergency vehicle stopped in the regular road surface.

Interviewer 12:32
Mm hmm. Well, it wasn't just in one emergency vehicle. There were lots of, you know, and people milling around with high visibility vests and all that. So it was pretty,… it was a pretty obvious scene of, you know, things up ahead. OK.
So would you ask a different question to the AI than you would the human?

Stakeholder7_Engineer 13:01
Good question. Well, an AI is interesting because you know, presumably it's…I… I actually,… I… I don't know. I don't know what information is saved in its memory state and about a decision that it made a few seconds ago or a few hours ago. But hopefully it does have that sort of memory versus a fallible human memory. You're not. You're saying perhaps no.

Interviewer 13:40
No. um… You're thinking in terms of a deterministic software you're not thinking in terms of stochastic software. Stochastic software do not keep logs like deterministic software do right? It's….It's opaque. It's a black box. People can't look inside of it. It's. You know they have video of whatever the camera is recorded, but they don't have a way to say, OK, I made decision X, at this point in time. I made decision Y, at this point in time that is not available to the designers. Hence why these systems are black boxes. They're labelled that way and hence the need for explainable AI which is.

Stakeholder7_Engineer 14:28
Right.

Interviewer 14:33
An initiative created by DARPA back ten years ago where they recognise, OK, all of these AI agents are doing things, and even the designers don't know how they're coming up with the outputs. They know what the inputs are, they know what the outputs are, they know how to tell it. Oh, this is good enough. This is bad. This is accurate. This is correct and … but beyond that, inside the box itself, so explainable AI is trying to break into that black box to get some of these answers.

Stakeholder7_Engineer 15:05
Right.

Interviewer 15:05
So again, no, that log does not exist in this form that you're talking about. OK so. Therefore, it's an agent acting in the world doing human tasks, and it did this. So, what do you want…to

Stakeholder7_Engineer 15:22
Yeah.
Interviewer 15:24
Know from it because other agents, human agents in the world. You ask explanations from all the time. What is… What is your main question and where… where your head is saying: OK, this would be an adequate explanation for this pattern of 20 crashes or 16 crashes.

Stakeholder7_Engineer 15:44
Yeah. I mean, I guess I want to know what were the different options that the AI considered before disengaging immediately before the …the impact? What other options were considered? When were they considered? And why? Why did it choose one option over another?

Interviewer 16:12
Mm hmm mm hmm.

Stakeholder7_Engineer 16:13
I mean I … I think AIs were kind of probabilistically, in determining that, you know, maybe 75%, this is the best option, 20….20% option B is preferable, 5% option C and so then it usually goes with option A.

Interviewer 16:35
OK, cool. Yeah, thank you. That's those are excellent, excellent inputs and questions. Anything else?

Stakeholder7_Engineer 16:54
I mean, I'd want to know if the AI has any capacity to learn from a particular scenario, perhaps even just within the same vehicle.

Interviewer 17:06
The answer is yes, they're always updating the algorithm always. Yes. Yeah. The more data it gets, the more it continues updating.

Stakeholder7_Engineer 17:11
Right, yeah.
Yeah.

Interviewer 17:18
Not necessarily in a good way, but.
Stakeholder7_Engineer 17:22
Yeah.

Interviewer 17:22
But it's updating.

Stakeholder7_Engineer 17:25
Yeah.
Interviewer 17:25
Yeah.

Stakeholder7_Engineer 17:27

So then I mean, if it was a human driver that was making these mistakes, I'd want to know what have you changed about the way you drive to make sure this doesn't happen again?

End Transcription for analysis .... secondary questions and general discussion continued until 36:13 when Interviewer stopped recording and transcription