Transcript

13 March 2025

Interviewer 0:23
OK, OK. So, the background information just kind of gives you an update and context about this study. So, as you and I have talked in the past, I'm working on explainable AI and I'm saying, hey guys, an explanation is not one uniform thing, right? We're always asking why questions, why did our friends get divorced? Why did an aeroplane crash? Why did the economy tank?

Stakeholder16_AI Developer 0:31
Yeah.
Mm hmm.

Interviewer 0:50
And depending on the situation and the person, the answer to these questions will be different. In fact, even if you ask the same event, different people will look at it from different perspectives like you will. I will, you know, a policeman will, you know, engineer looks at things differently. So basically, for explainable, currently, most engineers are building it for themselves rather than for a wide variety of people. So I put together a framework to help them identify exactly what model, what kind of model to build, what kind of questions to answer. And so this is more in the requirements management field, OK.
So basically the purpose of this interview is to gather in depth information on the subject of explanation from the perspective of humans. So I'm going to give you a case study and you'll give me, you know, details on what kind of explanatory information you want from it, and that'll help me verify and validate my framework so I can help engineers, you know, do a similar thing in the future. So remember, this is not a test for you. So you I need you to relax, OK? So yeah, everybody's going OK. I have a questionnaire. I'm like. nope, this is not a test. It's all only about gathering information about your views, your perspectives, your thoughts, your questions, your opinions about a real life scenario. OK, so there are no right or wrong answers, so I need you to take a deep breath and relax and just give me your opinions, OK?

Stakeholder16_AI Developer 2:38
Sounds good.

Interviewer 2:39
OK. You ready? I'm going to describe the scenario to you.
Stakeholder16_AI Developer 2:43
OK.

Interviewer 2:44
OK, so this case study scenario involves a real-life case within the AI application of automated vehicles, or AVs. It involves the occurrences of actual car crashes involving one particular AV brand which is Tesla and its advanced driver assistance system, ADAS, called Autopilot.
Stakeholder16_AI Developer 3:09
OK.

Interviewer 3:09
So Tesla's Autopilot system 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 controls to the human driver.
Stakeholder16_AI Developer 3:26
Yeah. OK.

Interviewer 3:29
So according to USA NHTSA, National Highway Traffic Safety Administration's Office of Defects Investigation, between January 2018 and January 2022, Tesla AVs, with Autopilot engaged, were involved in 16 as in 1-6 crashes where they struck highly visible stationary in road or roadside first responder vehicles that were attending to preexisting collision scenes. So, police, ambulance, fire, trucks, road maintenance vehicles, lights flashing, there is a collision up ahead. They're all there. OK, people with vests on highly visible right, OK. Furthermore, on average in these crashes, Autopilot aborted vehicle control less than one second prior to the first impact.

Stakeholder16_AI Developer 4:26
OK. So what I just want to confirm, I understand that the …the situation, So what … what those 16 crashes you were saying? It crashed into highly visible things?

Interviewer 4:41
Yes.
Stakeholder16_AI Developer 4:42
Highly visible humans or cars?

Interviewer 4:44
Cars. No, um… Yeah, actually, humans also, yes, police were. Yeah, in some instances the police were injured.
Stakeholder16_AI Developer 4:52
OK.
Interviewer 4:59
In some instances, OK, so I have. Yeah, I have … I have a website I created with news, reports and photographs. If you want to see that.
Stakeholder16_AI Developer 4:59
Sure. Yeah.
Interviewer 5:11
Of the crash scene. So various. .. Here we go. OK. There you go. So I combined, … there's even a video of one of them and you can see the crash scenes.

Stakeholder16_AI Developer 5:24
OK. And … and when you were saying. Uh… When you were talking about police cars, so you were saying it crashed into the police cars?

Interviewer 5:32
Yes, and fire trucks and or….
Stakeholder16_AI Developer 5:33
OK.
Ah, OK.

Interviewer 5:36
Do you see? Yeah.

Stakeholder16_AI Developer 5:38
Wow, I didn't know that.

Interviewer 5:39
Yeah, yeah. So these guys were attending to another accident, right? Fire, … fire trucks come. Ambulance comes, police comes, you know, tow trucks come, lights flashing. See all the vests, they're wearing. Highly visible. And Tesla just boom drives into them.

Stakeholder16_AI Developer 5:49
Wow. Yeah.

Interviewer 5:59
OK, real life 16 cases all over America at different times over this four year period.
Stakeholder16_AI Developer 6:07
Mm hmm.
Interviewer 6:07
So I chose this because a) it's not an anecdote, it is an actual investigation conducted by a government agency. It's a pattern of car crashes like Tesla has other car crashes with Autopilot involved, but this one in particular is about highly visible stuff. It's a pattern, it's all…. It's all over America.

Stakeholder16_AI Developer 6:26
Yeah.

Interviewer 6:32
… Lasted for four years, so it's a nice little case study scenario.

Stakeholder16_AI Developer 6:37
Oh wow, I didn't. I never heard about this in the news. I heard about Tesla running over like a woman that night or something. I never heard about this. This is crazy.

Interviewer 6:45
Yeah, well.
Yeah. So there are different Tesla accidents and you know I won't get into those other cases or studies ‘cause for this. I'm just kind of put a box around it for… for today anyway. OK. So do you have any other questions about what I've just described to you?

Stakeholder16_AI Developer 6:53
Mm hmm. No. I think that that makes sense, yeah.

Interviewer 7:08
OK, so it was driving in Autopilot. There was like a human. There was a human driver behind the wheel, but distracted didn't intervene. OK, through this whole scenario, the car just kept going, same speed and crashed into it and released control less than a second before the crash in, you know, on average.

Stakeholder16_AI Developer 7:11
Yeah.
Yeah.
Ah, OK, So what you're saying is one second before the crash, the car put the responsibility on the driver basically to say hey, Autopilot is done. You have to do something. Yeah. That's actually one thing I was wondering. So these car crash. So the way Autopilot works is it's driving itself. But then at some point, it has to give.

Interviewer 7:36
Yes, yes.
Yeah.

Stakeholder16_AI Developer 7:52
… The driving back to the driver.

Interviewer 7:55
Mm hmm.

Stakeholder16_AI Developer 7:56
And in this case, he said, it was about, like around one second.

Interviewer 7:59
Less than a second on average.

Stakeholder16_AI Developer 8:01
Yeah, OK. Less than a second. I'm gonna. … OK. OK.

Interviewer 8:04
For these 16 crashes, yeah.

Stakeholder16_AI Developer 8:07
OK.
Interviewer 8:07
OK, alright. So ready now for the actual questionnaire?

Stakeholder16_AI Developer 8:13
Yes, yes, I'm so excited haha….

Interviewer 8:15
OK, OK. So it's one big main question and then a few secondary questions about the stakeholder group you belong to. OK. So let's just talk about the first big question. So based on this scenario that we've just talked about, described to you, et cetera, you are seeking explanatory information about these car crashes from Autopilot. OK, so, you are asking yourselves, why did this car crash happen? But in your mind, you have some things that you really more specifically wanna' know about. OK. So from the AI system Autopilot that controls the steering, braking acceleration, i.e. the motor control functions of this AV, what questions do you have for it? Like, put everything else aside, put the driver aside, put the, you know, the.. the …the weather aside, assume everything the….

Stakeholder16_AI Developer 9:00
Mm hmm.

Interviewer 9:11
Cameras, all of the sensing equipment, all the hardware is working just fine.
What information do you have for Autopilot about the decisions it made or didn't make the actions it took or didn't take?

Stakeholder16_AI Developer 9:31
And in this case I'm... I'm ... I'm myself.

Interviewer 9:39
Mm hmm.

Stakeholder16_AI Developer 9:40
I'm not like a lawyer or something.

Interviewer 9:42
No, no. You are the ... you are an AI engineer, right?

Stakeholder16_AI Developer 9:46
OK, OK.

Interviewer 9:48
Yeah. Or ...or actually, what stakeholder do you belong to? Like you are a driver, you're a pedestrian. You're an AI engineer.

Stakeholder16_AI Developer 9:56
Yeah.

Interviewer 9:58
Let's call. Let's look at it from the perspective of an AI engineer, OK?
Stakeholder16_AI Developer 10:01
Yeah, yeah, yeah.
Yeah, like my first reaction to this is, I'm like, oh, how are we gonna' understand what happened? You know, like... I know that these systems are like really black box.
So from the beginning I'm like, well, think we'll get an explanation for this for a

specific incidents. However, my question is more how did they train this model? Because I know that it's deep learning and it's very black box, so I'm more like who the people who designed it. My questions would be How… Why was this? Like how different was this scenario in comparison to the training? So 'cause, I'm really just this this … 'cause in the case of for example, the… the woman. I remember that there was this news about the woman who got hit by a car at night. I watched the video of it and it was very obvious that most humans would have hit her also because it was from what I saw in the video, it was dark. She was just walking in the middle of the road. It was kind of like one of those scenarios where, OK, it's kind of understandable given the information I'm shown but in this case it's like this is really obvious to see. And the AI should have known. Should have seen this kind of situation in its training data. So what… How does this diff… How do these situations? I would probably talk to the engineers who made this and I would say how do you think this situation differentiates itself from the situations it saw during its training? So when you were kind of driving the car and getting sensory footage or sensory data to train on, how…. Why is this different from what it was trained on? That.. that's …that would be my first question. I'm not…. I'm now talking to the programmers instead of the system.

Interviewer 12:15
No, this is fine because you could … you could give me feedback on either or. In fact, I asked people. I asked people OK, what do you want to talk to the car about? Because laypeople don't think of the… from the engineer's perspective. Some people want to talk about the designers, I want to talk to the designers. Right. So if you, which is fine. So you want to talk to them now, which function are you talking about? Because….
Stakeholder16_AI Developer 12:20
Yeah.

Interviewer 12:41
… it performs motion control as well as perception, right? So are you focusing … which function are you focusing on? Which… which AI algorithm? 'cause you know, multiple algorithms….

Stakeholder16_AI Developer 12:51
Yeah, exactly.

Interviewer 12:52
Enable this car to drive. So can you be? Yeah. So can you walk me through that?

Stakeholder16_AI Developer 12:58
Yeah, yeah. OK. So I would, I'm looking probably from the perspective of object detection. So for example, if the car goes into a school zone, it should detect, the school zone sign. I assume that's they're probably using an algorithm like that. So, yeah, I... I think one of the first ways I would investigate this situation is I would probably question their object detection algorithm. And to see how they use it exactly 'cause, I assume they have a way to detect cars, signs, those kinds of things. So my ... so in this case we have this truck that's very visible, and so my first question was....

Interviewer 14:01
Multiple trucks, multiple trucks.

Stakeholder16_AI Developer 14:04
Multiple. Yeah. Multiple trucks with freaking lights going on... on top of them. So my first question is, did it detect the trucks? I assume it did detect the trucks. If it didn't detect the trucks, this would be really weird. Maybe it detected the trucks, but I thought it was like maybe it saw a row of trucks and thought it was a bridge or something. You know, I don't know, but I'm sure it saw something, unlike the woman walking at night, which it probably didn't see in the case of the woman walking at night. I assume it didn't see the woman and just hit her. But in this case, I guess it saw the fire trucks.

Interviewer 14:35
Mm hmm.

Stakeholder16_AI Developer 14:43
So my first question to the programmers would, can you go back into the programme or into the ... the ..the software and see ... did it.... what did it detect

when it saw the trucks? 'cause it probably saw them. That would be my first....where I would start. So for that... I that's how I would look at it from an object detection now, for me, I guess more, general point of view. Um... I assume it has some kind of algorithm also to react to things because they can't hard code the way it's going to react to every single possible things, that's something it needs to learn in a black box scenario like, I don't think they have a... so if you're in that school zone, go to 30 like, I'm sure they didn't write like code like that. I'm sure though, when it sees an object with the object detection algorithm it ...it's going to react to that object. But the way it reacts is something it learns. It's not something that's hard coded. And this is where I'm kinda' coming back to my initial thing where I'm like, what ... when it was trained to react to fire trucks, what situations did it encounter? Because I remember, II was listening to this podcast and it was talking about the situation of seeing ...you know a girl crossing the street. If you see a little girl ,11-year-old, crossing the street obviously needs to stop. But then there is the case of you know, if a....so the ... obviously if it's seeing an 11-year-old go crossing the street that has to stop. But let's say it sees a ball kind of bouncing on the street. A human sees this and they... they immediately think, oh, someone is gonna' go maybe catch that ball. And so this is one of like those scenarios where you know, humans will think of this, but a car might not think of it. It might say, oh, I see a ball where, you know, maybe I should stop, but maybe I shouldn't stop in every situation. I see a ball because it might injure the drivers also. But of course knowing that a kid could come pick it up, this is where you're like, no, you have to stop the car.

So you know this is ... these are the kind of situations where I think it's really tricky to kinda' train the car on how to react to them. So just coming back to the fire truck situation, I again, I would say like what did you train? What scenarios did you train the car to react to in these situations? Because I guess it's ... it uses some kind of reinforcement learning where if it reacts correctly, you're rewarded it and it doesn't react, well, you... you don't reward it or give it a punishment. So I was like, what kind of scenarios did it see during its training involving fire trucks? So, yeah, just to summarise, everything I've said, I would look at from 2 perspectives, the perspective of object detection. Did it detect the fire trucks and what did it see when it detected them? And then the more reaction perspective where it's like, what did it see during its training and how did you teach it to react to during those situations?

Because for those, that's how I would probably see it.

Interviewer 18:39
OK.
OK. Anything else you want to know from this algorithm or the engineers?
And I ...I should say multiple algorithms 'cause there's neural nets, RLs, right? CNNs, RNNs... I'm like, I have no idea. Right? So I from what I've read, it's... it's multiple ones doing different things, right, so.

Stakeholder16_AI Developer 18:59
Yeah. Oh, yeah, exactly. Yeah.
Yeah, yeah.
It's.
Yeah, but I mean, at the end of the day, it's all black box. You know it's ...it's, it's all just neural nets that we can't really know. Understand.

Interviewer 19:15
Yes, yes.
Well, even RLs ..RLs you can't probe either, right? Yeah, yeah.
Stakeholder16_AI Developer 19:22
Yeah, exactly so.
It it's always like, yeah, it so it it's, I'm sure they're just using some kind of either transformers or CNN's, but it's all just black box and all that stuff so.
Yeah, those are the two questions I have for now. I may if I if I think of other ones, I'll tell you, but this is immediately the ...the two things I would look at for explanations.

Interviewer 19:39
OK.
OK, so let's... let's take a break for a bit and talk about the secondary questions so.

{Secondary Questions and General Discussion}

Interviewer 24:04
OK, OK. Now why would you want this explanatory information? I was asking at the beginning of the question. If you want an explanation. Is it because of trust? Is it transparency? Other reasons?

Stakeholder16_AI Developer 24:20
Why would I want an explanation for the accidents?

Interviewer 24:23
Mm hmm.

Stakeholder16_AI Developer 24:25
Well, I definitely want an explanation. I just think… Well, there's a couple answers for. The first one is I guess from … I'm AI engineer point of view. This is like…you know, that's kind of what we do. You know, like we deal with issues like that. So it's like just from an AI engineer, you want to know what happens, right, what's the problem? What's the solution?

Interviewer 24:50
Yeah, yeah.

Stakeholder16_AI Developer 25:03
But from a more general point of view, I… I think like automated driving is one of the toughest applications of AI because humans…you know, because like the … the thing about deep learning is like because it's black box, there's always like those little edge cases that are gonna pop up later that you never really thought about. You know, it happens all the time for me or for my work when I'm dealing with AI stuff. It's like there's always, always those edge cases where something weird happens and you didn't see it coming because you know you train your AI, you don't hard code all the rules. So there's always little things that are gonna pop out. And when you're dealing with, like, language, that's kind of fine because you … you just tell the user is, yeah, it's the weird stuff will pop out at some points. Hallucinations, as… as …. as they call it like. It's gonna make up stuff sometimes, watch out for it. But with automated driving, those little edge cases, humans are at risk and this is like the big difference. So I just think it's important to really understand. Even when humans aren't injured or … or when there aren't big repercussions, it's really important to keep track of, you know, why these incidents happen, because that's what happens when you deal with black box models. It's like there's always going to be weird stuff that pops up at sometimes, and it's OK when it's like an LLM writing an e-mail or something, but when it's a car and humans are at risk, it's like even if….you know, no one was injured.

It's still important to like, keep track of these little weird issues. 'cause they're gonna' keep popping up with these black box models.

Interviewer 27:00
Yeah. So OK, so safety and also transparency, I guess improving the design is why you want to know, OK.

Stakeholder16_AI Developer 27:06
Yeah.
Yeah.

End Transcription for analysis .... general discussion continued until 00:46:24 when Interviewer stopped recording and transcription