# Transcript

24 March 2025

**{Salutations and general discussion re. opaque AI systems}**

**Interviewer** 0:27
And, so the field of explainable AI looks into trying to basically get insights into how the outputs that come about. So, the decisions made, actions taken by these AI in some cases agents that are acting in our world without human intervention. And so, my research falls under explainable AI. And what I'm doing is … I recognise that AI engineers are building explainable AI for themselves, mostly us being, you know, engineers. Being engineers, we build stuff for ourselves without looking at overall society, all of the stakeholders and their needs. And so, I took a step back and I realised hey, wait a minute, we don't even know what an explanation is because, you know, and I took a deep dive into the subject matter of explanation from thousands of years to now and basically…. although explanations are ubiquitous, they are not uniform, they depend on context --- depending on the situation and the person. And we're asking why questions all the time to each other, we're asking and giving answers. You know, why did our friends get divorced? Why did that car crash? Why did that aeroplane crash? Why did the economy tank?

**Stakeholder21_Judge** 1:46
Sure.

**Interviewer** 2:05
And so what I ended up doing is creating a taxonomy and a framework. A taxonomy of explanation from the social sciences and a framework for explainable AI engineers to use to determine exactly what type of explainable AI model to build. It's not just simply looking at "Does it function and how can we delve deeper to get it to function better"? So, I'm testing out this framework using this case study and interviewing people. I'm looking about gathering in depth information on the subject of explanation from the perspective of humans. So, it's not a test for you. I'm simply gathering your inputs, perspectives, thoughts, questions and opinions about a real-life scenario, and there are no right answers. So, everybody take a deep breath,

because one of the things I find is everybody just gets anxious and they want to answer questions. So, is that OK?

**Stakeholder21_Judge** 3:06
Yes.
**Interviewer** 3:07
OK. All right. I'm going to go through the case study scenario. If you want, I can copy it into the chat. If you need to refer to it, or we can just go forward,
**Stakeholder21_Judge** 3:19
Sure, sure.

**Interviewer** 3:25
OK. So let me just put it into chat. Copy the text. Here we go. It'll come in two chunks. Chunk one …Chunk two…Here you go. OK, so this scenario involves the real-life case within the AI application domain of automated vehicles or AVs. So, it involves the occurrences of actual car crashes of one particular AV brand called Tesla and its advanced driver assistance system called Autopilot. So Tesla's Autopilot controls the steering, the braking, the 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. Now the case study is based on USA NHTSA, which is National Highway Traffic Safety Administration's Office of Defects Investigation. So, between January 2018 and January 2022, 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 that were attending to pre-existing collision scenes. So, police, ambulance, fire trucks, road maintenance vehicles, lights flashing, people with highly visible vests on. And what Tesla's Autopilot did was just drive into that and on average into the in these crashes, Autopilot aborted, vehicle control less than a one second prior to first impact.

**Stakeholder21_Judge** 5:24
Yeah.

**Interviewer** 5:28
OK. So, any questions about this scenario?

**Stakeholder21_Judge** 5:29
No.

**Interviewer** 5:44

OK, so this questionnaire is going to be one big question and then some sub … secondary questions about your particular stakeholder category of legal expert.

**Stakeholder21_Judge** 5:43
OK.

**Interviewer** 5:44
OK so, the main question is so based on the scenario. You're seeking explanatory information about these car crashes from Autopilot, the system that controlled the steering, braking and acceleration functions. When you ask why did this car crash happen, what are you thinking in your head? What kinds of information do you have that you're, … you know, in your mind that you are looking for?

**Stakeholder21_Judge** 6:14
Is this in? Is this in a case that I'm hearing like is this in a trial?

**Interviewer** 6:21
No, this has not gone to trial at all, no.

**Stakeholder21_Judge** 6:25
So.

**Interviewer** 6:26
This is based on… this is the investigation that NITSA conducted.

**Stakeholder21_Judge** 6:30

Oh, but what I'm saying is you're asking me, what would I like to know? But as you said, you know, explainability like what? What constitutes a good explanation?

**Interviewer** 6:34
Yes.

**Stakeholder21_Judge** 6:44
Depends on context. It's highly contextual. So are you asking me what do I want to know from Autopilot as a judge, hearing a case? Or are you saying what information do I want from Autopilot, as you know, an interested consumer of news media? Or are you asking what do I want to know from Autopilot from the perspective of a family member pf somebody who was injured by Autopilot in one of these collisions with emergency vehicles? Like, I think what constitutes a good explanation is going to depend on who you are and what your relationship to the material you're seeking an explanation about is and why you you're looking for the explanation.

**Interviewer** 7:36
Exactly. You've hit it on the nail. My research topic, yes. I'm asking you as a judge.

**Stakeholder21_Judge** 7:43
All right. So presumably there's a case maybe of court action against Tesla for these Autopilot accidents. Somebody who's injured or something is suing Tesla?

**Interviewer** 7:58
Yeah, in these cases, in fact, some police officers were injured. Some firefighters were injured. There were no fatalities in these in these 16 crashes. And what I'd like you to focus on is these 16 crashes because as you are aware, AVs are crashing for all sorts of reasons in all sorts of places. So … this particular case study, … I chose because there's a pattern of really highly visible things on the road and it's… it's official, it's not anecdotal and NHTSA did the investigation. So in these 16 cases, if you're looking at this pattern over this four year period as a judge, what kinds of information would you be looking for?

**Stakeholder21_Judge** 8:35
Right. But in the context of a civil suit in the context of a criminal suit? Like what are

... what I need more contextual information to know what's relevant to an explanation.

**Interviewer** 9:02
Let's start with a civil suit, and then I'd like to know how you would... how you would address it if it were a criminal suit? Because I'd like to start with the civil suit first and then a criminal case.

**Stakeholder21_Judge** 9:17
I mean, I think if....if.... So in a civil suit, let's say an injured party is suing Tesla for injuries caused by these autopilot accidents, I ... I would be looking for information, ... explanations that are relevant to the question of civil liability. So, is it is... the functioning of Autopilot? Did it represent a departure from the standard of care that would be expected? And then I think there would be an interesting legal question about whether the appropriate threshold or the appropriate standard is. Whether Autopilot needs to ... and, you know, self driving cars, automated vehicles need to be held to the standard of care of a reasonable driver period. Which would typically mean a reasonable human driver, or whether there's a different standard of care for.... for driverless cars, for automated vehicles, but... but that sort of takes us a bit, I think, afield of what... what you're thinking about. So I would....

**Interviewer** 10:46
Yeah, let.... Can we hang on? Can we? I like that point. Can we focus on assuming this is an agent acting in the world without human intervention? That's what they're selling. So let's, ... let's say, OK, standard of care equal to if it were a human driver. And look at it from that perspective.

**Stakeholder21_Judge** 11:00
Right, right. OK. Alright, so I mean I think sort of at the most basic level, I would be interested in knowing how? Like what the explanation is for Autopilot driving into these very visible emergency vehicles at roadside? Like why do you say this happened? Is this... was this a malfunction? If it was a malfunction, why and how did it happen? What...what is like... what are... how does this system make decisions on the road? In so far as you can say, it's making decisions. I mean, it's not making decisions like humans are making decisions, but it's receiving inputs and processing them and making operational choices. I suppose based on the input. So you know, I'd be interested in information about how that process happens with Autopilot?

And what went awry here? I would also be interested in, and this may be beside the point for… for your purposes, but I would be interested in sort of what quality control Tesla does with Autopilot, what kind of testing they've done.  What the, you know, expected rates of error and accidents are? Whether this deviated from what they expected? Really, I would want to know how it's programmed. I'd want to know how it operates, how it receives information, how it processes the information, how the information it receives. In effect, how it drives. And how that…allowed, that process allowed it to drive into visible first responders at an accident scene?

**Interviewer** 13:30
Parse your how it makes the decisions or how it processes the inputs and come ups with the output.  So, let's talk about functionality. So there are various functions going on and there are various algorithms performing those functions. So there's … as… as you may be aware, it's not one algorithm, it's an algorithm for perception. It's an algorithm for steering, an algorithm for speed control and….

**Stakeholder21_Judge** 13:47
Right, yeah.

**Interviewer** 14:01
an algorithm for motion planning and an algorithm for actual. … combination of all of these motion planning. … So which of these algorithms are you asking this question of?

**Stakeholder21_Judge** 14:16
I think I would initially want to hear some information about what algorithms there are and what algorithms are responsible for what and which algorithms were the ones that were, which one or ones were the ones that were responsible that … that, you know, screwed up here and were responsible for these accidents?

**Interviewer** 14:46
Well these are black boxes. Even the engineers themselves don't know the answer to those.

**Stakeholder21_Judge** 14:52
Right.

**Interviewer** 14:53
So ask more specific questions so they can. Yeah. Would you ask more … what kind of more specific, … specific questions would you ask? So they can go dig into whatever they're digging into to give you these answers because these are generic questions you're asking of them.
**Stakeholder21_Judge** 15:15
Yeah, it's difficult to ask more specific questions when you don't know anything about the algorithms or how they work, right?
**Interviewer** 15:20
Yeah, yeah. OK, fair enough. So, let's go back to your first two choices and the second one where you're saying, is this a criminal case? So, if it's a criminal case, what questions would you have there?
**Stakeholder21_Judge** 15:37
And I think in a criminal case, you know the … the standard of proof is higher, so there's well not only a higher standard of proof of wrongdoing, but there has to be some culpable standard of fault that goes beyond mere negligence, right? So you know, Tesla could be held liable in tort for mere negligence, a departure from the standard of care of a reasonable driver, say in the circumstances. Whereas in a criminal case, I mean, I think there are a couple of interesting questions. I'm just I'm just kind of spinning this out in my mind as you ask the question xxx, but
I think there are a couple of interesting questions. And the first one is about like, can Tesla be held criminally responsible for the actions or failures of its tool that it's created and marketed? And if it is going to be held criminally responsible, there would have to be proof beyond a reasonable doubt. Not only that the… that Autopilot committed the acts of a … a crime, say, dangerous driving, or reckless or careless. I mean, in Canada, in Ontario, careless is not a crime, it's a quasi criminal, but let's say reckless driving or dangerous driving. And so not only the act, but there would have, there would have to be some blameworthy fault state. Or you know, what's called the Mens Rea, the state, a guilty state of mind, which doesn't necessarily only refer to intention or knowledge. It could also be criminal negligence. So a culpable failure to advert to what a reasonable person would have adverted to in the circumstances. So there, I would want to know again like basic I think details around how did this happen? What's the decision-making process or processes that… that led to this kind of driving? And you know that had these consequences of

injuring people ploughing into visible emergency vehicles and causing property damage and injuring people. So what… what were the processes that led to these to these accidents?  But also, was this foreseeable by the company? Was this something that they either adverted to as being a real risk? Or something that they didn't … didn't advert to, but they should have adverted to given the testing of the tools? And what… what they did know about it? That may be getting a bit far afield from explainability questions, but maybe not so much.

**Interviewer** 18:57
No, no, no, not at all. Not at all. Keep going. Keep going.

**Stakeholder21_Judge** 18:59
OK. Yeah.
**Interviewer** 19:02
OK. So, one question, I'm a layperson. What does adivert mean in legal terms, you use the… you use the verb adverted to.

**Stakeholder21_Judge** 19:12
Yeah. So, advert means to be to have some awareness of.

**Interviewer** 19:24
OK.
**Stakeholder21_Judge** 19:25
So it might be something less than like full out knowing you know.  You know what you had for breakfast, but you know where… where we're talking about risk and probability, the law recognises kind of different levels of knowing. There's full knowledge or full intent, you know full knowledge of a set of circumstances or full intention to bring about.

**Interviewer** 20:07
Mm hmm.

**Stakeholder21_Judge** 20:07
Talking about the risk of a harmful or prohibited consequence being brought about, someone might not know with certainty or even full awareness that their actions are

going to cause a consequence. But they might be aware of a risk. And so like when somebody adverts to a risk, it's kind of like saying they're they've got some sense that this risk is live. While they may not…..

**Interviewer** 20:32
Hmm.
**Stakeholder21_Judge** 20:50
You know, be fully in knowledge.
**Interviewer** 20:51
No, it's like … it's like … I'm an engineer and I took the iron ring ceremony oath. If you … if you're aware of it. And we actually swear on a … we hold a chain attached to a piece of the bridge that collapsed in Montreal, as you may recall, a famous case from back at the turn of the 20th century, where it was an engineer who was, in your parlance, adverted to… knew… had knowledge that the bridge was shaky. People died, so that's something that's ingrained in us in Canada, at least the engineering school. So it's advert. Yes, I understand. He didn't intend on the design being poor, but he knew the risks, right? And he was careless. He was careless.

**Stakeholder21_Judge** 21:28
Right, correct. Yeah. Of course, of course, right. And so right. And so a lot of law when it comes to, like, the kind of law that's about assigning or apportioning responsibility for harms looks at when is it appropriate to hold somebody responsible for running risk. And there are different ways of framing that, like in criminal law, there's a concept of willful blindness, and that's where somebody didn't know, or the prosecution can't prove with certainty that somebody absolutely knew. But it will be enough in some circumstances legally to establish that they were willfully blind. In other words, there were enough circumstances that should have alerted them to a high risk. And they, you know, it can be inferred from the circumstances that they made a choice to turn a blind eye and run the risk anyway. And so… so there are a bunch of different kinds of ways in law of … of characterising the running of risk and determining when a decision to run a risk is going to be culpable, whether in tort, like for civil responsibility or in criminal law. And in… in criminal law, we will hold people, although there's a very strong commitment in Canada to personal fault and personal blameworthiness before we engage the stigma of and the, you know, the heavy sanctions of the of the criminal process and

criminal punishment.  We do sometimes hold people criminally responsible for high degrees of negligence. So there is a concept of criminal negligence and I think. You know, there isn't a lot of law yet out there about criminal responsibility for the harms committed by AI tools.  There's,... there's a lot of scholarship about it and some of the scholarship looks at like when can you hold a maker, a coder, a user of an AI tool criminally responsible for the harms that an AI tool commits? And then there's also some ...some discussion, most of it at this point, still pretty whimsical and fanciful, but there is some scholarship around like when can you hold an AI tool itself... or is it possible to hold an AI tool itself responsible criminally?

**Interviewer** 24:35
You can't put these. You can't put these cars in jail.

**Stakeholder21_Judge** 24:39
No, you can't put the cars in hail, but there are I... I actually wrote a chapter in my book litigating artificial intelligence about you know, kind of looking at the question of could we attribute criminal liability to autonomous agents to ... to AI tools? And, know, ... you know you can't. You can't, it would be meaningless in some senses, punishment of a tool is... is meaningless, but there are penalties that have been talked about and bandied about, which include, you know, wiping the programme, turning it off. And I think mostly looking at criminal responsibility for AI tools is less about punishment and more about the symbolic effect of saying that an autonomous agent can and should be responsible for the harms it commits.  Whether or not you can ultimately meaningfully punish them, but anyway. So, in terms of the question of like if, if there were a criminal suit trying to hold Tesla, for example, criminally responsible for these accidents in your case study, I think I would want to know. I mean, I think in either case whether tort or crime, I would want to know what did the company know? About what did the testing of the tool show about how it would react in road conditions? What were the risks that were known? What assumption of risk? Like what ... what choices were made in terms of putting Autopilot out on roads?  What assumptions of risk were knowingly made? And yeah....

**Interviewer** 26:42
OK. OK. So, one more thing with respect to these assumptions of risk. Can you talk to me about .... well, a simpler question for you might be, are you worried about any particular function with respect to driving?  The visual? The motion control? You

know the speed control, the acceleration, deceleration gas pedal? You know, steering? Are you particularly focused on any one of those or is it just all of them in this in one umbrella? Under an umbrella?

**Stakeholder21_Judge** 27:25
I think it would depend on… it's a bit hard for me to answer that question in the abstract, not to because I think I would want… want to know more about what failed in these cases. You know what I mean? Like, I would want to know like what system was it that failed? Was it the perception? Was it the braking? Was it the decision-making? Was it the interpretation of the inputs that it was perceiving on the road? Like what system failed? And … and once I knew what system failed, then I would want to know more about that system. And it may be like if this were a class action suit where, you know, injured parties in all sixteen of these accidents over four years were all jointly claiming against Tesla, it may be that different systems failed in different accidents.

**Interviewer** 28:02
OK. Yeah.
**Stakeholder21_Judge** 28:26
But like I… I think I would want to kind of have a big picture understanding of what failed and then probably go deeper on those particular systems. And I would also be very interested in knowing more like your scenario adverts to the fact that… that the Autopilot can disengage and hand over controls to the human driver. So, and,… and you've got in here….

**Interviewer** 28:59
Which you did, which you did less than a second before crashing.
**Stakeholder21_Judge** 29:03
Right, right. And so I would want to know more about that system in particular, because it seems to me that that system exists as a fail safe right? Like if…
**Interviewer** 29:18
I … I… I… hang on, I wouldn't categorise it as a fail safe, but I … I… I won't go into those details until after we finish the questionnaire. Just… just a heads up on that, yeah.

**Stakeholder21_Judge** 29:18
OK. OK, OK. Like it would, it would seem to me intuitively and I gather from what you're saying that I'm wrong, but it seems to me that that exists in the event that the system feels it's getting itself into trouble.

**Interviewer** 29:44
Yes.

**Stakeholder21_Judge** 29:45
It can disengage and ask the human to take over, so I would want to know how that works, how it communicates to the human that it's going to do so. How the… like how its internal mechanisms work or how its levels are set to … to trip the threshold of when it's supposed to do that because you know, if the idea is to hand control back to the human once it gets into trouble, it did that too late here, right? Like, I know in driving situations, one second can be a long time, but, …
**Interviewer** 30:24
Yeah.

**Stakeholder21_Judge** 30:32
Probably one second prior to the first impact wasn't enough time for the human to react and avert the accident, and so I would want to know more about.

**Interviewer** 30:41
Yeah.
**Stakeholder21_Judge** 30:45
How its controls are set? How it … how and when it knows to trigger that how it's intended to communicate that to the human driver? Because I think … I think that would be important and whether you know what Tesla had to say about whether that malfunctioned in these cases.

**Interviewer** 31:05
OK, I'm going to take a breather, and then we'll come back to this. There's a lot to go over in what you said, but OK, so just a couple of secondary questions and then I'll go through why I'm asking those because some of these… the reasons for me asking you the secondary questions will go into quite a bit of what you've just talked about.

**Stakeholder21_Judge** 31:13
Yeah.
**{Secondary Questions and General Discussion}**

**Interviewer** 34:23
Yeah. OK. So this explanatory information we're talking about, why would you want it from an AI agent? Trust, transparency, other reasons.

**Stakeholder21_Judge** 35:00
So in the context of a case, I would want it like, you know, as a judge, hearing a case, I would want it for…

**Interviewer** 35:03
Yep.

**Stakeholder21_Judge** 35:11
Transparency and.
**Interviewer** 35:12
Mm hmm.

**Stakeholder21_Judge** 35:14
Assignment, I mean, I think it would be essential to determine the issues in the case. You know to assign fault apportion blame a fixed damages. And, I also think it's very important for due process reasons.

**Interviewer** 35:34
Hmm.

**Stakeholder21_Judge** 35:34
So you know it's and… and I mean, I'm not the first person in the world to have made this point. Legal scholars have written about this a lot, but the black box problem with AI tools creates due process challenges because it's very hard to cross examine a black box, right? It's very hard, like in a human driver accident. You can put the driver in the stand and say, well, how much sleep did you have the night before? And

were you drinking? Are you on any medications? Were you wearing your glasses? Was the road slippery? Was it foggy? Were your … was one of your headlights out? When's the last time you had the car serviced?

**Interviewer** 36:30
Right you are. You are able to query this agent and yet you're not able to query this other particular agent, right? And…

**Stakeholder21_Judge** 36:30
Exactly. So when it's a human driver, you can cross examine and you can get information that helps you….

**Interviewer** 36:43
Mm hmm.

**Stakeholder21_Judge** 36:47
Understand what transpired and from the perspective of the judge hearing a case helps you understand how do a fix blame to make decisions about fine facts and make decisions about the legal issues that arise out of an accident like this? But when it's an AI autopilot, automated vehicle, …
**Interviewer** 37:06
Yeah.

**Stakeholder21_Judge** 37:09
There's no one to cross examine.

**Interviewer** 37:12
Right.

**Stakeholder21_Judge** 37:13
And there's no one to hear on those issues.

**Interviewer** 37:17
So that's why you …so is that, would I be correct in assuming that's why when you began talking about blame or responsibility, you focused on the designers

themselves? On Tesla the ...the manufacturer, because the agent itself, you ... you went straight to the designer of the agent.

**Stakeholder21_Judge** 37:39
Yeah, because I don't. I mean, it may be that there's some recording device on the agent.

**Interviewer** 37:44
But yeah, there was, but it's...

**Stakeholder21_Judge** 37:50
Yeah, and it may, I would certainly be. I mean, I didn't say it before, but yes, I would be interested in knowing what information is on the agent. It's on the, you know, automated vehicle recording device. I would want that too.

**Interviewer** 37:50
Yep.

**Stakeholder21_Judge** 38:08
But I think in terms of fault, ultimately, yeah, that's why I would.... I would want to know from Tesla also.

**{General Discussion}**

**Interviewer** 49:44
What questions do you have to this agent for? You know, because, you're asking a human being a lot of questions, and yet you're not asking this agent the same questions?

**Stakeholder21_Judge** 49:57
I mean, you know, I ... I don't know to what extent it's able to answer. I ... I gather, as you said, there's some kind of recorder, so you know, I would be interested in looking at what ... what it perceived when it perceived it. What... what it thought it was doing in the moment and why? And, why it didn't abort vehicle control earlier? When ... when did it perceive that it was in trouble? To what extent was it in control of the vehicle in the period lead leading up to the collision? To what extent was it in control of the vehicle at the time of the of the collision? Like with one second to

impact when it aborted vehicle control, was it was that enough time for the human to assume control in any kind of meaningful way, or was it so late in the game that nobody was neither the agent nor the human was effectively in control of the vehicle at the time of the collision? I think that's it.

**Interviewer** 51:27
That's it. Yeah. OK. OK. Thank you. These are these are wonderful answers and you've given me a tonne of data. So thank you for this.

**Stakeholder21_Judge** 51:37
My pleasure. My pleasure.

**End Transcription for analysis .... general discussion continued until 00:53:36 when Interviewer** stopped recording and transcription