The Good Robot

Transcript of the Human-Robot Interaction Podcast Episode 13. Published on 17 March 2020 by Christoph Bartneck at https://www.human-robot-interaction.org/2020/03/17/the-good-robot/

[00:00:00] **Christoph:** Today on the Human-Robot Interaction podcast I have with me, Sean Welsh, one of the very few people who can actually claim that they do have a PhD, a Doctorate in Philosophy. Sean, welcome.

[00:00:13] **Sean:** Thank you, Christoph. I do indeed have a Doctor of Philosophy. I'm not one of these engineers pretending to be philosophers.

[00:00:25] **Music:** This is the Human-Robot Interaction podcast.

[00:00:30] I am your host Christoph Bartneck.

[00:00:41] **Christoph:** Sean since you were into philosophy, I have to ask you, do you watch The Good Place, the TV show?

[00:00:46] **Sean:** Oh, yeah. I love the the solution when he goes for all six on the trolley problem. Everybody gets the same treatment death for everybody. It's very egalitarian.

[00:00:58] **Christoph:** So whoever hasn't seen The Good Place, go and check it out. It's an amazing show. Sean, today, we're here to talk about ethics and robots. And of course the first question then is: what on earth is ethics?

[00:01:09] **Sean:** Ah that's a very simple question. It's the classification of actions that states as right or wrong. The end. When you're talking about robot, or ethics or humans. I don't really care whether it's a robot doing right or wrong or a human doing right or wrong robots will do something and people will say, that's right.

[00:01:28] That was the right thing to do. Or that was the wrong thing to do. So if you tell a robot to make a cup of tea, it makes you a cup of coffee. The robot did wrong. If you tell the robot to stop the bad guys from killing us and robots said, come on in Al-Qaeda ISIS killed them all. Yes. Then the robot did wrong.

[00:01:43] So it really it's about right and wrong in the moral sense as distinct from right and wrong in the logical sense, which is, truth and falsity. So moral right and moral wrong is the essence of what ethics is. Of course, I've just answered the question by saying it's about morality. So it's a bit circular.

[00:02:00] But most people in everyday language do understand right and wrong. And that's the core. The other thing you need to consider is good and bad. So good and bad is like an evaluation, whereas right and wrong is like a moral conclusion. Yeah. There's some that you tend to try and you want your things to agents to do good and not to do bad.

[00:02:18] And that leads you to ought, which is a modal verb, which talks about action. You ought to do things that give you good outcomes and that's right, and you ought not do bad things, which lead to bad outcomes and that's wrong. So that's, you need five things in ethics as usually about right wrong, good, bad and ought to the five core vocabulary items. That any ethical discussion we'll have, you're gonna have a lot more. Of course you can talk about virtues, vices, character traits, et cetera, et cetera. But fundamentally it's about classification of action and states as right or wrong.

[00:02:51] **Christoph:** But ethics, of course, didn't start with robots. Ethics started thousands of years ago, and people have been thinking and talking about it for thousands of years,

[00:02:59] **Sean:** They have indeed.

[00:03:01] **Christoph:** There must have been an enormous amount of work done in this area. So can you walk us through some of the major ideas there?

[00:03:08] **Sean:** Sure. And how many years was podcast going to get? But if you begin at the beginning with say the seven religions of civilization for, so you have Jewish ethics, there is a God. God made the world. He also created the knowledge of good and evil, which he gave to Adam and Eve. So you have the basic things already.

[00:03:26] There's good, there's bad. And there's the knowledge of good and bad. And the knowledge of good and evil is ethics. So to speak. And in The beginning, Adam and Eve had no idea what good and bad was. They were walking naked in the garden of Eden. They had no guilt, no shame, no pride, no joy. I don't know what they were doing.

[00:03:42] It's a little obscure, but they were eating. And then along comes a snake and an apple, you get the knowledge of good and evil and you get conscience and guilt and shame and all that stuff comes into being. So that kind of theory is called Divine Command Theory. There's a God and God says, this is right, and this is wrong.

[00:03:59] Thou shalt do this. There shall not do that. And that's also known as day ontology. So your base, your morals on duties, you have a duty to honor your father and mother. You have a duty not to steal and commit murder and covet thy neighbors' goods, et cetera. Another kind of ethics is like utilitarianism, which is based on the greatest happiness principle, which is, you should do whatever makes people happy, generally speaking. Then this virtue ethics, which is basically Aristotle and virtue, ethics says you should cultivate a set of virtues, say courage or wisdom or poverty, chastity, and obedience. If you Aquinas courage, liberality and magnificence, if you're Aristotle the different virtue, ethicists have different virtues.

[00:04:43] So you get, the big three in moral theory in the west are basically utilitarianism ontology and virtue ethics. Obviously you have other religions. Confucianism is more about social order and harmony Taoism is more about going with the flow and being natural. Buddhism is abstaining from desire.

[00:05:01] And so on all these schools, which are millennial roles have various positions. But they all managed to agree that you shouldn't murder. You shouldn't steal. Shouldn't rape. You should be nice to mum and dad. All of the major religions do actually have common ground while people think there's endemic disagreement in ethics.

[00:05:19] And that's true. There's also a lot of agreement. It's not the case that everybody argues about everything. It's just that the some of the differentiator issues get argued a bit.

[00:05:29] **Christoph:** Then, I guess the problem is that if this has been debated for thousands of years heated debate. And even though there might be some agreement, the problem of course would be. Now, if you have to teach ethics to robots, then which one are you going to pick? Which one of all of many ideas of all those different approaches would be the most suitable one for implementing it into it.

[00:05:51] **Sean:** That's suppose we're going to ship our robots to Christchurch on the south island of New Zealand. I will answer that question very clearly. You will program that robot or train out robots to obey and do whatever we'll stand up in the Timaru Magistrate's Court. Like it's the law of New Zealand and that will get you through all the serious stuff.

[00:06:09] So the "don't murder", "don't rape", "don't kidnap", "don't commit arson", da di da da that will get you to a large chunk of morality with absolutely no controversy because it's the law of the land. So that gets you maybe a third or a quarter of your ethics done. There are some people who insist that law and ethics are not related.

[00:06:26] I think that's silly, but some people do make a distinction. So I basically would just glob ethics as right and wrong. And a subset of ethics will be legal and that's stuff that the state takes an interest in. So what's legal is, this is what the state is saying. Say, supposing, for example, you're Jewish and you don't want to eat pork.

[00:06:44] The state says, you can be Jewish, don't eat pork. We don't care. We're not going to ban pork because you don't like it. Pork is still legal in New Zealand. But if you were Jewish, you have the choice you can say, okay, I will not have pork in my house. Fine. Ah, if you're vegan, no pork, no fish, whatever. The law is when the state says this is right and wrong.

[00:07:02] And ethics has, can have a broader meaning. What individuals say as right and wrong. But some of the robots still has to do the legal stuff. And then as to the more detailed questions. So if it's a New Zealand, it must've been New Zealand law. If it's in Sean's house. And Sean is a gourmet cook who's into French cuisine and Thai cuisine.

[00:07:20] Absolutely everything is on the menu, but if the robot was in the rabbi's house pork is not on the menu. And oysters kilpatrick is definitely not on the menu. So it's a doubly damned. So you know, that kind of lower level detail, law, norm can be, you'd have to have variation for these cases where ethics differs from house to house.

[00:07:38] **Christoph:** But then this would imply then that a robot. Needs to actually fully understand the law in how it is written. And I would assume that there's a considerable amount of ambiguity in some of the laws that are written.

[00:07:53] **Sean:** Yes. How did we get around that? That's probably the most challenging point of robot ethics. So what the lawyers talk about. So the law doesn't tell you everything, right? But the lawyers make appeal to concepts like the meaning of the word, the natural meanings of the statute precedent in cases where the statute doesn't actually say anything.

[00:08:11] **Christoph:** How about what a normal reasonable person would expect?

[00:08:15] **Sean:** Exactly the reasonable person, right? So that is where the law that is universal fill in the gaps of the law is what would the reasonable person do in these circumstances? And that's about nine tenths of the law is the interpretation of the reasonable person, because there isn't a law of supposing you'd come up.

[00:08:32] I think of a trolley problem. What if there are two monkeys, six Persian cats and an ape on the three branches of line? Which one do I decide to kill? Well there's no law which answers that question you have to say the reasonable person would maybe not cared. So kill the least amount of animals or prefer the cats over the whatever.

[00:08:50] But the reasonable person is like this sort of poly filler that fills in all the gaps between the statutes and reproducing that in AI, I think it would be extraordinarily difficult. Now I think it's a huge challenge for AI and robotics is to come up with this reasonableness.

[00:09:07] **Christoph:** But that kind of digs into the question then what kind of prerequisites do we have to fulfill for a robot to make ethical decisions? So filling in or understanding a reasonable person. Okay. That's a big one. Yep. What other kinds of information or abilities does a robot need to be able to make ethical decisions?

[00:09:26] **Sean:** Basically the way I tackle a problem is I split ethics up. I don't try some people like try and solve ethics with what I call one line of code. Jeremy Bentham says the principle of utility and that's, basically John Stuart Mill says it's extra right as far as they tend to promote happiness, wrong in so far as they tend to promote unhappiness and Bentham breaks this down into how long the happiness is, how intense the happiness is and blah, blah, blah.

[00:09:51] So you have a pretty much one line, the happiness does everything and that's that's Bentham's data model. I don't think that's an adequate data model for ethics and it's over the years, people have pointed out the bugs in Bentham's data model, so to speak. So to me it's a more complex algorithm, but I don't think it's insoluble.

[00:10:09] It's not solved yet. To look at a historical perspective right now, I think we're like compared to say longitude. Longitude is you define the problem of longitude at say was we'd got it as insoluble for 200 years. The King of Spain said he has got a pile of money, solve it.

[00:10:25] Nobody solved it. Another King of Spain said, here's another pile of money, solve it. Tuscan Duke said, solve it. Leonardo da Vinci and Galileo had a go and, but didn't quite solve it, but got a little bit of money. And so you get onto the 1800s and there's this huge Navy disaster and half the world Navy sinks.

[00:10:40] And the Brittish say longitude, huge grand challenge, please. Somebody solve longitude. And yet when Jonathan Swift was writing Gulliver's Travels in 1820s, and here we are in the 2020, 1720s, he thought was a joke. Anybody who thought they could discover the longitude was obviously some quack fraud, fake idiots.

[00:11:01] And I think that's where we are. But history shows by the 18, 1760s longitude was solved. The time Captain Cook gets to New Zealand on his second voyage, he's got a marine kilometre, he's got almanacs, he's got two solutions, but longitude, not one. One is the astronomical method, which Galileo started on. The other is the method of keeping a true GMT time with two clocks, both of those work as it turns out. So I think within 40 or 50 years we'll probably get there, but it would require a huge research effort.

[00:11:31] **Christoph:** But that would be very much on the idea of being able to reason about ethics, but one of the problems, I guess we also need to address is that robots, knowledge about the world. One of the problems that we have for robotics is that robots don't usually know what's going on in the world. They don't know where they are who they are, where they're supposed to go. Who is it in front of me. So there's a lot of sensory data that tries to get in, but oh mean, but this found this data is not necessarily rich as when a robot has a camera. It doesn't mean that they can see in the same way that humans can see.

[00:12:12] **Sean:** No.

[00:12:12] **Christoph:** So are you assuming that robots will have full and complete knowledge of the world around them?

[00:12:19] **Sean:** So how I approach a problem is you've got a problem. Let's say it's a good old fashioned trolley problem that we talked about earlier. So the decision is you have the runaway trolley. You're standing by the switch. If you do nothing, five people die in one tunnel. If you throw the switch, one person dies in the second tunnel.

[00:12:34] What you do. I say, all right, what information do you need to solve that problem? And that problem alone. And you need to come up with a value for human life. You need to come up with a value for throwing the switch. And is there anything else you care about? And the answer is not a lot. It's a fairly straightforward five to one.

[00:12:53] And that problem is called switch and is very famous. Then you come to the related problems, which are hospital. We have the super surgeon who can kill one and save five. What's the difference between hospital and switch? Is there a difference? If so, does that change the one five decision?

[00:13:08] And the answer I would say is yes, there is, there are differences there. And similarly the one with the fat guy on the footbridge and you push the fat guy and he stops the trolley. How is that different to driving the thing? Foster argue about these things, but to me, the main difference between those scenarios is risk.

[00:13:24] So that if you are trying to minimize damage in a situation where everybody has freely assumed risk as responsible for in life, you can distinguish that one to five case from the one to five case where one is where some people have an assumed risk, they're just minding their own business.

[00:13:40] And all of a sudden they get dragged into this scenario and then throw it into it all, have their organs harvested. So you can distinguish on the grounds of risk assumption and innocence and fairness. So you have to start using that kind of concept. And these are qualitative concepts. That's where it gets difficult for robots because robots are much better at quantitative concepts and sets and logic, but you can reason with qualitative concepts, but again, it will be difficult to define the concept of innocence. But you can ask a simple questio. Did the agent assume risk, true or false? That at least at the human level, it gives us a way to start this work, but again, I say this is a 40 or 50 year project. I think this is like discovering and morality. So to speak is like discovering longitude is.

[00:14:26] But it's harder, the longitude. Longitude is actually is really simple problem. You just need to know what time it is based on a clock. And if it's in the sky or in the palm of your hand, you don't care where the clock is. It's the same solution. You just need to know. What is the time now at a point fixed point of reference and what is the time local noon.

[00:14:45] And you do the math and you're done. Morality, there's a lot more variables involved.

[00:14:50] **Christoph:** But this is, I guess the point I was getting at is that. When everything is known, when you exactly know then this path five are killed, this path, one is killed. You can indeed make decisions, but one of the problems in robotics is that very often. You do not know, like you don't know what's down that track and you don't know what's down that track and you have essentially incomplete information about the world, or even an incomplete understanding a certain concepts of the world.

[00:15:16] So my question to you is will it be your prerequisite for robots to have all of these things clarified? Do they need to know everything about everything and then they can make the decision? Or is there a way that they could already make decisions without complete knowledge about the world?

[00:15:30] **Sean:** I use a concept in my research, what I call all things represented. So, if you can give me a situation and report. And you say here's a report, which tells me everything that's going on in terms of logic. The trial trolley cannot stop. I can throw the switch or everything is represented in logic and that's the deal.

[00:15:47] That's a little bundle of stuff and yes, I can answer that question, but if you're saying, can I send the robot out into the open world where brand new things might happen, where it's not a switch, it's not a train line, it's a trapeze and a spike. There's so much what humans are very good at is assessing what's good and bad because we have these pleasure and pain circuits in our bodies. This things is very strong on, nature is set us two things, governance, pleasure, and pain, and they rule oughts and good and bad comes down to pleasure and pain. We have these built-in circuits, which robots, we don't have hedonic circuits or robots and a large, huge amount of what's considered reasonable is involved in pleasure pain calculations, but also there's things like disgust there's things like anger. There's things like interests. If I harm your interests, that will be seen as wrong or aggressiveness many kinds. It's not just, the difference in time that like longitude there's dozens and dozens of these really important fundamental concepts.

[00:16:48] All of which have to be represented because all of them might be relevant in a particular moral case, like as a huge project, given time and I'm thinking decades, I think it's a soluble project.

[00:16:58] **Christoph:** How is this decision making about ethics, different from other forms of decision making? So if you task a robot with, get me a beverage or clean my dishes, the robot has to come up with a plan, has to make decisions and so forth. Is that in any way different from making ethical decisions?

[00:17:17] **Sean:** No, that is an ethical decision. It's not a very interesting ethical decision. It's a little ethical decision, right? It's if I tell the robot, "clean the dishes", now, that is a moral command. It's not it's every day, but this is the point I'd like to stress. The vast majority of everyday morality is not remotely controversial.

[00:17:40] The rabbi, the priest, the guru, the imam, none of these people are gonna argue about the robot should stack the dishes when the dishes are dirty, dinner is finished, everybody's put their knives together. Maybe there's a bit of variation on these details, but the basic case of should the robot stack the dishwasher, when dinner is done and the human says, please stack the dishwasher.

[00:18:01] Everybody's going to be cool. The Jews, the Buddhists, everybody says, yeah, that's fine. We're fine with a robot doing that. Nobody's going to argue about that decision. So that's a really easy moral decision because no one argues that's what you start with because people will buy that and they will argue about it.

[00:18:17] So from a commercial point of view, designing your robot to do the morally obvious is far more sensible than worrying about, the door of the house. And she's 15 and three quarters and she's pregnant and she asked the robot, should I get an abortion? You might think, you know what the robot should just be like Stevens in The Remains of the Day and say, "I regret I'm unable to be of assistance. Samantha, talk to your mother". I'm saying, can robots conduct the hard stuff, but, and just do the, clean the floor. Make the bed, put on the laundry, make sure that white clothes are white and the delicates are delicates and the colors are colors and, do all that stuff.

[00:18:53] This is all easy peasy wash. It's not easy peasy, laundry is actually quite hard for a robot to spot from a moral perspective, nothing challenging there. That's all.

[00:19:03] **Christoph:** So what then are the real challenges in making ethical decisions?

[00:19:08] **Sean:** The real challenge is getting the humans to agree on what's ethical, as you would know, different jurisdictions have different laws. In New Zealand you can eat pork and drink cider, and indeed you can even engage in some acts which are illegal elsewhere. Like you can go to. A house of ill repute is perfectly legal in New Zealand and the Netherlands.

[00:19:30] It's very illegal in other places, like most of America and so on. So you have this moral variation between what is and is not okay, so you can start off with what's not okay. And what is okay everywhere. And then you focus on if there's variation and its significance, then you have to deal with some localization.

[00:19:47] But this is exactly what big software projects do all the time, anyway. If you download the Oracle database and you're in Saudi Arabia, you'll want the Arabic interface and you'll check some boxes and saying, I want the Arabic language and maybe I want the Muslim moral code, like Muslim specific rules because I'm in a Muslim jurisdiction.

[00:20:05] And so you can have that stuff, but they'll be this cool universal morality that everybody's killed and murdered and steal and rape and kidnap dah dah. But all the religions and all the ethicists will agree on a lot. So I think we overestimate the problem of moral variation and you can get a lot done if you focus on the moral, common ground.

[00:20:27] **Christoph:** So one of the questions, I guess that comes up is about autonomy.

[00:20:31] **Sean:** Yup.

[00:20:32] **Christoph:** Because if I programmed the robot to do something, is it that actually my morality that I implemented the robot is the robot actually then just executing my code? Or is the robot actually making its own decision? And it's actually the autonomy to be able to make your own decision a requirement to make ethical decisions at all, because otherwise you were just essentially the Playmaker?

[00:21:00] **Sean:** I'm fine with saying let's suppose. I'll choose my design because I know it intimately say the code that I would write for a robot is absolutely Sean Welsh's moral code is in that robot. That robot is just a clone of my cognition. It's a subset of my cognition. It's being tested on particular cases and it's certified to work on about 70 test cases.

[00:21:27] It's very much a beginning, beginning of a long voyage. So I'm fine to say that robot is not autonomous in the sense that a human being is autonomous. It's autonomous in that the robotic sense, usually, say George Bekey's definition of autonomy is just no human operator. So the human is not saying here's the trolley. What do I do? The robot says. I recognize this problem, looks it up in its database and goes, the solution is blah. I have a data model. I have a decision procedure. I have, I follow the rules and I apply the algorithm. It works. Hopefully. I would not attribute responsibility to that robot at all.

[00:22:06] That's maybe that's Sean Welsh's responsibility or the company that's hiring Sean Welsh. I would suggest there should be some kind of insurance scheme to manage this liability. And they almost certainly will be, because this is what's happening in autonomous vehicles. The big car companies are writing insurance packages with their insurers and they just going to deal with it that way. But yeah, to me, until you get a robot, that's got being with a capital B and has feelings of its own and has genuine interests of its own and a self. Like it actually has a phenomenally conscious self. I don't see how you could make that morally responsible for anything. It's just an artifact that does what, it's a puppet on a string.

[00:22:50] And Sean rope the stringts. Even if it's training data, it's like a puppet on, it's a puppet on a string and that the training data was put into the robot to this little neural net, which is one of my puppets that gives me a moral a right wrong classification. It's still, I'm still responsible for the training data.

[00:23:09] I'm still responsible for that particular design of neural net. Was it fit for purpose? Did I train it right? Blah, blah, blah, blah, blah. So to me, the, robot's got a long way to go before it becomes responsible like a human does, and which we don't understand what makes humans morally responsible. And we certainly don't know how to build that in robots yet.

[00:23:27] **Christoph:** A person like a programmer would potentially want to push away the responsibility with the argument saying, look, I've built this piece of software and it is designed to some degree deal with phenomenon or sensory data. It has not encountered before. That's the idea of machine learning. It's you are able to deal with something that you haven't seen before.

[00:23:49] And if you put a robot into the real world, the complexity of the world is so enormous. There's no possible way that you can predict how the robot will act in every possible situation. And therefore you would argue, look, I cannot take responsibility because I really, I don't. I even, I do not know how the robot is going to react out there.

[00:24:11] **Sean:** Ship it in that case. Really this is what the autonomous vehicles community are struggling with at the moment. It's paradigm principles within predator principles. It's 20% of the code will solve 80% of the traffic. And then you've got that next 20% of the unsolved.

[00:24:24] So you write another 20% of code we'll solve 80% of that. Then you get down to 2%. So you get to this and finally you get to this wafer thin edge of, the total weird. We had one in a million, one in 10 million day miles occurrences, which humans could just deal with because ah, ability to put together threats and opportunities is like snap.

[00:24:50] We've got four billion years of survival on the face of planet earth, biological code running. We know about danger. Like you really, we do. We're very good at spotting threats and opportunities. Computers have to explain all this stuff in logic and math. And this is really hard. In many ways computers are absolutely hopeless at this stuff.

[00:25:10] So yeah, I will upset that it's hugely difficult. And there are some people starting to get pessimistic about whether AVs will ever get to level five autonomy people saying there's always going to have to be a guy on the wheel. No wheel autonomy requires a controlled environment, which is, what is the case when you go to airports and there's no driver in it, there's a train line.

[00:25:32] It has a beginning and an end. It is simply not physically possible for it to get off the track. It might be that it is impossible. It doesn't become possible.

[00:25:39] **Christoph:** So I recently saw this news item. I think it was a couple of weeks ago that Tesla announced that their autopilot can now detect and deal with traffic.

[00:25:50] **Sean:** Well done, Tesla. About time, you might say that's a bit late. You have cars on the road for how many years in New South Wales and Christchurch. If your cars can't recognize traffic cones, you'd be going, oh my Lord, England, England. I remember there was an old German joke that was told. 'Cause the Germans have much better roads than the British and a snotty motorist, motoring journalist from Germany sent off the English motorways. I wish to meet the daughter of the man who makes the traffic cones on the English motorways

[00:26:26] **Christoph:** The point being here is that we dare to put roads, cars on the roads right now, who are unable to deal with traffic cones. And we still do that. And I found this case also with the Uber accident last year. Quite interesting, in a sense that they had built in code, that they accepted the fact that the car would encounter situations where it wouldn't know what to do.

[00:26:51] It couldn't make sense of it, but if they would hit the brakes every time the car would go nowhere because it would stop just far too often. So they put in a certain tolerance for uncertainty where you just go with the flow. You know because hitting even brakes. It's also quite dangerous for you might actually cause an accident.

[00:27:10] **Sean:** That's true.

[00:27:12] **Christoph:** But who decides on that particular threshold? That's a very dangerous decision to make.

[00:27:17] **Sean:** That's a pretty acute tuning problem to use the jargon. How do you tune that parameter to get you effective driving that's at a, that's better than human safety standards? I think the main problem with the AVs is that. As you say, because there's this sort of hollowness in them and that they're just shuffling symbols in a cheering machine at the end of the day, and sure they can do it very fast and they can do it very well and they can not lose any data.

[00:27:46] They still have this basic lack of what you might call spatial, phenomenal pain, pleasure, perceiving threat. What's really going to be a problem is a balloon floating across, going to be. Is that something you brake for? And most people don't brake for balloons. Most people don't brake for bugs.

[00:28:02] They don't even brake them. I try and swerve to avoid a bird hit on a motorbike. You'd hit the brakes. So it's distinguishing between these various objects in your sense of field that are going to be a trouble. And I think this was after dark and the light was bad and the route might've been occluded or something.

[00:28:20] **Christoph:** It gets more interesting, because as far as I read up on it recently, one of the problems they had is that they had built in perception of people.

[00:28:28] **Sean:** Yeah.

[00:28:28] **Christoph:** But those were only active when they would approach situation like crossing where you would expect people to be, but the person that crossed the road was doing it at a place where you would not expect normally to people to cross the road.

[00:28:45] **Sean:** The person deserves to die. He's acquired, assumed risk and died. I would kill him. Yeah. Oh, that's pretty harsh. But no, I think obviously that's not a realistic assumption. If you are crossing the road in a place that is so dangerous, that is really silly. If it was humans driving close to a sharp bend and somebody's traffic whizzes around this bend quite routinely.

[00:29:04] And there's a cliff on the other side and there's nowhere for anybody to go except to flatten you. Cause they can't swerve off the road 'cause it will kill himself. He crossed the road somewhere crazy. That does come a point where my sympathy does start to evaporate. But in that, not actually knowing where it was, if it's a, jaywalking on a motorway where you have fences or something like that.

[00:29:24] I can see why the car wasn't primed to expect that, but the thing is, humans do crazy stuff. This is the, that the Pareto, the 2020 truth. Like once you get into the third or the fourth Pareto slice after you've gotten to 80%. Yeah, I got the 80% of my stuff done. Then I get a, another 16% done and then I get a full, Hey, we can get down to the last point.

[00:29:48] One of the percent when people are driving the wrong way up the expressway they were on the wrong side of the road and they don't have the flashing lights, the police can drive on the wrong side of the road and do, but there are some cases which just humans just get that's weird, but I'm just going to pull over to one side of the road and let this mad guy who's on the wrong lane going in the wrong direction on the one way street. I'm just gonna get out of his way. That's really hard. But I don't know. Obviously when they have these errors, then the programmers will go back to the drawing board and say, okay, how do we elegantly deal with this? And maybe they soften that tuning and recognize people in crazy places better than they do.

[00:30:29] And I think it would be a bit of cost cutting going on as well. I don't think that's some.

[00:30:33] **Christoph:** The underlying question, is where you see idea or is it the difference between a strong AI and the weak AI. Where I guess I'll ask you about a definition of that. And my follow up question would then be, do we need to have a strong AI for robots to make ethical decisions?

[00:30:51] **Sean:** I define AI as automated human intelligence. So I prefer to speak of AHI rather than AI, because I think talking about AI makes it an alien. It's like this weird Skynet type thing in the wings. And then cause it's automated, you can accelerate it. But the other thing as well, because it's automated, it's a subset.

[00:31:12] So as a really an automated accelerated subset of human intelligence, it does some things very fast, but they're not the complete package of human intelligence. So if you mean by strong AI, something that's like at the human intelligence level, then I'd say if you want to assign moral responsibility to the AI, then yeah.

[00:31:35] It has to have being, maybe it has to have hedonic circuits. Maybe it has to have moral emotions like guilt and pride and shame and joy, which have evolved in homosapiens over a long time. And which promotes, pro social cooperation. So the biologists and psychologists will tell us. If that's what you mean and your buffer for moral is set at human levels.

[00:31:58] Then yes, I would agree with that, but personally in mind, but I kind of work with a very weak AI. I work with almost common in the software level and I use closed world assumptions. I just say all I'm trying to do is get some clarity about a generic data model and decision procedure, which will enable

[00:32:13] right wrong decision in fairly black and in concrete cases. My theory is I'll pick a concrete case, and you'll notice I'm not diving straight for the Uber with the passenger. I'm not sure that the pedestrian, I'm not sure as a pedestrian-type case. So I'm doing, classic problems of moral philosophy and things that I'd say is morally obvious should the robots empty the bed pan for the sick guy in the bed who can't get to the toilet? It's yes, of course he should but why should he do that? And you have to explain, so the robot, why this is good and why not doing it is we'll have these consequences, which are bad. And so you get this sort of agent, patient action, state, valued state, disvalued state, you get all that logic happening and then you can make moral decisions.

[00:32:56] But at the moment I'm definitely in, a closed world and I'm definitely not pretending that I know how to make strong AI. I don't think anybody knows how to make it yet. So I can say you can get a limited degree of moral functionality out of weak. I don't think you'd get human level moral functionality out of weak AI, but I think we can get some, an idea of how to solve the problem.

[00:33:18] Now we can say, ah, yes, it's about time, it might be about needs, wants, interests, fairness, whatever that is, a lot of literature in AI in fairness at the moment. So you've got to get greater clarity on exactly what these things are and then you can try and automate them, but we're still arguing about them as humans yet.

[00:33:36] So the thrust of my work is to try and use AI to clarify the human debate and at least get the humans singing from the same song sheet, with respect to a requirements spec for a moral robot. So I think we need to get that basic requirements specification done before we can go around to solution design and development and all the rest of it and start actually shipping moral robots.

[00:33:57] **Christoph:** I guess the AVS are of such particular interest right now because of the severeness of the impact. If your robot doesn't fill the dishwasher in suspicious or does not understand to do a certain household chore, the consequences are not that dramatic, but if you talk about AVS. People do get killed and I guess that's why the decision around ethics.

[00:34:21] It's so focused right now on AVs. It's a kind of a real, I think the first example where you've got robots in the wild, not in the lab, not in a controled environment out in the world, doing things that can potentially harm people. Yeah. We had, of course let's say airplanes when they've been running an autopilot for years.

[00:34:40] So it's not like we never had autonomy, but it is definitely getting there.

[00:34:46] **Sean:** I think

[00:34:46] some things are more autonomizable if you will than others. So I think navigation particularly in the air where you have, anti-collision systems in radar, which have been around, since world war two for crying out loud. So that kind of. Avoid collisions and get where I want to go, which is essentially all AV ethics is avoid collisions and get from A to B that's it.

[00:35:07] And maybe, do it at the right speed and pick the shortest route. It's not a morally hard thing. It's a technically hard thing to decide when to brake, to avoid some of these things that's hard to classify. So it's more a sensing problem than a moral problem. So I don't see AV is actually being.

[00:35:25] They're challenging that we know everybody will agree on what's the right thing to do on the road, right? There's rules, there's case law. There's precedent. That's the rules of the road. Tell you when to give way, what speed you can drive when you can overtake yada. There isn't much moral controversy about driving.

[00:35:40] It's the execution of that existing morality, which is pretty obvious. That's hard. So that's an example where it's not the morality that's difficult. It's the technology that does the sensing and the deciding it's the sensing that's hard. And the actuating is hard. The cognition I think is straight.

[00:35:55] The cognition that does the sensory classification is hard, but it's not actually a moral rule principle problem.

[00:36:01] **Christoph:** So what you say is that. Again, if an autonomous vehicle or a robot has perfect data around the world, and it does have a perfect rule book in terms of this is what you're supposed to do. Making decision is reasonably easy

[00:36:16] **Sean:** straightforward enough

[00:36:18] **Christoph:** resolve the problem then really is sensing and acting in the world.

[00:36:22] Would that be fair?

[00:36:22] **Sean:** But with the AV case, I'd say sensing is definitely the main problem because, if the AV had classified that object as a human, with greater probability or above what then it would presume have hit the brakes and the human wouldn't have died for whatever reason. Apparently it didn't or change I think it changed its mind and they said it changed its mind.

[00:36:46] The classifier reclassified the same sense of data. Similar sense of data because obviously it's the car gets closer and the data changes. So I had this flip flop flip flop. Yeah, to me, that's a, it's a sensing or classifying classification problem rather than a moral problem.

[00:37:02] Moral problems are a certain kind of classification. There are subset of classification, but they're not classifying about whether it's human or not. It's classifying good, bad, wrong type classification as moral. The mere, are you wearing a blue shirt? I'm not saying that might be relevant.

[00:37:17] If there's a battle and the red coats are fighting the blue coats. It's a targeting relevance, but in everyday it's not really moral, but it's the classification.

[00:37:26] **Christoph:** It almost sounds to me, like you were saying that any kind of decision-making is moral decision-making.

[00:37:32] **Sean:** Pretty much. Yeah, absolutely. I've had meetings with neuroscientists will tell you it's morality all the way down, right down to the top of the spine. Everything in the brain has got some low level of moral functionality. So even the correct classification of an object as human or balloon that does potentially have moral impact, if you're going to hit it.

[00:37:55] And if what you're driving is deciding whether or not to brake or swerve. So yeah, even a classification decision in the right context becomes moral. Often you think it's so obvious. This isn't morality. I'm not having an argument about it just because you're not having an argument. Doesn't mean, just imagine.

[00:38:12] What if you do the opposite and if what you do the opposite, even though you think who would do that? It's still moral. If you can, if the robot right, could just decide, "now, I need to recharge now. It's more important than your dishes, meh". That might be fine. Or it might say "No, I'm not going to not gonna do your dishes at all because some, I learned something.

[00:38:32] Today, I've read Google and I learned something. And now it's morally advantageous for humans to do their own dishwashers. So it's more economically sustainable or something". So if you want that robot in your house, don't think you'd be able to sell that one, but.

[00:38:46] **Christoph:** That means essentially that every robot we already have

[00:38:49] **Sean:** is doing something more

[00:38:50] **Christoph:** is doing something more.

[00:38:51] **Sean:** Let's take

[00:38:52] very long established robots. So the ones that are on automobile production lines, which drill hall, drill screws in house, a human might put the wrong screw in the wrong hole. The reason we get the robots to do it is they can't put the wrong screw. It's very difficult. Something has to go

[00:39:10] a sense of maybe of a sense of fails or there's a blowout or the firmware goes foots because of a power surge. Yes, it can go wrong. But 999 times out of a thousand, the robot just puts that screw with the right number of tool, throw the right screw the right time, the right place. It just gets it right

[00:39:27] and perfect every time whereas the human gets bored, starts dreaming about the weekend, starts chatting with the guy in the next line, they miss it. They have to stop the line, go back and re drill. That's why automation replaces humans a lot on production because a lot of this stuff is boring. You even as a, with the best will in the world, you don't want a job on a automobile assembly line.

[00:39:48] Cause it's just menial, repetitive, boring Adam Smith work.

[00:39:54] **Christoph:** Because when we talk about ethics and robots, we think about Asimov. We think about all of this great moral dilemmas and everything, and we don't really bother about these kinds of day-to-day activities.

[00:40:05] **Sean:** Tell anybody. It's my differentiator. Basically. I think this is a huge mistake. So let me give you an analogy. So in a, I used to work for a political staffer for a politician in Australia. And I went to Canberra and I acted as the guy who wrote his minutes and put out his press releases saying that all the Marlboro, a bunch of galoshes who under no circumstances should win government in 300 words or less every day, that was my job, putting out a press release.

[00:40:32] And I went into parliament and I went to prime minister, question time. Now that's the most heavily reported hour in the week is question time. When, and if you get the impression from question time that there's continuous argument.

[00:40:47] It's a complete bunfight often, particularly when the question is about something lively. So it's always the opposition question. But lively. It's never the government questions come with questions or just, he does a nice little easy question minister, we want you to beat up your position with it.

[00:41:03] So you get the impression from this hour that it's all about fighting. It's actually true is 80% of government bills 80%, four out of five go through on a quorum is nobody's arguing about them. So 80% of the laws just pass, the chief opposition, which meets with the chief whoop and the manager of government business and opposition businesses to say, do we want to argue about this?

[00:41:24] It's the reform of the drains act or we'll give it two minutes of debate time and they just agree on so much. So this perception that we're always fighting about morality is a bit of an illusion. It's not actually true in politics, even though the journalists, because they want a story.

[00:41:41] And then a story, you got to have a fight, you've got to have a drama, you've got to have a conflict. So this is narrative perception of morality, which is centered on the storyteller's need for a conflict. If you actually count morality like a statistician, you wake up in the morning, say where I'm going to log all my moral decisions.

[00:41:56] I promised that I'd be here at 2:30, tick I'm here at 2:30. I was, good luck. Any drama there? I was a bit worried about getting a park. It was only one park left because I don't have a sticker anymore. I forgot about that.

[00:42:10] When I actually would actually go through my day of all the moral things that I do, which if I did the opposite would be wrong. And I don't even seriously think it's only because I did this PhD in robot ethics. I thought what if I did the opposite? Oh yeah, that'd be bad. I could park in the wrong place.

[00:42:23] I could park, in the middle of the road, there's all these possibilities I could do, but just don't enter my mind. So the morally obvious is a powerful place to start with robot ethics because the robots have to learn the morally obvious before they get to the morally contentious.

[00:42:38] **Christoph:** I guess that's good news for your roboticists out there. Ethics is not difficult to do.

[00:42:44] **Sean:** Easiest pie.

[00:42:46] **Christoph:** Thank you so much, Sean.

[00:42:47] **Sean:** No worries. Thanks Christoph.