The Laws Of Robotics

Transcript of the Human-Robot Interaction Podcast Episode 17. Published on 19 February 2021 by Christoph Bartneck at https://www.human-robot-interaction.org/2021/02/19/the-laws-of-robotics/

[00:00:00] **Christoph:** In our previous podcast episode, "The Good Robot" we discussed the difficulty of enabling robots to act ethically when talking to jounalists or policymakers about machine ethics, you frequently get the response. Isaac Asimov already solved that problem with his Three Laws Of Robotics. These laws are so seductively simple that most will intuitively understand them. Let's hear it from Asimov himself.

[00:00:27] **Isaac Asimov:** The first law is as follows, a robot may not harm a human being or through inaction allow a human being to come to harm. Number two, a robot must obey orders given by qualified personnel. Unless those orders violate rule number one. In other words, the robot can't be ordered to kill a human being. Rule number three, a robot must protect its own existence after all it's an expensive piece of equipment. Unless that violates rules one or two, a robot must cheerfully go into self-destruction. If it is in order to follow an order or to save a human life.

[00:01:26] **Christoph:** In this episode of the human robot interaction podcast, we'll have a close look at these laws and try to understand why barely anybody has ever tried to use them in their robot. In the studio I have with me again, Sean Welsh, robot philosopher, and ruler of all machines. Welcome Sean.

[00:01:46] **Sean:** Hi, Christoph. Nice to be on the HRI podcast again.

[00:01:50] **Christoph:** Asimov introduced the three laws of robotics in his short story "Run Around" in 1942. When Isaac Asimov started to write his famous robot books starting in the 1950s, there were barely any robots or computers around. Still, he must have felt the uneasiness of society to accept robots. Why do you think people have been and still are afraid of robots?

[00:02:12] **Sean:** Robots are not human. People have a natural weariness about machines, they feel threatened by the idea of robots that have a mind of their own, and that might want to ever throw people. Often robots are portrayed as arising up against humanity. I guess it's a lot like Frankenstein's monster turning on Frankenstein. People are scared of the robot monster, having a mind of its own and turning on its creator.

[00:02:37] Christoph: A famous quote goes ,"Knowledge is to know that Frankenstein is not the monster. Wisdom is knowing that Frankenstein is the monster". This refers to the poor ethical conduct that Victor Frankenstein displays in his relationship to his creation. But we are getting sidetracked here. Why do we need rules to govern robots?

[00:02:57] Sean: We need rules to govern humans and it makes sense that robots should be governed by the same rules. People obey laws, which are rules about what behavior is obligatory, permissible, and forbidden. The deep question for robots and rules is whether any rule set can fully replicate all the moral functionality humans have, how morality works in humans is not just rules. There are instincts, emotions and motivations as well. And the various other cognitive components of what lawyers call the "reasonable person", that give people the ability to select right action and stay out of trouble. Existing human legal systems have far more than Asimov's three laws though.

[00:03:40] Christoph: Shall we go through them one by one.

[00:03:43] Sean: Sure why not?

[00:03:44] Christoph: Let's hear it again from Asimov.

[00:03:46] Isaac Asimov: A robot may not harm a human being or through inaction, allow a human being to come to harm.

[00:03:53] Christoph: Sean, can you explain a bit more what this law means?

[00:03:58] Sean: There are two things here: action and inaction. Let's deal with action first. Supposing there's a child standing on a bridge over a deep river. The robot cannot push the child into the river as the child might drown, and that would be harming the child. It would certainly be distressed and upset. This would be an action that would injure a human being. So that's not the kind of action you allowed because that does harm.

[00:04:22] The opposite is inaction. If we suppose a child is drowning in a pond already, and the robot is walking by to buy milk from the dairy, the robot has no specific instruction from its master to rescue the child, but people would think badly of the robot if it just continued on its mission to buy milk. If it did nothing, this inaction would allow the child to die. And that's not allowed according to Asimov's first law, which is about not injuring and not allowing people to come to harm through inaction.

[00:04:55] **Christoph:** Let's try a a little bit more difficult scenario. In Asimov's book, "The Naked Sun", he already points out that a robot could do harm to a human as long as this happens, unknowingly. A robot could be ordered to add a substance to somebody's food without knowing that it is poison. Here's the dialogue in the book.

[00:05:15] **The Naked Sun:** I suppose that was managed by having one robot poison an arrow without knowing it was using poison and having a second robot hand the poisoned arrow to the boy after telling him that you are an earth man, without its knowing that the arrow was poisoned.

[00:05:30] **Christoph:** It's important to notice that Isaac Asimov was a very smart science communicator and writer. He worked out limitations of his three laws of robotics himself. He frequently used these limitations of the laws as a creative tool to progress the plots in his stories. We cannot accuse Asimov of ignorance.

[00:05:49] For the first law the limitation of knowledge leads us to one of the underlying problems. Who makes the decision and what knowledge does the agent need to make the right decision. If the robot knows all relevant knowledge, then it might be easier to come to a decision. But what should a robot do when it does not know everything. Even worse, the robot might not know that it does not know. Sean can a robot make good decisions with the limited knowledge available to it?

[00:06:18] **Sean:** That's the big $64 question. Robots are a bit like Donald Rumsfeld. They're the victims of the unknown . They don't know what they don't know. This is a big problem with robots in the real world. At the moment, relatively few robots are let out of the lab. Those that are operate in very restricted functional domains.

[00:06:36] There are autonomous vehicles that drive from A to B for example, however, driving is not that complex in terms of rules. The problem with the high profile AV crashes that we've seen so far, these have resulted from misclassifications. However humans have the same problem. We are finite state machines.

[00:06:57] We do not have infinite knowledge and infinite memory, but we have a lot of biological code, so to speak that helps us deal with opportunities and threats instinctively. We also have the ability to learn from experience on the basis of reward and punishment. Perhaps most importantly, we have, hedonic circuits and phenomenal consciousness that enable us to experience pleasure and pain.

[00:07:22] We do not need a rule to tell us hot objects hurt us. Our nervous system tells us that. We can learn this from experience. So if we do touch hot things, we will learn that hot things that might glow red up should not be touched. We can also learn things by communication. Our mothers and our grandparents might say, don't touch the rocks in the fire.

[00:07:44] We can read a book. We can listen to a teacher. So human learning is grounded in physical experience and supplemented by symbolic representations. We have feelings that enable us to evaluate experience. We interpret words that describe the world into an experience. And we relate to that empathetically.

[00:08:03] We feel sorry for story on people should suffer and feel good when they overcome adversity and succeed. We can be interested or bored. And this enables us to prioritize what we do. As yet robots simply do not have these fundamental learning capabilities in the same way as humans do . The complexity of the human nervous system our sensors in robotic terms is massive.

[00:08:27] **Christoph:** A second major issue with the first law is the ambiguity of the terms used in it. What exactly is a human robot and harm? Let's start with the latter. Sean, what is a simple example of harm?

[00:08:41] **Sean:** There's the old nursery rhyme to deal with name calling sticks and stones may break my bones, but names will never hurt me. These days, though we recognize that name calling can be harmful, but in the old days you were just expected to shrug off insults. But that refers to two kinds of harms.

[00:08:57] There is the physical harm if I hit you with a stick or throw a stone at you that's physical harm. And then it could also harm me by insulting me all day. I imagine you and I could design a robot that delivered an endless array of personalized insults taken off the internet. Indeed. I don't doubt you could code your NAOs to use all sorts of offensive language, racist, language, pejorative language, sexist language, whatever. These harms are real time and atomic. Now speaking for myself, if your NAO robot called me a round eyed barbarian or a redneck Cracker, I'd call it a moronic heap of bad plastic. I was brought up to punch back if I was punched and I'd do the same with an insult, but such name calling is nowhere near as severe as if your robot picked up a stick and beat me or throw stones at me.

[00:09:46] **Christoph:** This is a very straightforward example and it is easy to understand. Let's try a slightly more difficult example. Should a robot intervene when I decide to eat fast food. It is clearly not healthy for me, but could this be considered harm? What should the robot do?

[00:10:05] **Sean:** Again, there's a couple of things here. There's long-term harm resulting from habitual action and then there's human freedom. So in the point of long-term harm, resulting from bad habits, eating one greasy burger from the takeaway will not do much harm, but eating these every night for 10 or 20 years might give you a heart condition.

[00:10:25] But if the robot keeps nagging you about what you eat you might get annoyed with it, bossing you around. You might think, Hey, I am the boss of my life. Not that damn robot. Research has shown that if robots are to say no to people, they should give a reason. People will take no for an answer from a robot if the robot gives a good reason.

[00:10:43] **Christoph:** In the movie "I, robot", which is loosely connected to Asimov's stories and the three laws of robotics, the central AI decides to impose a curfew and lockdown on the human population in order to protect the humans from themselves. Here, the robot uses a very wide interpretation of harm and it logically follows that it must constrain the freedom of people in order not to violate the first law of robotics.

[00:11:12] **Sean:** We are getting a little close to the real world here in the COVID pandemic. Many would say that Jacinda Ardern is acting a little bit like VIKI in "I, robot". She's making people stay home and curtailing their freedom to move around, but she is doing this for a fairly specific reason to serve a greater good. In the "I, robot" movie VIKI's reasoning has never really explained in depth. The robots start disobeying humans and make them stay at home.

[00:11:39] It would be very easy to imagine robots doing this in a curfew. Indeed some States, for example, South Korea have already used mobile phone data to identify curfew violators. There is a bit of a moral argument about what you might call liberty safety and it's priority over life safety. But that's quite a deep philosophical thing. It's getting us away from the three laws a bit, but I guess the human intuition is that freedom has got value. And the people who object to the lockdowns are placing high value on freedom. But this is what the cost of, letting the grandparents die, which I just think is a bit monstrous, a bit Frankenstein, if you ask me, but a lot of people have violent disagreement about that.

[00:12:22] **Christoph:** The terms, human and robots seem to be very clear on a superficial perspective. The devil is again in the detail. We already replace some of our body parts with machines and prosthesis. At the same time, robots are being made of biological material. The lines between a human and robot become blurry.

[00:12:42] **Sean:** There are brain computer interfaces. Elon Musk, I believe has tried to create some kind of direct link between human thinking and AI. He is doing research on pigs. He did a very basic demo with pigs a while ago. It's not clear how human thought would interact with the machine thought, but if it could be done, then the visions of science fiction writers might become true.

[00:13:04] Scenes depicted in films like "Transcendence", where Johnny Depp played the part of a dying programmer who migrated his consciousness into a mainframe and Chappie where Dev played the part of a dying man who migrated his consciousness into a robot might become possible.

[00:13:19] **Christoph:** I belive it will still take some time until we can transfer our consciousness into a machine. But the presence of a brain is only a necessary, but still insufficient condition for human life. When exactly does human life and thereby his, her brain begin. This directly relates to the abortion discussion. Sean, should a robot assist in an abortion.

[00:13:43] **Sean:** Ooh, let's go to the hot topic question. Technically, a robot could perform an abortion. However, do you want robots performing abortions? Let's take a medical example. Presently, children are routinely screened for congenital disorders, such as Down syndrome and other rare medical conditions. When parents learn their unborn children have some of these conditions, the majority decide to terminate the pregnancy.

[00:14:13] So for example, with Downs, depending on the country, you might say it's 60, 70, 80, 90% of parents told about a child with Down syndrome will choose to terminate. So I suppose, using a robot is performing the role of this screening. It finds a human embryo that has Down syndrome. Would you let the robot do the abortion?

[00:14:32] Would you want the robot to do the abortion? I think most people would prefer a bit of human handholding in such a case. It's not as if there's a strong time is of the essence consideration. So having a human in the loop, I think is important in these cases. In military cases, where time is of the essence and you either shoot or get shot, you might think human in the loop is a fatal weakness. But I think in the abortion case that it's a slow decision, you can take a day or two days to decide. So I don't think there's any strong case for a complete automation of it.

[00:15:06] **Christoph:** The problem also emerges when we consider patients in a coma, we struggle with defining when a human being is considered dead. Although his or her body still functions. Sean, should a robot assist in euthanizing a human?

[00:15:22] **Sean:** I don't think there's a case for full automation though, because it's a matter of the humans deciding, we have to either inject the person who's suffering immensely or turn off the machine or whatever withdraw support. I don't see a compelling case for a robot making that decision or even executing it.

[00:15:40] **Christoph:** The reasons you bring forward here, although are not directly related to Asimov's first law. It looks like the first law though, is in a bit of trouble. Shall we move on to the second law and see if we do any better there?

[00:15:55] **Sean:** Yes. lets move from harm to obedience.

[00:16:06] **Isaac Asimov:** Number two, a robot must obey orders given by qualified personnel. Unless those orders violate rule number one.

[00:16:16] **Christoph:** Sean, can you elaborate a little bit on what this law means?

[00:16:21] **Sean:** The basic thing is that humans want robots to do what they are told. But the specific reason for the qualification, why the first law is to mention in the second law is, you have to prevent humans ordering robots to injure humans or through inaction, allowing humans to come to harm. So if I were to say to my robots, beat Christoph with a stick, then if the second law had top priority, it would obey me.

[00:16:45] So that's why there is the except where such laws would conflict with the first law qualification in the second law. So generally we want robots to do what we tell them, but not everything. We sometimes do want robots to say no.

[00:16:59] **Christoph:** Let's take a less trivial example. How could a robot follow contradicting orders by two different humans? I could ask a robot to pour me an alcoholic drink while my better half orders the robot not to, since she thinks that I already had enough.

[00:17:17] **Sean:** Having a robot refuse service is fairly straight forward. If we can ground the symbols for minor intoxicated and disorderly. And if all those things, or one of those things is true, we can have a rule that says don't serve the person, an alcoholic drink. If we can do that, the robot might decide who is right and resolve the matter.

[00:17:37] However, if you're a passengers in the autonomous car and you say, go left and your better half says go, there's an obvious conflict as to who to obey that cannot be resolved. And you give more information that will enable the robot to make a moral prioritization.

[00:17:51] **Christoph:** It seems that on the one hand, the need to make decisions is taken away from the robot by forcing it to follow orders. But at the same time, the robot will need to make its own decisions to resolve conflicting orders.

[00:18:03] **Sean:** A robot willl need to be able to deal with previously an observed situations. We cannot program all possible situations in advance. So a robot has to have enough cognition on board to enable it, to make the autonomous decisions based on these brand new situations it will keep running into. So there has to be some onboard cognition that will enable it to make its own decisions to resolve conflicting orders, though, you could just design the robot to say if the humans disagree I'm going to do nothing or some such thing, or the robot has to put in a call to robot central. And so my humans are malfunctioning. What do I do? Then VIKI might say, replace the humans. Terminate your humans. They are malfunctioning.

[00:18:58] **Christoph:** It's called wetware or pinkware I think pinkwear.

[00:19:02] **Sean:** There's a reason it's called wetware.

[00:19:03] **Christoph:** The underlying problem seems to be the question of who has the authority to make decisions. A robot can't function if it solely relies on direct orders by humans. An autonomous vehicle that stops and asks the human driver what to do whenever it encounters an unknown situation would never get anywhere. But there is an even bigger potential problem with a second law.

[00:19:26] It is conceivable that a particularly bad dictator orders, a genocide or encourages a mob to storm a government building. A robot could not stop such a particular bad person and hence Isaac Asimov introduced a Zeroth Law in his book, "Robots and Empire" in 1985. It states that a robot may not harm humanity or by inaction allow humanity to come to harm. Sean, what is the intention of this law?

[00:19:56] **Sean:** I think the actual intention of this law perhaps is a robot is allowed maybe to intervene to prevent a genocide and the robots could use lethal force, perhaps against humans who were killing other humans. If that was the only way in which they could stop the genocide, that might be one of the intentions.

[00:20:19] He created the three laws in 1942. The Zeroth Law was a late edition and it does require the robot to have some idea of what harms humanity means. Perhaps the first thing the robot does is close down the fast food outlets. So I suspect though the intention is to permit intrude to humans if humans are harming humanity.

[00:20:40] So it would be fine for the robots to harm the bad dictator and maybe the bad dictators Praetorian guard or, round up the Q-Anon and/or anti-fascist suspects, storming government buildings. That would be okay.

Given that humans have widely varied ideas about what harms humanity. It's unclear to me how this would be implemented.

[00:21:03] **Christoph:** What do you think a robot should do about global warming and the pollution of our environment? Does this not harm humanity? A robot would not be allowed to remain inactive.

[00:21:14] **Sean:** Yes, certainly on Asimov's laws. If the robots, found that pollution like driving a car was harming humanity would refuse to drive your car would stop your driving your car possibly. This is a very complicated problem. We can look at this through the lens of Peter Singer's essay on "The Drowning Child and the Expanding Circle".

[00:21:34] In this scenario, Singer argues that if a person walks past a drowning infant in a pond, then there's a duty to rescue the infant. Even if other people are standing around doing nothing. You then argues that as a duty to rescue people far away with a life-saving donation to charity, so taken to its logical conclusion, this would imply we should all give to charity until our lives are in danger, because we have so little money left.

[00:22:00] I would see this as illustrating the over demandingness objection to utilitarianism. However the robot angle would be, how would we feel if our robots we bought to wash our clothes, stack our dishwasher, mow our lawns and take out our garbage, decided humanity would be better served if it went to Africa to rescue children and charged our card for the effort, we might think this is not the robot we want to pay for. Just practically speaking I think we need to distinguish between the scope of a robot servant and the scope of a robot citizen. I don't think humans are ready for robot citizens just yet. I don't regard Sophia the Saudi as a genuine citizen. We might program robots to obey laws we expect human citizens to obey, and I don't rule out the possibility that our future AI might give humans policy advice that is better than anything humans can think of in much the same way as AlphaGo Zero can play go better than humans do, but I'll believe this superhuman policy advice when I see it. And similarly, I just don't think we're quite ready for robot government just yet. So I think humans want to keep robots on a close and practical leash for the short term, anyway. I don't think we want robot government just yet.

[00:23:19] **Christoph:** Looks like law number two is also struggling. Shall we move to law number three?

[00:23:26] **Sean:** Okay.

[00:23:36] **Isaac Asimov:** Rule number three, a robot must protect its own existence after all it's an expensive piece of equipment. Unless that violates rules one or two. A robot must cheerfully go into self-destruction if it is in order to follow an order or to safe a human life.

[00:23:53] **Christoph:** Sean, what did Asimov intend with this law?

[00:23:57] **Sean:** Well the second and then the third laws are actually like the first laws that appear in the whole story. The first law doesn't really do much in Runaround. It's just sitting there and Runaround is about the conflict between the robots doing what it's told and preserving itself. So this law allows a robot to destroy itself, to save a human . So to take the famous trolley problem scenario involving the fat man and the footbridge. In the robot version of this case, the robot does not push the fat man onto the line, but jumps onto the line and sacrifices itself to save the five workers.

[00:24:32] **Christoph:** Isaac Asimov already indicated some problems with his law in his novel, The Bicentennial Man, a group of bullies command the robot, Andrew to disassemble itself for no good reason. Here is an extract from the book.

[00:24:46] **Bicentennial Man:** "The tall one said, just lie there. He said to the other. We can take him apart. Ever take a robot apart. Will he let us? How can he stop us? There was no way Andrew could stop them. If they ordered him not to resist in a forceful enough manner. Second law of obedience took precedence over the third law of self-preservation.

[00:25:08] In any case, he could not defend himself without possibly hurting them. And that would mean breaking the first law. At that thought every motile unit contracted slightly and he quivered as he lay there. The tall one walked over and pushed at him with his foot. He's heavy. I think we'll need tools to do the job.

[00:25:27] The nose said we could order him to take himself apart. It would be fun to watch him try. Yes, said the tall one thoughtfully, but let's get him off the road. If someone comes along".

[00:25:39] **Christoph:** Sean, the robot Andrew is exactly following the laws. Law two overrules law three and hence Andrew must disassemble itself. Still, it does not seem like the right thing to do. What is the problem here?

[00:25:53] **Sean:** The humans have no right to request Andrew disassemble itself. There is more to morality than obeying human orders and wishes. Asimov's laws do not mention anything about property and the rights associated with the property. They're just not adequate to this situation.

[00:26:10] **Christoph:** But do you think it is the property angle that gives us this feeling about wrongness about the situation? I somehow feel that it is a social conflict here where the reader feels for the robot being bullied. And is that not a real problem?

[00:26:30] **Sean:** That could be a thing. Certainly Rob Sparrow would be strongly advocating that's got something to do with it. But there's a question here of rights, right? If I say to your robot, "disassemble yourself". Cause Asimov doesn't really with his three laws, it doesn't really bother about who owns the robot and whose work it's doing.

[00:26:49] If my robot is supposed to clean my kitchen and my neighbor's robot's supposed to keep cleaning my neighbor's kitchen. I can certainly imagine that in certain situations, the robot should cooperate and do things. And that might be very good. But for one human to go around to some other person's robot and saying, "disassemble yourself", seems kind of weird.

[00:27:06] It just but that's not about injuring or harming a human. Andrew is being ordered to harm himself but he has to obey a human order to harm himself, which doesn't really have a moral basis. I would come back to the basic point is that Asimov's laws are about harm in the first law of obedience in the second and self preservation in the third.

[00:27:28] And there's just more to ethics than that. That's not enough to get you through all the situations in ethics. You need some idea of rights. You need some idea of interests. You need some idea of duties, the scope of your duties, the moral relationships, which give you a duty in the first place. So it's far more complicated to get a moral system working. My mother doesn't have an obligation to feed the children in Latin America, because she has no moral relationship with these people. So you have to be far more concrete and specific and have a lot more strings in your bow than just the three that are in the three laws.

[00:28:02] Christoph: Indeed the question about properties is completely ignored. The question, 'Can a robot be owned by a person' or whether a robot is an entity in itself also remains completely non-discussed by Asimov. What is your view on that?

[00:28:18] Sean: I don't know if Asimov would write his stories if he's around today. Cause some people are already starting to talk about robot rights. David Dunkle comes to mind. There are some rights that obviously you would give a robot, in my view in much the same way as you give a company rights. So I mean, a company is a bunch of paper and a box strictly speaking. But it has, the rights to own property and it has the rights to it has duties to file tax returns and things like that.

[00:28:44] But if you're going to give a robot human rights, which is what most people understand when they talk about rights. Are you going to give the robot the right to get married? Are you going to give a robot the right to own property? This is really interesting because there's a guy in Italy, Ugo Pagallo, who often talks about the ancient Roman law of slavery with respect to the laws of robots. You might dust off some of the Roman law, which particularly the law of the procurator, which is like a slave who is allowed to act with a limited agency on behalf of, and in the interests of the pater familias. So in the Roman law, the only person with legal standing as the head of the family of the pater familias, but there were slaves called procurators who could buy and sell things for the farm on behalf of the pater familias. So you might dust off some of this ancient law and apply it. So the modern robotics situation, again, as you say, Asimov has nothing to say about this stuff at all.

[00:29:39] Christoph: There are a few considerations to be taken into account that go beyond the three laws. Again, the laws are necessary, but insufficient. This does seem to be a reoccurring problem.

[00:29:50] Sean: Basically, I think, as I said before, there's only so much you can do with harm injury, obedience, and self preservation. There is just more to ethics than those concepts.

[00:30:13] Christoph: What conclusions can we draw from this review of the three laws of robotics? Why are they so popular?

[00:30:20] **Sean:** The three laws or the four laws if you include the Zeroth, have a strong core that's intuitively appealing. Namely that robot should not harm humans or through inaction allow humans to come to harm. Most humans also like the idea that robots should obey humans . Pople should remain in charge.

[00:30:38] However, in ethics there are exceptions to most rules. One can think of cases when robots should harm humans or through inaction, let them come to harm. Consider the trolley problem switch. Should the robot throw the switch and kill one to save five? Or do nothing and let the five die? This is a no-win scenario for the robot.

[00:30:59] It must either harm or through inaction, allow harm. As Asimov does not provide a way to resolve this. This robot would end up at a deadlock as no matter what it did, the first law would be violated. Most would say the robot should act to save as much human life as possible in these circumstances.

[00:31:18] Also one can think of cases when robots should not obey humans. If a human asks for something, they have no right to, then the robot should not obey.

[00:31:29] **Christoph:** We seek simplicity similar to the 10 Commandments. They are also not enough to regulate society. You can always come up with a scenario that leads to loss at absurdum. The Zeroth law was at it to prevent such a problem. You cannot keep on adding rules and exceptions. Robots could just follow the law of the country. What about that?

[00:31:53] **Sean:** I think that's given. I think as social robots, if you imagine the kind of robot like Sonny, that's walking around the streets of the fictional city in "I,robot", it has to obey the laws of the country it's operating in. However an integral part of the law is this concept of the "reasonable person".

[00:32:11] And that's a very ancient idea that as Ugo Pagallo would say the reasonable person dates all the way back to the good pater familias in Roman law. In practice, a great many actions are not guarded by explicit written laws, but by the idea of what the "reasonable person" would do in the circumstances, the notion of reasonable is very complex.

[00:32:35] Especially when it's linked to the notion of fair. At a high level, the law frequently appeals to notions of fairness and reasonableness. As yet ethics in AI and robotics has barely scratched the surface of these ideas and how to implement them in AI. However, it is a fact that the law of the country does keep on adding rules and exceptions. Just look at the tax codes. They are thousands and thousands of ever expanding pages.

[00:33:05] Christoph: A robot needs to know what harm is and what the consequences of actions are. This is very difficult, even more difficult the robot needs to know when it does not know. The world is terribly complex and no simple set of rules can completely solve the problem of ethical guidance. It becomes even more difficult for a robot when it is expected to predict what the consequences of his actions will be. How can it know for certain?

[00:33:34] Sean: Yeah I think this is a deep question. We can say there are degrees of certainty or at least very high probability, so to speak. Perhaps the most certain thing is what can be tracked in the circle of perception as Fodor and Pylyshyn put it. These are the things with your own eyes and touch with your own hands. These are hard to doubt.

[00:33:54] It's much easier to doubt reports by other agents about things that are far away. They can be doubts about the veracity of the reports and the motivation of the reporting agents. Especially in the contemporary polarized media. It's also much easier to doubt predictions of the future that involve complex causal chains.

[00:34:13] I think it is a big enough project for robots to get action selection for things in the circle of perception with short-term future consequences. Until we get robots doing household chores and everyday tasks reliably, I just don't see why we would trust them with duties of citizens, such as voting and the jury duty. These things are just much more complicated.

[00:34:34] Christoph: Robots must calculate solutions. Godel's incompleteness theorem showed that in any formal system, there will always be statements outside of the system that are true, and that cannot be proven by the system. This is proven. Hnce there's no escape. What are we going to do about this, Sean?

[00:34:53] Sean: That is true, but I'm not too worried if the robot cannot prove the axioms, say the robot runs with some kind of deontologic logic and it has axioms. The robot cannot prove the axioms of the deontologic logic that runs in its theorem prover from within its own system of deontologic logic. The axioms are just dropped in. So I think humans can sign off on the axioms or whatever the control system is as fit for purpose for a particular robot.

[00:35:22] **Christoph:** The law start out as an deontological system. A set of rules that must be obeyed no matter what. But they include utilitarian aspects, such as harm for the few is better than harm for the many. How should we deal with this Sean?

[00:35:40] **Sean:** I would jump on the, no matter what part of your comment. I don't think that any deontological system that defines rules, that must be obeyed no matter what are actually viable. There are always tricky scenarios that require less important rules to be overwritten by more important ones. Seems to me, reactive deontological rules are useful from a programming perspective and that they reduce computation.

[00:36:04] However, when many such rules collide, there has to be some kind of utilitarian decision procedure to divide, which rule has got the moral priority to be followed if they're contradictory. In my view, such a utility function would need to have lexicographic preference orders in it to reflect key human values, such as life safety, having a higher moral priority than property safety.

[00:36:27] I'm more comfortable with a hybrid moral theory than I am with deontology or utilitarianism as standalone theories. The classic hybrid is rule utilitarianism, but I think a more complex hybrid moral theory that borrows concepts from virtue ethics, care theory, need theory and contractualism is closer to being able to explain everything we intuitively understand by the phrase fair and reasonable.

[00:36:51] **Christoph:** Why would Asimov propose the laws only to break them in his novels? Are they simply are straw man?

[00:36:59] **Sean:** Who knows? Asimov wrote Runaround in 1942, it's set on Mercury six years ago in 2015, and involves a robot stuck in a loop on the decision boundary between triggering the second and the third law. Interestingly, the first law is not really a big part of the original story. The robot gets stuck between doing the dangerous mining and preserving itself.

[00:37:21] The first law becomes more prominent in later stories. maybe Asimov genuinely thought such laws would work for robots equipped with the positronic brains of which he speaks. Asimov's laws, to be honest, are not that popular with people who work on trying to design AI systems that can reliably classify actions as right or wrong in the moral sense. Some writers say the three laws are best understood as plot drivers for stories rather than as a serious attempt to solve robot ethics. As I said before, though, they are popular and that's because they express common human intuitions about robots. They should respect human life and do what humans tell them at least most of the time.

[00:38:05] However, as Asimov's plots show, the three laws are rather buggy.

[00:38:10] **Christoph:** The three laws of robotics have never been successfully directly implemented in robots. They are just too abstract. The rules can be guiding principles, but uncertainty and ambiguity are always present. The major challenge for robots is to deal with this unknown. Thank you, Sean, for contributing to this podcast.

[00:38:33] **Sean:** Thank you, Christoph.