Panel 2:
Accountability for the Actions of Robots

Moderator: Reux Stearns

Panelists: Ryan Calo, Howard Jay Chizeck, Elizabeth Joh, and Blake Hannaford

Calo: So, my question for you has to do with, what is the sort of range of human control over robots, right? It's not all like the robot, you wind it up and it goes. Like what are the different kinds of configurations for human control of robots?

Chizeck: This is also a taxonomy question. The first class is the industrial robot that you see in a factory. There is no human controlling it except to push its buttons, and then the original programming operates; there's yellow lines on the floor that you're supposed to stay behind for safety. And as we heard earlier, there's established law for how to deal with injuries for this kind of robot. The next kind of robot is what we would call a telerobot, or a *human-operated robot*.

Think of it as a puppet. Your automobile, an older automobile, is that sort of device. You turn a steering wheel and you push on some pedals, and mechanical things happen in response. An airplane that you, the pilot, are flying by yourself is that kind of system. Telerobots became an accepted term when people started to deal with manipulating objects that were too hot or cold or too radioactive or too biohazardous or too far away for a human being to directly handle.

Thus, we have telerobots for mining and for explosive-handling, and we have telerobots for surgery into small spaces, and telerobots for operations in low earth orbit. The key idea of the telerobot is that there is a human operator. Most of the military drones that are flown are telerobots. There's a human operator in Texas or Oklahoma flying these things in Central Asia.

Now, some telerobots have a degree of autonomy, a shared autonomy with the human operator. That is, some of the tasks are done by the device and without specific direction by the human operator. You see this in the

newer automobiles with anti-lock braking, which takes care of the braking for you when you start to slip. Next, you have vehicles that can stop for you in time to avoid a front-end collision. If you buy a high-end car now, it can prevent you from drifting out of your lane or changing lanes if there is a vehicle in your blind spot. Thus, the human driver now is not completely in control. There are sensors and automatic systems that drive some driving decisions. This, of course, changes responsibility, and there are legal implications, I presume.

You don't see autonomy in telerobots for surgery yet, but soon you might be able to say, "suture from here to here," and then the device might do that. Shared autonomy exists in commercial aircraft. The pilot can engage the autopilot and the plane flies itself. These systems can fail, especially if there are sensor errors or human–computer interaction confusions.[1] A challenge in these systems is: how do you switch back and forth between human control and robotic control?

The robot knows what the human is doing, always, if there are good sensors. The human may not know what the robot was doing or intending to do, especially if it's an emergency hand-off.

The next type of robots are devices that are a part of people, such as implanted devices. A new hot topic is brain–computer interfaces. *The Economist*, on January 4th, interviewed most everyone working on this in a survey article of the field.[2] These devices that affect your physiology, they respond for you, and they are a part of you.

This might be a pacemaker or defibrillator that's implanted and you're not in control of it, or it could be a prosthetic device, where the human is sending commands to grasp the object. But the human is a part of the overall system, a part of it physically.

A final class of robots is the fully autonomous robot. This would be your self-driving car. This would be something where you may be telling it in the beginning where to go, but everything afterwards is done by the device. So, now is it the manufacturer or the maker of the software, or the owner or the driver, who is responsible if something goes wrong?

---

1. William Langewiesche, *The Human Factor*, VANITY FAIR (Oct. 2014), http://www.vanityfair.com/business/2014/10/air-france-flight-447-crash [https://perma.cc/7YBK-DY6A]; Gerald Traufetter, *Death in the Atlantic: The Last Four Minutes of Air France Flight 447*, SPIEGEL ONLINE (Feb. 25, 2010), http://www.spiegel.de/international/world/death-in-the-atlantic-the-last-four-minutes-of-air-france-flight-447-a-679980.html [https://perma.cc/2L9B-7QWD].

2. *See Eberhard Fetz*, ECONOMIST, Jan. 6, 2018, at 10; *Grey Matter, Red Tape: Looking for Serendipity*, ECONOMIST, Jan. 6, 2018, at 11; *Headache: Non-Invasive Devices*, ECONOMIST, Jan. 6, 2018, at 5; *Inside Intelligence: Implants*, ECONOMIST, Jan. 6, 2018, at 7; *Thought Experiments: Brains and Machines*, ECONOMIST, Jan. 6, 2018, at 3; *Translation Required: Data Processing*, ECONOMIST, Jan. 6, 2018, at 9.

These become interesting questions. A different taxonomy is to consider there are robots that interact with people in a communicative way and robots that don't. A tremendous amount of research is now being done on facial expression and posture and hand motion and voice recognition as ways that robots and people can interact. Now, I don't think of Alexa as a robot. I think of it as an operating system.

But, as soon as you tie Alexa to the door and the lock and the heating system and the lights, then it's kind of a larger robot. And here we get to the ideas that you heard in the earlier panel of artificial intelligence. That's where AI touches into the physical. You know, there is a joke that goes, "What's the difference between AI and machine learning?" And AI is when you're trying to raise money, and machine learning's when they try to hire people. The truth lies somewhere in-between.

Deep inside the system, there are algorithms to do things, but these algorithms are learning from data, which might be biased data. And when this intersects the physical world, then we get intelligent hardware agents. There is also a kind of *collaborative* robot under development, where humans and robots work together on tasks. A local startup in Fremont, BluHaptics, is working on this for space applications.

Imagine robots that are working with astronauts, outside or inside a space vehicle, carrying and fetching or holding tools. They have to communicate with humans. That requires a certain amount of social interaction. There's no yellow line to avoid human injury now. The robot has to know enough to do the right thing.

All of these different types of robots, I think, are relevant for different areas of law, which I know nothing about. There are people who program the robots, there are people who manufacture them, there are people who maintain them, there are people who own them, there are people who choose to operate them, and issues of liability and safety and responsibility are intertwined throughout.

Calo: So, I have a follow-up question, and I also wanted to give, you know, Blake, is there anything you wanted to quibble or add to that? It's actually nice to have two engineers. You get two perspectives on that. Anything you wanted to add to that taxonomy that Howard put down?

Hannaford: Right, so we could get into an argument and use up all the time.

Calo: You could! Yeah.

Hannaford: Actually, Howard and I, our offices are right next door, so we've had about ten years to get all those arguments done. So, I actually completely agree with Howard's classification.

Calo: Good. Excellent. There you go. So, there you have a consensus, for once. My follow-up questions is that, you know, lawyers

being so, what's the non-pejorative way to say this by ourselves? We like definitions, you know? And we like to sort of think about definitions, and often, there's almost a fetishism of definition, you know? And so, I've been involved in countless conversations around how to define artificial intelligence, which is sort of what we're calling, you know, what used to be called statistics, and then became analytics and big data that we now think of as machine learning or algorithms. Do you have a favorite definition of artificial intelligence that you want to hazard?

Chizeck: I don't know if it's a definition, but I'll give you an example.

Calo: Okay.

Chizeck: In July 2017, Facebook decided to have two chat bots, programs, negotiate as they traded objects. And Facebook neglected to constrain them to negotiate in comprehensible English. And after a few hours, the programs developed their own language, and it was a sing-song language with repetitions, and they were chanting it with each other. But they were coming up with the trades. When the Facebook engineers discovered that they couldn't understand it, they turned it off.[3]

This was an emergent property. Unexpected. They told the chatbots to optimize for the negotiation, to optimize the speed of negotiation. There's transcripts on the web, but it was like a lot of repetition and key words were taken out of English and lot of objects and verbs and things were thrown out. They didn't need those. So, to me, artificial intelligence becomes interesting when you get emergent properties that were not programmed, that were not scripted.

Alexa doesn't quite do this yet. Everything in Alexa is scripted, but the chat bots that some of the Alexa competitions developed, like the winning UW one,[4] go to Reddit and find topics and talk with you about them. Still un-emergent, but when you start to get new strategies because of unknown optimization, I think then you start to cross the line.

The singularity fear is about that. I think it was two or three years ago, we had three singularity movies come out at once. One was *Her*,[5] which was an operating system very much like, but before, Alexa.

---

3. Andrew Griffin, *Facebook's Artificial Intelligence Robots Shut Down After They Start Talking to Each Other in Their Own Language*, INDEPENDENT (July 31, 2017, 4:10 PM), https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

4. Jennifer Langston, *UW Students Win Amazon's Inaugural Alexa Prize for Most Engaging Socialbot*, UW NEWS (Nov. 28, 2017), https://www.washington.edu/news/2017/11/28/uw-students-win-amazons-inaugural-alexa-prize-for-most-engaging-and-conversant-socialbot/.

5. HER (Warner Bros. 2013).

One was *Ex Machina*,[6] where the evil robot is going to replace us. And one was *Transcendence*,[7] where a well-meaning human merged with a computer to help save the environment and the world but was misunderstood and killed. Sorry to ruin the movies. But all three of those were singularity stories in the same year, and their underlying question was: "Do you need the human part to get these emergent properties? Or can the code itself do it?" And I don't know how you define that in legal terms. But that's the nightmare or the promise that is coming.

Calo: What's interesting is—they overlap because that is also where I personally believe the most interest legal questions are, but let me turn it over to—thank you very much for that taxonomy. So, Elizabeth, one of the things I know you study most closely is a particular use case for robotics that we're seeing more and more of, which is the use by police. And I wondered if you could talk through a little bit, how are the police thinking about using robots and what sort of challenges of law and policy are they beginning to pose?

Joh: Thank you for inviting me. As I mentioned, I'm not a technologist or an engineer. I am somebody who's interested in teaching and writing about Fourth Amendment law: how human police officers are regulated. But increasingly, I've been interested in how the police think about adopting technology, and of course, the new buzz words of the day: AI and robots.

I thought, "What would headlines from the future look like?" I came up with three, off the top of my head. One is, "Calls for Investigation into Seattle Police Robot's Decision to Fatally Shoot Unarmed Man." That's one. Number two: "Ford Offers to Repair All 50,000 Autonomous Police Vehicles After Automated Ticketing Errors Found in Software." Number three: "New Claims that New Patterns of Neighborhood Security Robot Platoons Are Menacing Local Homeless."

You know, you could come up with a lot of sort of dystopian headlines like that, but I created these because they're the kind of questions that are somewhat familiar to us about policing. These are questions of force, accountability, but they also involve the kind of what would happen if these new capabilities were adopted by the police? I think it's safe to say there aren't any autonomous police robots roaming the streets of America, not in the way that we think about them and certainly not as a singularity.

But police departments are in fact interested in robotics. They're interested in what might happen or how this might help their mission, and I think the way to look at it is through the way that the police often get their technology, and that's through the military. A lot of technology is

6. EX MACHINA (Universal Pictures 2015).

7. TRANSCENDENCE (Alcon Entertainment 2014).

often adopted—either directly or in modified form—from military options. And the military, as Ryan has already noted, is quite interested in robotics. They have spent literally millions and millions of dollars in robotics research.

Some of you may have heard about what kinds of thing they might do, including doing benign things like having medic robots to help wounded soldiers, but also more troubling things, like having armed robots to replace the kind of normal things that human soldiers do. One of the big issues in military robotics is, "Should we have autonomous armed robots?" Here are just a couple of things I wanted to mention. In January, just last month, the head of the Australian army said that autonomous robots were a threat to any nation state that believes in the law of war and ethical conduct.

And at a security conference just this week in Munich, the head of German Cybercommand said, "We have no intention of purchasing autonomous weapon systems." Now, that's not to say that other countries are not going to. There are some countries that have already invested in that technology. I know you said in Israel, there is an autonomous weapon system. I don't know a whole lot about it.

But what I like about its description is that, and this is another old cultural reference, but does anyone remember the Ronco oven? You set it and forget it, right? Well, it's been called Fire and Forget It, and it just kind of goes on its own. There's another one at the Korean Demilitarized Zone, which has the ability to identify threats at the DMZ. It has an autonomous mode, but apparently it's not being employed right now. So, those capabilities are out there, so I think the real questions are really about should these be adopted?

When they're adopted by the military, the American military perhaps one day, will they eventually be adopted by local police departments? And if they're adopted by local police departments, well, what would that look like? And what kinds of legal questions are there? There are certainly questions like: who's responsible if somebody dies as a result, or someone is seriously injured?

And that's a problem in the sense that Ryan's keynote talked about. The singularity has the potential to sort of break our ideas of law, but certainly, even in the short term, if we have some kind of autonomously armed police robot, it has the potential to really undo our ideas about why we give police officers, human ones, the deference we do in their ability to use force. We do so because they have very human fears about saving their own lives or saving the lives of others, but it's very hard to translate that to a machine, right?

You can see the appeal of having a police robot from the perspective of the police department because a police department certainly wants to save lives of their officers in the same way the military does. You know, anything that would increase the safety for police officers is a no-brainer. We want that to happen. I think there are basically some very important questions regarding adoption, and if we adopt robot forces, should they be armed when it comes to policing? And what would being armed mean?

And I think the paradigmatic example is always a fatal shot, but even if we didn't want police robots to have lethal firearms, would it be okay for police robots to have non-lethal force? Non-lethal force is still force. It hurts a lot to be tased, and it's not a small thing, and would we want that? I mean, again, even the simple justification of increasing human safety may not be enough of a question or the appropriate question.

So, not only do we want to think about the why, the why adopt this kind of question, but who should have a say in how these kinds of robots would be adopted by the police. You know, policing is a very special case. We think of it as a public service. If a vendor comes to a police department and says, "we want to sell you this autonomous police drone," I think we want to have a lot of stakeholders in thinking through the questions of who gets to make those kinds of decisions.

That's just policing. My imaginary headline also considered the use of armed robots in the private sector. There's already a market for that as well. There's a market for autonomous security robots. Many of you may have seen pictures of that: a 300-pound Knightscope robot that's used in California and elsewhere. People want to use it because it's a cheaper option than paying a human being to do the same thing. And one of the things about law, and I think Ryan will agree with this, is that the law is very flexible.

So flexible, in fact, that maybe it folds in robots too easily. If you think about our ideas about self-protection and what we're allowed to do to protect ourselves and our homes, it's not clear that many of these uses would be obviously barred. And so, we need more than the common law. We need expertise, we need ethical considerations beforehand, rather than waiting for someone to be injured, or waiting for questions of harm or inequality to appear.

But those are some of the kinds of questions that are coming up with policing, and policing's such a fundamental issue here. I think that a lot of what we're talking about here are very American problems because in our system, policing is very much a local activity. We can't have a kind of top-down approach that might be possible in places like Japan. I think they have an AI panel in Japan, or a task force. We simply don't have that kind of top-down control possible in the United States.

Calo: So, in a moment, I'm going to ask Blake about, to talk through some of the value kind of decisions that roboticists have to think about because I know some of you are thinking about them. But I have one or two follow-up questions for Elizabeth. Your point is such an interesting one about, I do agree that there is a sense in which, you know, it does fold rather easily into constructs and perhaps it shouldn't, right?

One place where I would anticipate resistance, and this is something that I'd like to get your view about as well, is what about how robots feel to us? In other words, your question is about like, okay, so if the officer's not physically there, you know, if she's in the precinct and the robots on the street, then that would remove that "I feared for my life" rationale. You know what I mean? I get that. And that feels like an extension of the sort of telepresence.

If a system is just fire and forget, that raises questions about who is responsible for the decision that happens. Kind of an emergence, as you put it, and as I think about it, my work too. But what about this idea that robots feel like people? You know what I mean? And so, how should we think about, is there a sense in which either the interaction between the military and our domestic police or a sense in which, I mean, because I don't think that ordinary citizens want to get involved in every question the police ask about procuring.

You know what I mean? Like we want to replace all of our floor mats. You don't have to have a town hall. So, what is it about, is there anything about robots that is wrapped up in this way that we feel about them? Does that make any sense? And their social feelings to us?

Joh: That's a good question. It certainly depends on how they'll actually come to us in the domestic sphere, right? I mean, everybody likes to think about this kind of like, "Oh, Robocop! I sure don't want a Robocop in my neighborhood!" Well, it's not clear to me it's going to be Robocop. If you look at the prototypes in the private security sector, they look like big heavy lumps. Big heavy lumps where a little door can come out with a cattle prod.

You know, people are not going to feel like so warm and fuzzy about that. I think it will scare people a little bit, but it won't have that kind of uncanny feeling, like, "Oh, this looks like a human. I'm confused." It's not that sort of an issue. But the distance issue is really interesting because it raises questions about: what should police do for their communities? Under our Constitution, there's no minimal guarantee to policing.

There's no minimum guarantee to security. That's not the way we define our rights, our federal constitutional rights. You can imagine in the future, what if you just had a police department that was just totally outsourced to robots? Like it's just so cheap? Like the community says,

"Why should I pay for human beings when these guys are going to do it for free? Or like virtually for peanuts?" Right? Do we owe people the right to other humans to enforce the law and having them have other kinds of community interactions?

You can imagine a kind of rosy future in which robots could do a lot of the boring stuff, a lot of things that police officers don't want to do anyway, and that frees up time for cops to do stuff that we want them to do, like community policing. But the problem is, if you look back at the history of policing, the two biggest twentieth century innovations in policing, technologically, were the car and the two-way radio.

And they did one big thing, and that is they removed human beings from neighborhoods. It did a lot of other good things. It allowed for management to control what individual officers did. The technology gave them this enormous distance, and they weren't as involved in face-to-face interactions anymore. And robots can kind of have the same effect.

This telepresence idea: a big development in incarceration is forcing the only kind of interaction with an inmate to be through video conferencing. That seems good enough, but not quite, right? And you can certainly think of communities saying, "It's not enough for me to just talk to a machine when I need help. I want to talk to another human being." And that's the kind of question we'll be forced to consider.

Calo: Okay. Well, thank you very much. I think we'll come back to aspects of that conversation. So, Blake, I know you've been thinking about values in technology and new robotics in particular, and I wonder if we, you know, one of the things I think about, to sort of tee this up, is what kinds of obligations do engineers have? Especially in a world where you're just not sure how your technology's going to get used?

You know, for example, it wasn't necessarily that the folks who made the police robot in Dallas thought that the police were going to strap some explosives to it and blow up an active shooter. I mean, that was not a use that they contemplated. I'm just wondering, how are you thinking about the role of ethics and values in the designs of these systems please?

Hannaford: Sure.

Calo: And would you prefer to just answer, sir? Or do we want to do slides?

Hannaford: Again, I have some slides because I have taught a course in professional issues for engineers for our undergrads, and this is only one tiny part of it, but what was missing, I felt, was there's a lot of emphasis in that kind of course on, you know, avoiding conflicts of interest and sort of being a professional as an engineer and what are your obligations.

But what's new, coming down the pipe, is if you design an artifact, what are the ethical implications of that artifact having the ability to make some decisions or autonomy? So, being very much more of an engineer and much less of an expert or scholar in the ethical side, I was eager to get reaction from all of you guys to what I'm teaching. So, I have a few slides.

So, again, the context of this is just one lecture in a course on a lot of other topics for undergrad engineers. And also, my research area is medical robots, and on the spectrum that Howard laid out, the medical robots that are right over here at VM and Swedish and UW and Harborview and used on over three million patients so far worldwide are purely teleoperated.

They're the puppet end of the extreme, where not one little millimeter of motion is made that doesn't come directly from the surgeon's hands. But as a researcher in academia, we're looking ahead to adding these automated functions that may support the surgeon. And you know, it's way too far off in the future to think about an administrator just pushing a green button and then the patient goes in on a conveyor belt with an appendix and comes out without an appendix and all sewn back up. You know, nobody's talking about that.

Nobody even sees any reason that that would be good today. But it's a nice, extreme case. So yeah, really, please help me improve this material. And one inspiration is, a student of mine, Andy Lewis, one day after spending a whole day trying to get a robot working and just getting a ton of frustration, came into my office, and said, "It's days like today that I'm pretty sure the robot uprising isn't happening any time soon."

It's a very important quote for me because you kind of have to be down in the weeds and actually working on a robot to realize how far away we are from the scary robots, in my opinion.

Except for this kind of thing, like maybe strapping a bomb into what was called a bomb-disposal robot and turning it into a bomb-delivery robot. So, there's all those things to consider. And I want to acknowledge these wonderful people I ran into at a workshop, Susan and Michael Anderson, who actually are professional scholars in this area. So, we start with Asimov's Three Laws.[8] Everybody knows it. It's really accessible. It's a really durable idea. There's a great cartoon about it, why they're in the order they are given in, okay?

So, if you say, if you mix the three laws sequence up just a little bit, you get a kill-bot hellscape instead of the goal of Asimov's Three Laws. So, you know, is it sufficient for the robot to behave ethically, like following rules, or does it have to have ethical reasoning? Does it have to

8. Susan Leigh Anderson, *Asimov's "Three Laws of Robotics" and Machine Metaethics*, 22 AI & SOC'Y 477, 477–78 (2008).

be able to explain why it should not take this step? And those are very different ways that you might have to design a robot. You know, humans aren't really all that ethical, right? So, are they the standard?

Maybe robots, straightforwardly, would behave more ethically than humans do? We all take that little shortcut and cheat sometimes, right? So, it's not necessarily all doom and gloom. They introduced these two theories of robot ethics, teleological and deontological. So, in the teleological approach, you know, only the outcome is what we care about, regardless of the means, and with the deontological approach, as long as you follow these laws and principals, it's ethical. Like the soldier kills somebody, but they're doing it in an actual war, according to the laws of war, according to their orders and training.

So, that's sort of two different ways that robots could be coded. There's a mathematical definition of the teleological thing, which we're trying to emphasize because that'll make the engineers go into their comfort zone. Even though this is kind of an obviously trivialized equation, nobody knows where all the numbers would come from to score all these things. But the problem is that, you know, maybe drinking soda (quite legally relevant today in Seattle) has a pleasure intensity, but what about longer term problems?

And also, if you just add up the total net pleasure or benefit or utility to people, what if that really screws one member of that group? So, the total utility is really maximized and that's great, but only because one person gets thrown off the bus.

By contrast, deontological is following duties, and you can emphasize rights and justice, and you avoid the need for this fuzzy, weird, phony math, but you can also ignore serious consequences.

And this is the famous ethical problem, the Trolley Problem, and we kind of go into which way you should throw this lever. The two approaches are that sure, throw the lever because if you throw it, then five people are saved and one person dies, so you've got four net lives saved. But if you just follow the principle "don't take any action which kills somebody," then that's the deontological approach, and I had one student in the class who raised her hand and said, "Yeah, that's what the nuns taught me in Catholic school."

So, we began a discussion and realized that robots probably should have downloadable ethics packs that are generated by places like the Vatican and the Imams in Saudi Arabia and so forth for the different religious beliefs of the owner of the robot.

A really wild idea I don't know much about is "ethics mining," where you mine the web with software for examples of the moral dilemma that

you're thinking about, and then look for good outcomes and bad outcomes according to some rules, and make decisions that way.

So, quickly, on to the singularity. I don't need to introduce it, but one of the ideas is that as these major increments in human evolution occur, they get closer and closer together. And there's a lot of skepticism, though, about whether the singularity is really a thing.

So, there's really no evidence for it. It's just a hunch. That's what Pinker says. "Machines have no beliefs, desires or motivations," so they never will supplant people who do have those things.

Very high machine intelligence will cause economic chaos before the singularity and it will sort of implode. Or, you know, observing that all of those paradigm shifts are getting closer together in time is biased because we look at the most recent ones as more significant than they really are. So, my personal view is that intelligence is overrated, okay?

If it was such a great thing, if intelligence was so powerful that we have to fear it, then the smartest people would be our leaders and our leaders would be our smartest people and probably all political sides would disagree with those assertions. So, I think we have more to fear from ignorance than from intelligence. Thanks.

Calo: Okay, great. Thank you very much. Okay, so, thanks a lot, Blake. And so, in the spirit of that you put it forward, which is, in terms of soliciting feedback, I hope that folks, when it comes time for Q&A, will offer Blake some thoughts. I mean, my thoughts are, number one, is that as your presentation hints, there's a rich and diverse community around this, and indeed, I hope you can stay for the next panel, which focuses specifically on ethics.

And so, Anderson and Anderson is one take, and there's a number of them, is what I would say. And by the way, for the organizers, how much time do we have? Because I'm not sure that I'm tracking the amount of time.

Reux: You have about fifteen minutes left before we start Q&A.

Calo: Fifteen minutes? Okay. Great. Before we start Q&A? Okay, perfect. That's awesome.

Speaker 2: Can I ask a question?

Calo: Yes, please.

Speaker 2: So, I'm fascinated by your talk and thank you. Do you think that that framework works best where the robot's objective is something that we could all agree is the thing the robot's supposed to do?

Hannaford: Sorry, which framework?

Speaker 2: I mean, just having an ethical, you mentioned, maybe robots could have an ethical app that they could download?

Hannaford: Oh, yeah.

Speaker 2: A set of considerations? That works best when we all agree what the robot's supposed to do, with precision?

Hannaford: No, no. I was saying that this would be religious beliefs. So, literally, the Vatican would have a little branch of presumably Jesuits, I guess, who would actually code this, but a little branch that would code the ethical rules that the Pope thinks are correct, and then Catholic owners of robots would plug it in. And you know, maybe the law pack comes from your local government, and then the ethics pack on top of that comes from your religious leaders.

Calo: I mean, do you want to jump in now? Or do you want to wait until Q&A?

Speaker 3: Yes. I guess I think, it's interesting, Blake, you started at one point saying, do we want ethics where the AI or the robot is following a set of rules? Or do we want ethics where the AI or robot could explain why they did what they did? And I think that's an important distinction, and I think we're kind of blurring some lines when we say the Vatican would have an ethics download. Religious beliefs are not ethical beliefs necessarily, though some ethical beliefs are informed by religious commitments, right?

So, I think we could make some distinctions there. But I would argue you should probably veer to that second fork, where AI or the robot would be able to explain their behavior and explain the principles by which they made that choice. And we'd be in much safer territory than looking for the one download pack that everyone on the earth would agree are the right set of rules.

Instead, we'd say, "This is the principle, and you may not agree with it, and we might have to course correct," but I can explain where I'm, the entity can explain why it made the decision it did, and then we can have an intelligent debate about what to do next, right? And I think that's where we want to head with ethics.

Hannaford: But the nature of that explanation could be quite different. There would be a lot of different ways that could happen, like, "I followed the law until it was ambiguous, and then I used my Islamic ethic pack to resolve it." Okay? And that's an explanation. But it's not quite the same as saying, you know, "I prioritized these principles." They're not quite the same thing, but it seems, in an engineering sense, the former seems a little more achievable.

Calo: Howard, do you want to jump in?

Chizeck: Yeah. So, I'm going to kind of bring this down to an actual case. I think something that's missing here is the importance of the individual. So, in my lab, one of the things we do is, we're building devices that adjust stimulation in deep-brain stimulators for people with

Parkinson's or essential tremor, and they do this by some machine learning that matches patterns that say it's time to turn on stimulation and time to turn it off. And the question is, what does the patient think about that?

They can get a clinically available device now that's always on, but what about this thing that's in their head that's turning on and off when they don't expect it? So, we actually have a philosophy student, a PhD student, and embedded emphasis in the lab asking questions about this from patients, and we get different answers from different patients. And a version of this device allows the individual to turn up or down their stimulation by thinking about it or by moving their hand in such a way to trigger that, which would be a very good thing for someone who wanted that control.

It might be a very bad thing for a person with Obsessive Compulsive Disorder to have available to them. And so, somewhere underneath all of this, I think, in response to Blake, is the idea of who's the individual involved or the population involved in policing this? What's the right thing to do might depend on the individuals involved, so I'm questioning whether there is an absolute set of rules that can be followed or whether it's got these adjustments.

Calo: So, I do want to take a moment. Obviously, clearly, the questions of ethics and the questions of law are intertwined, right? And it can be somewhat artificial to pull them apart. But that's the decision that we've made for the symposium. And so, what I wanted to do was table this direct conversation about ethics and pivot to a legal question I think is closely related, that we can hash out among ourselves, and then hopefully return to this rich conversation about ethics in the next panel, which is dedicated to it, if I may.

And the question, though, that I wanted to pivot to now, and I'll start with Elizabeth, if I may, is one of the things, that conversation . . . is happening in robotics law but also in neuroscience, as you know, and you know further as a criminal law and criminal procedures scholar, is this idea that suddenly, because of technology, we can control aspects of ourselves or might lose control over them, right? And so, a famous example would be where a person had sexual predatory behavior due to a tumor, and when they removed the tumor, the person ceased to have that predatory behavior.

And then, the predatory behavior began to come back in and they found that the tumor had come back, right? And this poses a pretty deep challenge to the law that thinks about this idea that people intend what they do and it's not like some material, sort of fatalistic idea, but rather that there's some meaning there in terms of what we do.

But what Howard's work and Blake's work make me think of is there are going to be situations where people can control their own behavior and

their own mindset in ways they never could before, and also situations such as with use of prosthetics, where there might be even a disconnect between what the person intended and what actually happened in the world. Do you see that as a deep problem for criminal law that's on the horizon? Is it something you think about?

Joh: So, I haven't thought about it deeply, but my first response is I think it is a challenge. I mean, one of the basic sort of concepts in criminal law is the voluntary act, which is always premised on this idea of kind of free will, right? That you as a human being have made this decision, that you can actualize into the world, right? But if there are situations where there are enhancements, or even the robotic enhancement has some capabilities that are not fully in control by the user or the human being, you can imagine situations where harm results, right?

And this person is now in some criminal trouble, and they try to make some sort of argument that, "Well, it's not truly me," which would seem like an odd argument to make, right? This thing that you sort of colloquially think of as yourself is now, for a criminal law purpose, not yourself. I think that is a pretty big challenge—this idea of enhancement, that we could enhance ourselves. But I think one of the biggest things about, I was thinking of this example that Howard used, about emergent properties in robots.

I mean, I think that's another angle to bring into this criminal idea of responsibility. You know, what if machines talk to machines, they help us as individuals, but there is something emergent about their—I don't know—machine conversation that might also lead to certain kinds of social harms. It's unclear whether that would be addressed through criminal law or through tort law liability, and one could certainly think about that.

Now, I'll just swing back to my policing interest, and that is this idea of individual responsibility. Like in policing, we think of police officers as making decisions about where to police and how to police and discretion. With robots, you know, we need to think a little bit about how much discretion. You know, all the steps are here. Now you need to make that final decision. How do you do that? If there's emergent behavior, whether that's an extension of an individual officer or a completely autonomous robot, that's also a kind of collective question that we don't have a good answer for.

Right? Like a quick example would be what if a group of autonomous police robots, or let's not think of human beings. Cars, right? Autonomous police cars just decided we would just, twenty-four-seven, drive around the worst neighborhood in Seattle, right? And the people there were like, "What are you doing? You're like fencing us in here."

And the human beings behind it are like, well, I don't know, they figured it out, like, we don't have an answer, right? I mean, it really cuts against this idea of individual human responsibility in some kind of way that we don't have a good answer for, right? And as I said, you don't need to get to the far future to get to that.

Hannaford: I agree.

Joh: It's coming pretty soon.

Hannaford: Here's a really quick example of that. Imagine somebody has a prosthetic hand with dexterous fingers, which exists already, and they hear a noise in their house, so they grab their gun. And then, they shoot this intruder voluntarily by activating the trigger finger on the robot, and it turns out to be the repairman. So, they're in trouble. But they say, "No, I didn't do that. I was holding the gun. I admit it. I was pointing it. But there was a glitch and it just pulled the trigger."

Calo: Right. Right. And I guess to turn that into a question for the engineers, to what extent are you all thinking about this? In other words, is there . . .

Hannaford: Unfortunately, you just saw it.

Calo: Well, because I mean, I have a project at the tech policy lab that Howard is affiliated with, and the project app is basically looking into the design of systems that are more algorithmically favored, you know? That is to say that it's a project looking at when people design systems with whatever values they are, whether they come from the Vatican or wherever, when you're designing the values into these systems.

And one of the questions that we're posing to folks who design these systems is: to what extent are you anticipating legal challenges? Do you see what I mean? So, you deploy a system that's supposed to help people make decisions. Sometimes those decisions will be challenged because the person didn't get the credit they wanted, or they got denied benefits, or they wind up in jail.

And preliminarily, we're not seeing a lot of thought about them being challenged. So, I mean, as engineers maybe one thing that you could teach your students or think about yourselves is what happens when there's a challenge to something that happened due to my technology? Howard, can you start with that?

Chizeck: Yeah. I think there's a couple of things. We're seeing this now with databases used for decision making. I mean, there have been legal challenges on credit worthiness where it turns out, because of the data that's put in, people of certain racial groups are excluded or given lower scores for unjustifiable reasons, but because the data's there. One of the camera companies had a problem. Their cameras didn't work for people with dark skin.

And so, all of their photo recognition failed for a population group. But the reason was they didn't have a sample of enough diverse people when they were training the algorithm. So, the idea that the data that you provide for the training for AIs has to be representative of the appropriate need, I think, is now out there, and you know, Google and Amazon, they're all hiring database and AI emphasis all of the sudden because they know there's a legal situation here.

In the automobile, self-driving car area, Waymo made a decision that they weren't going to allow humans to take over for the cars because it turns out that that transition is very, very difficult, and the accident rate for automated vehicles, fully autonomous, is way, way lower than humans. So, they made the decision, okay? Other companies said, "No, we have to let the humans take over, despite the risk." People are thinking about it, but there isn't an answer because I think part of it is there's clear law, and there also isn't an experience base.

If you haven't had it long enough to know what the right strategy is, you're kind of making it up as you go along. And I think that is a challenge in engineering for these things; when you're making new things that don't yet exist, there isn't an experience base to draw on, which makes it very, very difficult.

Joh: And I think just a brief thing. I mean, that's such an interesting comment because the world that one could imagine of robots in policing could lead not just to the big philosophical questions about ethics, but practically speaking, what would happen if every autonomous police vehicle gave you a ticket? Every one of us sitting in here gets a ticket.

The minute we get in our car, we're all speeders, right? We haven't anticipated this world in which all of a sudden there's perfect enforcement of the law. That's not a robotics decision. That's a legal question. Do we want to live in a world of perfect, total enforcement?

Calo: Assumingly not, because . . .

Joh: Of course not.

Calo: There have been challenges to the stuff that can, the cameras, right? And there was a great paper for a conference that I helped organize, it was by Karen Levy and Meg Jones, and the paper was drawing from referees and the debates in sports about whether or not we should do everything with just cameras. And a lot of times the idea was that that athlete should have a sporting chance in order to steal that base or whatever, they get away with it, and that's why people objected, and they were analogizing.[9] Okay, so I think it's time for us to open it up for questions.

---

9. Meg Jones & Karen Levy, Sporting Chances: Robot Referees and the Automation of Enforcement (Mar. 2017) (unpublished manuscript), http://www.werobot2017.com/wp-

Joh: It would be great if we did.

Calo: Yeah, absolutely. You know, we're back on time. Yeah, go ahead.

Speaker 1: I want to address the taxonomy question because we moved on and kind of dropped this. When I was writing *Robotica*, I was really struggling myself with how to identify the state of robotics currently, and then what people were talking about was futuristic robots, and you know, I thought this was a very large category of first-order robotics and second-order robotics. But how I ended up, or we ended up, defining second-order robotics really made it almost unlikely that anything that we have today would be deemed truly second-order robotics.

I mean, and what we were talking about is the point at which foreseeability stops, and thus human liability for a creation stops. I mean, in the case that you were talking about, with the hand that pulled the trigger, if he says that was a glitch, it's very clear we have a way of determining whether the glitch really existed. And if it did exist, there's someone who's clearly liable, and that would be the producer. But that robot is not a truly autonomous, truly intelligent robot.

The truly intelligent robot has to be one that operates in an unstructured environment that is entirely adaptable and that is self-corrected. And the only example you gave that even remotely touched on that were the robots that were creating their own language because, somehow, they were adapting to achieve an end that was given to them, and they realized they could get rid of prepositions and articles and all of that and move more quickly.

But what they were doing, in a sense, was self-correcting. So, I'd like to know two things. Is this dream of, well, I don't know if it's a dream or fear or thought, of second-order robotics a realistic one? Are we likely ever to get to a point of truly intelligent robots with a level that we would call consciousness, and the ability to self-correct, to rewrite one's own code, to adapt to unstructured environments? Is that realistically likely?

And if that is true, then that's the point at which the law breaks off because at that point, the creator no longer has foreseeability, and it seems to me, the only way of dealing with that would be some massive insurance system that says, "If we sell you a robot that becomes self-correcting, does something that is not foreseeable, you know, we're going to have to have the insurance system where everybody pays, essentially, for the error of the robot." But I'd like your comment on that.

Hannaford: Yeah, that stimulates a lot of responses. I mean, we insure our cars routinely, but I think, again, going back to Andy Lewis's

---

content/uploads/2017/03/Jones-Levy-Sporting-Chance-WeRobot-draft.pdf [https://perma.cc/X2NM-CFLB].

remark, I think we're a lot farther from that than people say. And you know, Ryan, you introduced the feeling people have about robots, and I would say those feelings change just as rapidly, even more rapidly than the technology changes. So, when the internet was a new thing, it seemed kind of crazy to type your credit card into a website.

It was like, "What? That's way too risky." Now, you can't really function unless you type credit cards into websites. It's even automated for you by the browser and you just take it, right? So, you know, these emotional responses are like, for example, think about the viral videos from Boston Dynamics; probably everybody has seen one of the them. The robot doing a backflip,[10] the quadrupeds which work together and open a door[11]—that was last week's one, right?

And as a roboticist, I would caution you to watch those and really not extrapolate, okay? The robot, it's an amazing, that is a wonderful company with some of the absolute top smartest robotics people anywhere working there, and the video is not fake, okay? It really did that. But extrapolating even one little thing beyond that video, don't count on it at all, okay?

Calo: Was it autonomous? Blake, do you think? I couldn't tell.

Hannaford: Yeah, those are autonomous.

Chizeck: No.

Hannaford: Those are autonomous.

Calo: So, *Wizard of Oz*?[12]

Chizeck: No.

Calo: It's a term of art.

Hannaford: But I just want to say that if that door handle that it opened was like three inches lower, start over, okay? And if that door opened to the left instead of the right, the engineers would have to get everything ready, make a few tests, have a few failures until they get the tape that shows it working, okay? So, people don't do that, though. They see a humanoid robot doing a backflip off of a box,[13] and that's an incredible feat of technology, but there's no evidence that it could do another thing or maybe even just off a different-height box. So, just take those as they are and try not to extrapolate.

Chizeck: I think here I will disagree with Blake a little bit. So, Google has a room, Google X, I guess, alphabet, but they have a room where they have a bunch of arm robots, manipulating arm robots picking

10. Boston Dynamics, *What's New, Atlas?*, YOUTUBE (Nov. 16, 2017), https://www.youtube.com/watch?v=fRj34o4hN4I&ab_channel=BostonDynamics.

11. Boston Dynamics, *Hey Buddy, Can You Give Me a Hand?*, YOUTUBE (Feb. 12, 2018), https://www.youtube.com/watch?v=fUyU3lKzoio&ab_channel=BostonDynamics.

12. THE WIZARD OF OZ (Metro-Goldwyn-Mayer 1939).

13. Boston Dynamics, *supra* note 10.

things up, learning how to pick them up, and the experiment was just to let them try and learn how to do it. So, they didn't give them instructions, and they've been doing it for a year-and-a-half, two years now. I don't know how they're getting on that. But in a sense, Blake is right.

Each new task someone has to program in the beginning, but the quadruped robots that see the door handle can adjust for where the height is and where it moves itself. The thing is, it can do the door handles, but there are other things it can't do. That's what, to me, makes these operating systems like Alexa so exciting and interesting—that there is a collection of skills, thousands of them now, that people are writing for it, and every time anyone talks to Alexa, that's information that could be used to improve Alexa for how Alexa answers and whether it's understood.

Same thing with Siri. So, we are all collectively teaching these large AI operating systems. Now, we're not yet to your taxonomy, but as they learn to speak with us more effectively, or to translate more effectively, you see gradual improvements in these things, and it may just be a lot like how kids learn to do stuff, okay? There's a learning process going on, but the thing the artificial devices have over organic is that they think much more quickly.

So, the learning time could be in months or in days perhaps, instead of in years or decades. So, I suspect that somewhere out there, there are stock-trading robots and teachers-training robots and economic-balancing robots that are doing a lot and have learned a lot. And they are, I hope, not self-aware. But there are things happening now that were not happening five years ago, or three years ago. You know? And my last comment will just be it's always the definition of what is human, right?

So, now that these devices talk to us and listen to us, we no longer say that only humans have speech. We just keep moving where the goal post is for what it is to be human, right? And before it was with vision. So, I don't think robots are ever going to be human. They may be intelligent in a sense, and maybe they'll fit into a taxonomy we could construct, but they're never going to be us because they're not.

Calo: I think we have time for one more question. Yeah, go ahead. Please.

Speaker 5: Yeah, so speaking of sheer responsibility, when you're trying to assign blame, or something did a wrong thing, often with that comes a system of sanctioning that to deter a future behavior—to modify the behavior the machine makes, and beyond just incapacitating a machine that does something wrong, essentially just turning it off and starting over. How would you modify machine behavior, from an engineer's perspective?

Chizeck: MIT did an experiment where they monitored a human observer's mental response, and there are certain signals you give off when you're surprised or when something fits right. You build responses, and they use that as input to the robot.[14] So, when the robot did it the right way, the human was giving it approval and that was reinforced.[15]

Now, underneath there, there's an algorithm like the thing that Blake had, where you have some summation of net good versus net bad. But the idea here is to have a critic, whether it's a human critic or another computer system, that keeps saying, "Oh, you did it right," or, "You did it wrong." Maybe not motivational—no "Atta boy." But telling it, "Yeah, you did it right," or, "You did it wrong."

Hannaford: Yeah, that's the notion of reinforcement learning. So, when Google's recognizing cat photos, it's not reinforcement learning. It's like, "This photo's a cat, this photo isn't a cat" times ten to the seventeenth, or however many pictures they have. But reinforcement learning is more like, there's a goal, and let's say a set of actions that are needed to achieve the goal. And you get a reward when you achieve it, and you don't get a reward or you get a negative reward when you don't achieve it.

So, that notion would be built into reinforcement learning if you can, but the problem is that you sort of have to generate a lot of negative examples to get it learned. It'd be nice to learn it without the bad outcome that you're punishing.

Speaker 5: It's like a child in that way.

Calo: One last comment about that, and then I'm going to turn it over to Reux to close us out and to thank the panelists, but think about this though. I think that the social balancing of robots will matter early in the following way: Imagine you have some humanoid police officer that does the wrong thing, okay? And they bring it back to the lab and Blake and Howard and Maya and other people work on it and say, "Okay, we've optimized the right thing. Now we've corrected the problem."

And then, you go back to the public and you give them the exact same robot the next day, and you're like, "All good! We've fixed him!" You know what I mean? That is actually not going to be satisfactory, and that's, I think, what's interesting, whereas if you say you fixed the car so the brakes work now, that might be satisfactory. And that delta, to me, is very exciting. Anyway, with that, please join me in thanking an excellent panel.

---

14. *See generally* ANDRES F. SALAZAR-GOMEZ ET AL., MIT COMPUT. SCI. & ARTIFICAL INTELLIGENCE LAB., CORRECTING ROBOT MISTAKES IN REAL TIME USING EEG SIGNALS (2017), http://groups.csail.mit.edu/drl/wiki/images/e/ec/Correcting_Robot_Mistakes_in_Real_Time_Using_EEG_Signals.pdf [https://perma.cc/6XFV-BFC7].

15. *Id.*