S2E7: Transcript
Machine Morality
with Dr. Walter Sinnott-Armstrong, Dr. Jana Schaich Borg & Dr. Vincent Conitzer

Tavia Gilbert: Welcome to Stories of Impact. I'm producer Tavia Gilbert, and in every episode of this podcast, journalist Richard Sergay and I bring you conversation about the newest scientific research on human flourishing, and how those discoveries can be translated into practical tools.

Every episode of this season of the Stories of Impact podcast explores the questions: What are diverse intelligences? What do we understand about ourselves by studying them? And, most importantly, how those findings support human flourishing?

We'll spend today with an ethicist, a computer scientist, and a neuroscientist who have teamed together to research how we might build human morality into an artificially intelligent machine. And we'll discuss a surprising outcome: that when we do imbue machines with the capability of acting in accordance with what human beings generally judge to be moral, those intelligent machines can help us deepen understanding of our own moral philosophy and behavior.

I'll start by introducing the three scholars who make up an unusual interdisciplinary team of researchers, all of whose careers have led them to expertise in machine learning, or artificial intelligence. The three voices we'll hear from in today's episode are educators and researchers affiliated with Duke University in Durham, North Carolina. Lead researcher, ethicist Walter Sinnott-Armstrong, is the Chauncy Stillman Professor of Practical Ethics and the Kenan Institute for Ethics in the philosophy department, as well as being affiliated with the Duke Institute for Brain Science and Duke's law

school, computer science program, and psychiatry department. Neuroscientist Jana Schaich Borg is Assistant Research Professor and Director of the Master and Interdisciplinary Data Science. And computer scientist Vincent Conitzer is the Kimberly J. Jenkins University professor of new technologies, as well as a professor of computer science, economics, and philosophy.

With such a wealth of expertise across such a broad range of subjects, it's no wonder that this is an ideal team to explore a common and complicated vision. This threesome focuses their research on how, in the future, the artificial intelligences we deploy might serve humanity and exhibit behaviors that align with broadly defined "human morality." And their work goes a step further: It also considers how AI can support and enhance humans in their own moral evolution, especially in decision-making scenarios in which overwhelmed humans are faced with ethically complex dilemmas that demand a solution that will benefit the greatest number of people.

First, let's establish why this work is important now. Here's Dr. Walter Sinnott-Armstrong, discussing the origins of this project:

Walter Sinnott-Armstrong: Well the genesis is simply our concern about the role of artificial intelligence in our society today. It seems to be pervading many, many aspects of life, some that are quite surprising and raise some concerns. What we want to achieve is to figure out which of those uses are ethically acceptable to people and which we think are morally justified, but also we're going to try to program morality into computers so the computers will be less likely to make decisions or perform actions we find morally problematic.

Tavia Gilbert: And here's Dr. Vincent Conitzer:

Vincent Conitzer: One thing that we're seeing now is that as AI is being broadly deployed in the world, the kinds of problems that we are dealing with are changing. And now, as these systems are being broadly deployed, we actually do care about the objectives that they pursue.

Walter Sinnott-Armstrong: Computers are getting used more and more, in more and more different areas. And so we want to get a little bit ahead of the game so

that when computers start getting used in questionable areas, we'll be more likely to do it right.

Tavia Gilbert: And here's Dr. Jana Schaich Borg, who highlights how this team uniquely responds to the complexity of the question: how do you program human morality into an artificial intelligence?

Jana Schaich Borg: It's a very big challenge and I don't think it would be possible without a very interdisciplinary approach. Our project is dedicated to developing a moral artificial intelligence with a small 'i' rather than a big 'I'. So we're not trying to create something that is equal to human. Intelligence is something that solves problems. And so I don't put a capital "I" on it when I'm thinking about it. We're trying to create something that can help solve problems. And so in that case, what we are putting in there is what humans think is morality, not what the machine thinks is morality. And then whatever the machine thinks is morality is what we define for it.

Walter Sinnott-Armstrong: It's artificial intelligence that we view as doing moral things, but it's also building morality into artificial intelligence so the system itself is making moral judgments.

Tavia Gilbert: If the team is considering how to program artificial intelligences with a human sense of morality, then we need to take a step back. What is the definition of morality itself?

Walter Sinnott-Armstrong: Morality is impossible to define. But the kind of morality that we're talking about, which is not what everybody considers to be morality, is composed of two elements. One, reducing harm to others, and respecting other people's rights.

Jana Schaich Borg: A lot of moral judgment, not all of it, is harm-based. It's based on trying to avoid harm. Usually there's some part of us that is turning the moral scenario into something about thinking about how someone else would feel, what kind of pain it would cause them, what kind of distress it would cause them. And we as humans, and it turns out other species, too, are wired up to avoid others' distress.

Walter Sinnott-Armstrong: So you want to reduce the amount of harm that people suffer in their lives. But you want to do it in a way that doesn't violate the rights of individuals along, along the path.

Tavia Gilbert: So is the team trying to program in a universal morality? Do they believe there is such a thing as a universal morality?

Jana Schaich Borg: No, I don't think there is. But it doesn't bother me because everyone has some sense of what morality for them is. So experimentally, for example, when I ask questions, I don't ask about the universal morality, I ask about what's moral to you. And everyone has some sense of what is moral to them and that's what we go after. Is that the exact same feeling, the exact same concept? I don't think so. Probably not, but that's ok.

Tavia Gilbert: So who does Dr. Schaich Borg think defines morality?

Jana Schaich Borg: You define it. Morality is a very personal thing. And so what feels moral to you is the important thing.

Tavia Gilbert: The team doesn’t need to arrive at a single definition of morality. In fact, their research relies on gathering data about lots of humans’ viewpoints about what constitutes moral thought and action.

Walter Sinnott-Armstrong: We want the computer to reflect human morality in general, rather than a particular individual.

Jana Schaich Borg: So part of what we need is a lot of data about how people make judgments and what their judgments are. And in the past we've always had to do that by inviting people into a lab and asking them some questions. But now there's been a lot of both work done, but it's kind of just new creativity around the ideas of how could you collect data from everyone? What we've seen in the past couple of years is that you can crowdsource science. And so with that development, we then started to realize well now we can get the data we need to actually train a moral AI.

So as technology started to develop, and it started to become possible to start thinking about something else making a moral judgment, it was a very natural switch to think well first of all, how would you design one? But then second of all, how could we use the new, developing technology

to make us make better moral judgments. And for me, it was also natural to move from thinking about how does the brain make these computations to how would you make something else make these computations.

Walter Sinnott-Armstrong: We are not trying to take our favorite moral theory, build it into a machine, have it apply to problems and have everybody agree with us. That's not the goal at all. Our project looks at survey data about what a wide variety of people take to be morally relevant features and then develops an algorithm again based on experimentation with actual humans alive today, to determine how those different features interact. And then the algorithm predicts which judgments humans would make.

Tavia Gilbert: In order to research how they might program an algorithm to be moral, the team needed to study over time a real-life situation, one that was ethically complex, demanded difficult decision-making, boasted ample data points to evaluate, and offered a space in which emotions, empathy, instinct, and, potentially, errors in human cognitive evaluation, were at play.

Walter Sinnott-Armstrong: An example that we have focused on is kidney exchanges. Sometimes when you're distributing kidneys to potential recipients you have to decide which one gets it because you've got one kidney and lots of potential recipients. So should you give it to younger people rather than older people? What if the person has been on the waiting list longer than another person? What if this person's, they're responsible to a certain extent for their own kidney problems because they did something that helped cause those problems? What if they have large numbers of dependents at home? There are lots of different features that people might take to be morally relevant.

Vincent Conitzer: And now you can imagine that if you're running an exchange, you get all of this information from all these different people what their medical data is, who might be matched with who. And now you have to sort through this enormous number of people and figure out what is the best way to match them all together. And if you think about it for a while, this quickly becomes very overwhelming. You might find something reasonable, but how would you know that you found the best way to match them all? Even if what best means is just maximizing the number of people that get

a kidney, that already becomes a hard problem. And there is a sense in which it's actually formally, computationally hard.

So this is where algorithms come in because they don't mind searching through a large base of possible alternatives, right? Whereas you or I would get bored of this and overwhelmed and might start making mistakes, computers are very good at systematically searching through a space of alternatives. So that's I think the primary reason that AI is getting involved in these kinds of things.

Jana Schaich Borg: So the main reason to not leave moral judgment to humans is because humans are fallible. So, we get tired. We get tired, and we lose cognitive resources and emotional resources. So by the end of a day, when you're tired, when you're sick, when you've just made too many decisions, you often can't make the decisions as well as when you started. And so what we're trying to do is just build some help. It helps you be efficient. So you don't want to be taking into consideration every single detail of every single moral judgment all the time. And that's why instinct can be good, but then sometimes it gets you into trouble because sometimes you need to be taking into consideration things that you aren't.

Tavia Gilbert: And in the context of a kidney exchange, relying on human intuition can be complicated, at best.

Vincent Conitzer: So in this particular context, people's intuitions obviously disagree. Some people would, for example, say that you should take into account whether a patient has dependents, like small children that they're taking care of, and other people would say that you should not take that into account, right? So that's a controversial feature.

Jana Schaich Borg: And in the field, there are some people that say you should not be making emotional judgments. Others would say actually what matters is that you have the right emotions and that your emotions match the actions that you're performing.

Vincent Conitzer: We have to deal with this problem of how to set policies or objectives for algorithms based on the opinions of multiple people. Which are the people that should make those decisions, and what kind of information

should they have access to? In the context of kidney exchange, you might think that the best people to make these kind of calls are people with medical training. And there seems to be some truth to that. But on the other hand, for determining whether it's relevant that a person has dependents or not, it doesn't seem that medical training particularly qualifies them. And so maybe what you really want is kind of a jury of your peers, randomly selected members of the population, that nevertheless have access to expertise that allows them to understand which of these features are relevant, what we need to think about in the context of kidney donation. But I think this is still a very open question, and I think that's one that really as a society we need to be thinking about.

Tavia Gilbert: Whether it’s humans who prioritize or shun emotion, medical professionals or a “jury” of non-medical peers, until AI can take its place in complex decision-making, human beings will continue to wrestle with ethically difficult choices, including in scenarios like the kidney exchange.

Walter Sinnott-Armstrong: So humans are going to make those decisions currently, and they're going to be subject to kinds of mistakes that we think a properly programmed computer could help avoid. They make a number of different mistakes. They, for example, overlook morally relevant features. They get confused by very complex problems, and they have biases. And they're not very quick. But the machine can give you a better sense of which judgments humans would make if they considered all the morally relevant features, that is the features that they themselves take to be morally relevant, which are legitimate. They don't think those are illegitimate biases, and they're not getting confused by the complexity of the situation.

Tavia Gilbert: How does Dr. Sinnott-Armstrong define bias?

Walter Sinnott-Armstrong: Bias is when you make a moral judgment or make a decision on the basis of factors that you should not be making. Now, you recognize it as a bias when you yourself recognize that. So for example, if I walk into a room and there's an African-American and I sit five feet from the African-American, whereas if this person was European-American I would have sat one foot from them and started talking to them, then that's a

bias. Now, I don't want to be like that. And yet the data shows that most people are like that. They've got biases of that sort that are going to affect their daily lives even though they're not aware of them and even though they think they're wrong. And so it's not unusual to think that these hospital administrators would have biases that are affecting their decisions that they're not aware of and that they themselves think are wrong. It happens to all of us. We're all human, and the computer can help us figure out when it's happening and then correct for it.

Tavia Gilbert: Dr. Sinnott-Armstrong believes that AI can act as a safeguard against human bias, and can even enhance the development of human morality. AI can act as a check:

Walter Sinnott-Armstrong: A check in the sense of preventing the most common errors. Ignorance, confusion, and bias. And enhance? Absolutely. Because computers don't forget morally relevant features. Computers don't get confused by complex situations. Computers don't have the biases. If you don't enter the person's race into the dataset, they can't be biased. But humans, they see the patient, they know what race they are. You can't you can't avoid it.

I don't want to be ignorant. I don't want to get confused. I don't want to have biases that I myself see as inappropriate. So if the machine can point those out when I'm doing it, I'm going to learn to avoid them, and my judgments will be better. That's the sense in which my judgment will be enhanced. I have to make the decision in the end, but I want to make the best decision, and the machine can help me do that.

The human still has to make the decision, but the computer can say this is the decision that you would make, given your values and the things that you yourself take to be morally relevant, given them interacting in the way that you yourself take to be appropriate, and getting rid of all those biases that you yourself take to be features that should not figure into your moral judgments. Then the computer can say this is a judgment you would reach if you corrected your own judgment in that way. And then if the judgment you think is right or wrong disagrees with the computer, now you've got a problem. But if it agrees, you feel, ok, the computer has helped me confirm I'm going to be more confident. If it disagrees, then we have to do some research, we have to think about it more.

One upside is that it's going to improve human moral judgment, because the hospital ethics committee now is making these decisions about who gets a kidney, they've got no check on them. How do they know whether they got it right. There's, there's nothing to tell them. I've been on these hospital ethics committees before and you reach a consensus, but you still think, I don't know. It was a tough case. The computer can either agree or disagree, and when it disagrees you have to think about it more carefully. And what that means is that you might have to recognize, I forgot something in that case. I got confused in that case. Oh no, I had a bias in that case that I don't want to have. And so it can train me to be a better version of myself by reflecting the judgments that I would make if I were not ignorant, confused, and biased.

Other upsides is, people get kidneys when they ought to get kidneys. If there are fewer mistakes, then the people who need the kidneys and who deserve the kidneys are going to be more likely to get them. The machine can do that. So the machine can actually lead to better decisions in a medical way as well as getting rid of bias in a more moral way.

Tavia Gilbert: So will AI one day entirely replace humans in decision-making?

Walter Sinnott-Armstrong: Will people stop thinking about morality because they've got the computers doing it for them? And I want to say, the answer to that is no, because the machine will not be replacing the human. The machine will be operating as a check on the human. If you simply handed over the decision about kidney exchanges to the machines, you wouldn't need the hospital ethics committee anymore, but that is not at all what we're proposing. What we're proposing is that the committee use this machine to correct them, to enhance their judgment and to learn from. Then they're not going to stop making judgments themselves. So I think this fear that we're not going to understand morality or even be able to make moral judgments, is misplaced. In addition, I want to say that these programs can tell us a lot about our own morality.

Vincent Conitzer: Today's kidney exchange algorithms, while they're very sophisticated in searching through this large space of possible ways of matching people, they have no concept whatsoever of what it's like to be a person, what even a kidney is, right? To them it's just zeros and ones that they are

interpreting in a particular way just for the purpose of this matching problem, but they have no concept of what it means, right. So somewhere at this point humans still need to come in and make those value judgments. But they need to be somehow formally encoded in the algorithm.

Tavia Gilbert: Using AI to support decision-making might actually accelerate the advancement of human moral judgment and behavior.

Walter Sinnott-Armstrong: I think these computer programs can help us figure out that very deep problem in moral psychology, and thus understand our own moral judgments in a much more profound way than we've ever been able to do so before. So I actually think much to the contrary, that these programs will not reduce our understanding of morality or our tendency to make moral judgments, instead it will improve and enhance our understanding of morality, and also enhance the judgments that we make.

Vincent Conitzer: Imbuing the machine with morality, it's still in a very limited way that we encode the morality into the machine, right, because even after we set these kind of priorities to the algorithm, it still has no real understanding of what it is like to be a person in need of a kidney.

That being said, we're increasingly finding the situations where AI systems have to make these decisions that in our eyes have a significant moral component. So in some sense, we still need human beings to decide what the relevant features are and how much they should be weighed. And because one thing about AI today is that even though we have made tremendous progress, one thing that it does not yet have is a broad understanding of the world. So we humans are very good at having a broad understanding of our world, especially our social world. It's very broad, it's very flexible, it's very integrated, and that is something that we have not yet replicated in AI systems.

Tavia Gilbert: Dr. Schiach Borg anticipates that at times it will be difficult for humans to accept the moral judgment of an intelligent machine.

Jana Schaich Borg: If you have a strong moral conviction or strong emotion associated with the judgment that you intend to make, and all of a sudden, there is something else, this artificial intelligence, that's telling you that's not the

decision you should make, or the decision you feel so strongly about is actually inconsistent with your values. That's-- will cause a lot of cognitive dissonance. Both emotional dissonance and cognitive dissonance. And it will make it very hard I think for the decision maker to actually receive the information that's being given. So figuring out how to overcome that, I think, is going to be one of the biggest challenges. And from a neuroscience point of view I know how strong those circuits are that bring the emotion about and what happens when those emotions and those circuits are so activated they short-circuit in some ways.

One of the biggest contributions is not just in how to imbue a morality in a machine, but in the second step, how do you use that information to impact our moral judgment.

So a very big part of the project is figuring out that second step of how do we translate the AI to humans in a way that the two intelligences can work together and together be better, because there's a big risk that the artificial intelligence will not be integrated into human decision-making at all.

Tavia Gilbert: The relationship between information and moral behavior is Dr. Schaich Borg's particular area of expertise.

Jana Schaich Borg: Most of the time, despite what we'd like to think, we make our moral judgments based on intuition and emotion. When we make, it's not moral judgments as much as the behavior that we then perform, usually that's based on empathy more than other things. And so the more we can either tap into empathy or understand the influence empathy will have, the more we'll be able to understand the behavior you will likely, that will likely follow.

When I started studying moral judgment, everyone kind of believed that it was a very cognitive process. And now people laugh at that. But we still really did. And so we thought, well, if I wanted to change moral judgment then I should understand how we make conscious moral decisions. And as I started to do more and more research, it became more and more clear that only a very small subset of our moral judgments are actually based on things that are at the forefront of our mind. Often it's based on

kind of what we feel in our gut responses, and certainly our moral behavior is much more related to what we feel than what we say.

And so that's why I moved from studying moral principles, which I still do, and it's still important, but-to actually try and understand the motivations that correlate more with actual behavior rather than just what you say. And empathy is one of the things that seems to correlate most with your behavior, your moral behavior.

For me, I'm more interested in trying to understand the relationships between what happens and what we actually do. How do we make better judgments, how do we treat each other better, and honestly how do we not be jerks to each other, especially when it comes to violence?

And for me, because the reason I get out of bed in the morning is to prevent violence and help us treat each other better, I care about the behavior more than the judgment. I also care about the judgment; the behavior is the most important thing and if I care about that, it was clear that the cognition wasn't the best way to get there.

Tavia Gilbert: The team wants to find ways that humans can successfully incorporate the informed analysis of artificial intelligences in difficult decision-making. They're simultaneously working on ways to build in moral, ethical limits to artificial intelligences — limits that will keep a separate intelligence, something you might even call a separate *consciousness*, in check.

Vincent Conitzer: That's not to say that AI systems, even today's AI systems, can't be creative. They actually display some kinds of creativity that within the domain that they're working in, they might find new solutions that none of us have ever thought about. But usually they're restricted to think within a particular domain. And that is something that we haven't quite figured out how to give AI systems this kind of broad and flexible understanding of our world.

Jana Schaich Borg: Because we are making an artificial intelligence with a small 'i,' we're training this intelligence, so it's always going to have the same goal which is, based on one of our applications to predict what your judgment would be. So that goal is never going to change. So it's not going to suddenly have a morality that has a different goal. Where things might get a little

interesting is that the way it gets there might be different than how we make our moral judgments. And so if, you might say that that's a different type of intelligence than ours. But it's not going to be different type of moral intelligence that tries to achieve a different type of goal.

Tavia Gilbert: Still, do they have any real concerns about the future of AI implementation?

Vincent Conitzer: The kinds of decisions that we make now and maybe more generally the frameworks that we come up with for making these kind of decisions, for setting these kind of policies, I think will have a long-term effect and will shape how we make those decisions in the future as well, for applications that maybe right now we cannot even imagine yet.

I think there's a lot of opportunity here to do good for the world. I think AI systems will make human lives better. We have to be very careful and make sure that they don't make human lives worse. I think there are real things to worry about. We can worry about technological unemployment, autonomous weapons systems, large-scale surveillance, these are all issues that tie to AI and that AI could be abused to create a worse world for ourselves. I think in large part, that is up to us as humanity to determine how we're going to use this technology, what uses we will permit and which ones we won't permit, how we design the systems that we do allow out there. This is exactly the right time for society to be thinking about these issues.

Walter Sinnott-Armstrong: So you put the wrong kind of morality in, you get disaster. Isaac Asimov wrote about this a long time ago with his laws of robotics. But if we put the right kind of morality into computers so that it reflects what intelligent and unbiased humans would want to be done, and would think it's the right thing to be done, then I'm not sure what the downsides are.

I first programmed computers in 1971. They've come a long way since then, an amazingly long way. So yes, I'm an optimist about what they'll be able to accomplish in years to come. I'm also a pessimist about them accomplishing everything. I think there are going to be limits. What those limits are, I don't think we know yet, that's what makes it exciting, is that

we don't know. And that's why I think this field is very interesting and important to explore.

Tavia Gilbert: Like any excellent scientist, Dr. Sinnott-Armstrong is comfortable, even cheerful, about not having all the answers.

Walter Sinnott-Armstrong: If morality is put in a machine today it will be different from what happens 100 years from now, because we haven't got it all right yet. But what you have to do is do the best you can with the views that you have and admit that some of them might be wrong. Again, there's no perfection, there's no guarantee, there's no certainty. The trick is to use machines to help us get a little bit better.

Tavia Gilbert: Next week, we bring you the final exploration in our diverse intelligences season, when we meet three researchers affiliated with the Templeton World Charity Foundation's Diverse Intelligences Summer Institute, scientists like Dr. Erica Cartmill, who talks about the plural term "intelligences":

Erica Cartmill: I think by purposefully using the plural term, we're highlighting from the very beginning that this is something that manifests in a multitude of ways, that we're open to the constellation of different kinds of experiences, of ways of knowing, of ways of engaging with the world, whether that's across species, across time points over someone's life, across different modalities, different technologies, different societies across the world.

Tavia Gilbert: We look forward to bringing you that final diverse intelligences episode, and hope you'll continue with us into our upcoming third season, where we'll shift our focus to citizenship.

In the meantime, thanks for listening to today's Story of Impact. If you liked this episode, we'd be thankful if you would take a moment to subscribe, rate and review us wherever you get your podcasts, and if you'd share or recommend this podcast to someone you know. That support helps us reach new audiences. For more stories and videos, please visit storiesofimpact.org.

This has been the Stories of Impact podcast, with Richard Sergay and Tavia Gilbert. This episode written and produced by Talkbox and Tavia Gilbert. Assistant producer Katie Flood. Music by Aleksander Filipiak. Mix and master by Kayla Elrod. Executive Producer Michele Cobb.

The Stories of Impact podcast is generously supported by Templeton World Charity Foundation.