Can artificial intelligence be ethical?

Guests

Dr Jean-Françoise Bonnefon (IAST/TSE), Professor Daniel Chen (IAST/TSE), Rory Cellan-Jones (Host), and Professor Diane Coyle (Bennett Institute for Public Policy).

STARTS

Rory Cellan-Jones

Hello and welcome to Crossing Channels, a podcast collaboration between The Bennett Institute for Public Policy at the University of Cambridge and The Institute for Advanced Study in Toulouse.

This series is all about using the interdisciplinary strengths of both institutions to explore some of the many complex challenges facing our societies.

I'm Rory Cellan-Jones and in today's episode we're going to be talking about: *"Can Artificial Intelligence be ethical?"*

Today we'll discuss the ethics of AI, including why we need to care about it, who is responsible for it, and whether there's a double standard for AI and humans.

To explore these issues today, our first guest is Jean-François Bonnefon from the IAST.

Jean-François, remind us briefly of your main research interests.

Jean-François Bonnefon 0:53

I'm a psychologist and I study how people would like artificial intelligence to behave, especially in situations where AI has to make choices that humans themselves find morally challenging.

Rory Cellan-Jones 1:05

That's going to give us plenty of meat for our discussion, I can tell. Our second guest is Daniel Chen, also from the IAST. Tell us, Daniel, about your research.

Daniel Chen 1:13

So I'm an economist and a lawyer, and currently interested in the promise of machine learning and AI to improve justice systems around the world.

Rory Cellan-Jones 1:21

And our final guest is Diane Coyle, from the Bennett Institute. Diane, what's your focus?

Diane Coyle 1:27

I'm an economist, I've spent many years looking at the digital economy, how to measure it, how it's changing market structures, but also now thinking about the value of data now that AI seems to be becoming so much more widely used.

Rory Cellan-Jones 1:40

Thank you, Diane, in the interest of transparency, I should disclose that Diane and I have actually been married for more than 30 years. But that does not mean I'll be any less diligent in interrogating her position on AI.

So let's get going. Let's start with a few basics why now, the field of Artificial Intelligence has been around for more than 50 years. So why has the debate about ethics really only taken off in the last decade?

Jean-François Bonnefon, do you want to get us started on that? What's happened in the last decade that we're all now talking about ethics?

Jean-François Bonnefon 2:17

My feeling is that it happened organically from the many fields of applications that AI was suddenly deployed in. And so people were suddenly worried about the ethics of AI in transportation or in medicine or in yet other domains. And all these debates coalesced into this field that we know now as AI ethics.

Rory Cellan-Jones 2:37

Daniel, have there been big advances in AI, though, that has suddenly come along? That's been my impression that for years, there was what was called an AI winter. And now we've been in high summer for quite a while.

Daniel Chen 2:48

Yeah, exactly. So I think 50 years ago, there was a lot of interest in AI that was very top down where computer coders would come up with, you know, these are the rules through which we'll want the programme to output certain responses to particular inputs. But now with the advent of things that vastly large amounts of data and much better compute power, people are able to start

to bottom-up generate, what would be the decision support tools that I'm hoping can improve and reduce bias going into the future.

Rory Cellan-Jones 3:22

Diane, you're an economist, some people might be surprised that economists were taking such an interest in this area, they might think: separate field, what have economists got to bring to this? When did you start getting involved and interested in it?

Diane Coyle 3:29

Well, I would say the more it got used by businesses and policymakers because the datasets have become available, as Daniel said, the computer power is there. And people develop these business models that made it possible and profitable to use AI. And so one of the questions becomes, is it getting better or worse public policy decisions? But also, how is it shaping market structures? Is it helping contribute to great concentrations of market power on the part of the big tech companies? And so it's clearly... as it becomes more widespread, reshaping the way we need to think about lots of things.

Rory Cellan-Jones 4:06

Jean-François, were you even thinking about this 10 years ago? What made you start thinking about it?

Jean-François Bonnefon 4:13

Well, I'm a psychologist, and my field is morality, how people judge whether an action is moral or ethical or not. My epiphany was when I looked at self-driving cars and realised that while self-driving cars could avoid many accidents, they would still have accidents, which means that we had to decide as a society the kind of accidents we would allow them to have. And that was my entry point, suddenly realising that everything I knew about how people make these kinds of moral decisions would affect their consumption of self-driving cars and their policy preferences.

Rory Cellan-Jones 4:47

So why does artificial intelligence need to be ethical? Where could it be unethical? There's not a great drive to make an evil sentient AI. So what are the big challenges out there that are being posed? We've already mentioned self-driving cars Jean-François. Diane, where do you see the pinch points here where we need to particularly think about ethics?

Diane Coyle 5:08

I find it really interesting that it's become a debate about ethics, above all else, that it's so prominent in the field of AI, because a lot of my best friends are computer scientists or machine learning folks, and they're not particularly evil. And as you said, they're not out to design evil beings who are going to rule us all. So I think it's because it's crystallising the kinds of things that we didn't pay enough attention to perhaps previously, because if you like AI systems, machine learning systems are like economists or like Mr. Spock in Star Trek, in that you give them an objective and they go about satisfying that objective, meeting it as efficiently as they can. And they do it really quickly. And they got all this data. But what if it goes wrong? What if the data is biased? What if actually, we quite like fudging decisions, because that helped create consensus among people who had some disagreements among each other. It forces if you like a moral

debate, that hasn't been as explicit previously, in thinking about, for example, criminal justice decisions, the kind that Daniel has been working on,

Rory Cellan-Jones 6:20

Daniel, talk us through that you focus on these ethical areas in relation to criminal justice. When did you start thinking AI is beginning to influence my area of research, I need to worry about it?

Daniel Chen 6:33

Well, one way to think about this, as I think I've always been interested in the topic. So I'm trained as a computer scientist, a long time ago before as an economist, as a lawyer.

When I was in law school I was really interested in understanding, you know, people's normative commitments, you know, how do they form these? What are the consequences of these, and in particular, how we can measure these. And so...

Rory Cellan-Jones 6:54

I'm just going to interrupt you there for people like me, who are not quite sure what is meant by normative commitment, what are normative commitments?

Daniel Chen 7:01

What people think is the fair and just thing to do in different circumstances. So at least in a US legal education system, you're often eliciting from the students their perceptions of whether you think the judge's decision was right or wrong, and how it could be improved.

At the same time, the datasets on judges' decisions became more easily accessed, and we could start to understand them or understand them in an applied econometrics fashion.

And so I guess one thing leads to another and we started to think how can we compare the humans who may be not as ethical as you might like, or maybe not cognizant of the potential lapses of ethicality and the data that could potentially support them to make better decisions, decisions that society would broadly feel is more just and unethical?

Diane Coyle 7:50

Criminal justice is a really interesting example. First of all, because it's so consequential in people's lives what the decisions are. Secondly, because it has adopted quite quickly, certain AI tools, but also because it's an area where there are what political science literature calls incompletely theorised agreements. And what that means is that people can agree about certain decisions without agreeing on the basis on which they're being made. So you can think of criminal justice as being either about rehabilitation, you want people not to reoffend or about retribution, you want to punish them and lock them up. But you might nevertheless agree even from very different perspectives that a sentence of three years is right for theft. And if you've got a machine learning system, you need to tell it what its objective is. And even if it's 50% of one and 50% of the other, you've got to, if you like, remove the fudge that's so common in public policy debates.

Rory Cellan-Jones 8:45

Isn't it the case that AI is often just replacing something else, which is itself imperfect? If you think for instance, of a machine learning recruitment system, whereby it potentially imports

human biases? Are we expecting a higher ethical standard from AI, than we went from humans? And is that reasonable? Daniel.

Daniel Chen 9:05

I guess one question I might have is I you think of the advent of toaster ovens or microwaves from a long time ago, you know, we're worried about the potential risks that these new technologies brought about. But at the same time, we had an easy way to think about is it something that's ethical? Or is it you know, the builder of the machine or the user of the machine that can lead to adverse ethical consequences? That would be one way or paradigm to think about the ethics surrounding AI.

The other direction, of course, I think is something you alluded to in one of the earlier questions is “compare with the status quo”, you know, how ethical are the current humans that are making decisions? And I think society should come to a recognition that the current system might not actually be as ethical as one might hope, and then use that as the comparison? And if indeed, we're just saying there's going to be a perfect system to which we're comparing the AI system, then? Yeah, you're right, then there is some degree of differences in standards.

Rory Cellan-Jones 10:06

Jean-François, I mean, there is a danger is there not that we will reject AI when it's actually producing not totally ethical outcomes, but better outcomes than humans are?

Jean-François Bonnefon 10:17

Well yes, we're also very happy not to look under the hood, when a human makes a moral decision, we might think that t a judge or a doctor could make, you know, better ethical decisions that they're currently making. But we cannot reprogram them, even if we're unhappy, we have to rely on them, we could tell them to try to improve, but then we have to rely on them to do it. And so maybe that's why we're not looking too much under the hood, because there is very little prospect of improvement. But with machines, we can actually reprogramme the machines if we're unhappy with them. And with that power comes the responsibility to laying out what exactly we're trying to achieve. So I think that absolutely, we should hold machines to a higher ethical standards than humans. That's the purpose. But that's also the very big difficulty. As Diane said, you know, there's a lot of fudge here that we're happy to, you know, leave alone when humans make decisions. But we cannot have that fudge, when we programme machines

Rory Cellan-Jones 11:13

Isn’t autonomous driving a good example here, we know that human drivers that are messy, dangerous, may choose to speed, whatever the danger to their fellow passengers, but we expect autonomous cars to be a lot better before we're going to accept them.

Jean-François Bonnefon 11:28

That is correct. I mean, mostly, I think, because people themselves completely overestimate their safety as drivers. And so if you tell them that, for example, a self-driving car will eliminate 20% of crashes, compared to the average human driver. Most people think Okay, so that's not relevant for me. Because I'm way more than 20% safer than the average driver. So I will, I will wait for a car to be even better than this.

But this is performance, right? This is not really ethics. I mean, I guess the example here would be to say, look, we think cyclists are dying too much on the roads. And so we want to do something about this. And so we could tell drivers, please be mindful of cyclists give them more space.

We cannot enforce that, really. So we're sort of hopefully telling them please do this. But if we actually programme a car and say the cars have the objective of killing fewer cyclists, then reprogram them to give cyclists more space, but then we have to decide how much space exactly what is the target? What is the objective, how much fewer cyclists do we want dead on the road every year? And how do we solve the trade offs considering the safety of other road users? So these are very, very hard questions that we didn't have to ask as long as only human drivers were on the road. But when machines start to drive, then we have to solve these really difficult questions.

Diane Coyle 12:48

The idea that any government minister is going to say: “Oh, yes, we'll aim for only 100 cycle deaths this year” or whatever the number would be, it's just not going to happen, is it?

Jean-François Bonnefon 12:56

Well, yeah, I maybe I'm a little bit more optimistic and thinking that we could agree on something, like, for the moment like 20% of death on the road are cyclist, but cyclists are only using 10% of the road time. And so the goal is to try to decrease the number of dead cyclists to a number that matches their use of the road, for example, to correct an existing injustice.

Rory Cellan-Jones 13:21

Daniel, have you got to take on this? I mean, obviously, the way the law of the road is framed, is coming under huge examination, as we look forward to an autonomous driving future.

Daniel Chen 13:32

Yeah, I was just thinking about Diane's last comment. Is it feasible to imagine policymakers or society deciding, you know, 100, or 110, or however many?

And how do we very quickly elicit people's preferences on these in real time and thinking about digital democracy tools, but that's a whole other podcast.

The other thought I had was, when you asked us earlier about ethical AI, I would have thought that from a philosophical and also legal perspective, that there's this concept of intent, as well as the consequences. And most of our conversation so far, has been focusing on the consequences and not so much about the intent. And I wonder what the others on this podcast think?

Should we say that? Well, no, there's not really such a thing as ethical AI, from the perspective of the intent.

Rory Cellan-Jones 14:22

It comes back in most cases, not to the AI itself. But to the designers out there. Is that what we're saying?

Diane Coyle 14:28

What is it the designer? Or is it the company that's deploying it? So I guess that merges into where does legal responsibility lie? And I don't know in the autonomous vehicles area where that debate is going.

Rory Cellan-Jones 14:41

Professor Stuart Russell from Berkeley is currently lecturing on living with AI for BBC Radio, he makes the point that the issue is with designing the objectives, the machine, you'll give it an objective, and it will complete that objective, drive you from London to Paris. And if you don't tell it not to run people over, it might run people over on the way. So it feels Jean-François as if it always comes back to the humans rather than the AI?

Jean-François Bonnefon 15:09

Yes, and about responsibility and blame. The worrying thing I'm seeing is that when people hear about these complicated scenarios, where there was a crash, and something was done wrongly with the AI, but also maybe the humans involved in the situations, didn't do really what would be expected from them, either the safety driver in the car or pedestrians, that was it, then people latch on to human responsibility of the human agents involved in the crash, because it's easier, right? That's how we function, we don't have a mental model of what the algorithm of the car is doing. So it's very hard to intuitively hold the car responsible or the AI responsible.

What people seem to do is to immediately latch on to what the human safety driver was doing was that person attentive, what the pedestrian was doing, was this pedestrian cautious and that's a very human reflex, I think. But that will be an obstacle. When we want people to really fully realise the responsibility of the other agents in the chain, the one more invisible, the one we promoted, programmed made decisions about the software of the car.

Diane Coyle 16:14

The point about what you're programming the AI to do there is… reminds me of a debate about something called new public management, which said that people in public service roles like teachers and doctors will do much better if you set them explicit targets. And in this country, in the UK, we went through a period when there were loads of loads of targets. And it turns out that people are highly incentivized to hit their targets. But that doesn't necessarily correspond to what you actually want, you want them to get better. And that might involve not waiting more than four hours on a trolley, but it might involve something else.

So it took out the ability to make those judgments which humans make so very well. But that problem about defining an objective function is a really innately hard one to solve. Because life is complicated. There are lots of things we want to get taken into account.

Rory Cellan-Jones 17:01

Isn't part of the problem that we've come to accept that computers are already more intelligent in many areas than humans. A case in point from the UK, there's been a big scandal about the Post Office, which had introduced a new computer system 20 years ago, which appeared to throw up that lots of the staff the sub postmasters, we're committing fraud, because the computer system came up with figures showing the shortfall, whereas the staff said: "No, there is no shortfall" and we believe the computers.

Is that not an issue that we don't quite understand computer systems and we're going to understand them less and less, but we tend to have excessive faith in their accuracy and their wisdom?

Diane Coyle 17:40

The mistake might have been calling it artificial intelligence. It's not really the same kind of thing as human intelligence is it?

Rory Cellan-Jones 17:46

Jean-François?

Jean-François Bonnefon 17:48

It might be true, but then the antidote, if you will, is that we know that people lose faith in computers much faster than they lose faith in human decision makers. When a machine makes mistake, people tend to jump to the conclusion that if machine makes a mistake, that is a diagnostic of bad coding, and that this mistake will repeat itself over and over.

Rory Cellan-Jones 18:10

You may say that, but it took 20 years for the British Post Office to accept that its computer system was making mistakes. And in the meantime, 700 people were taken to court and some of them were thrown into jail.

Jean-François Bonnefon 18:20

That's not a psychological failure then… that's management failure.

Rory Cellan-Jones 18:25

Daniel?

Daniel Chen 18:26

I was wondering, why don't we build machine learning algorithms that predict the current human decision makers, and then use the historical data to see how much better the humans are relative to the predicted self?

So, Jean Bonnefon was saying: "look, maybe people just trust the humans more, and they give them leeway". But you could also imagine creating a machine learning model of the humans series of decisions like in criminal justice, and then compare, well, if the human is very attentive to things that aren't included in the programme. You know, your machine learning models not that accurate, then yeah, the human could do better than the predicted self.

Daniel Chen

But if instead the human is very, you know, they're hungry before lunch, or they're tired, they're moody, then the predicted self might actually do better than the human self.

Wouldn't that actually create, like, a different angle to which we understand and appreciate the fallibility or infallibility of humans? Yeah…

Rory Cellan-Jones 19:23

Jean-François?

Jean-François Bonnefon 19:24

I think that would work when there's a replacement process where you actually are replacing humans by machines, might not work in situations where machines take up decisions that actually humans find it too difficult to make.

I'm thinking for example of kidney transplants, where you have this huge database of kidney donors and potential recipients and you're trying to maximise the number of successful transplants. And this is just simply not something that a human can do. It's way too complicated. So you have to trust the machines.

But I agree that in other examples, yes, that could be an option. Although I could not see this be done for cars, for example, for driving, I mean, just thinking of the database, you would need to actually simulate what a human driver would do in a given situation that makes my head spin.

Diane Coyle 20:10

Aren't some of the systems being used in contexts where it's not clear that humans are doing worse, I'm thinking of automatic facial recognition, I don't think there have been good enough evaluations about how accurate that is compared to how good humans are at recognising people. And yet, it's being deployed by police forces around the world with, again, quite significant implications for people if they're identified by a computer.

Jean-François Bonnefon 20:33

That's a very good example. Because again, you see that the AI is not really replacing someone, you're just scaling up things that we use facial recognition to check like millions of faces at the same time, and we could never have a human doing that.

Rory Cellan-Jones 20:48

We're talking about individual sort of areas of AI, presumably, we want to work towards something that means we can trust in general, the AI around us, and how would we make that happen? And who should be in charge of doing that? Is this a great United Nations project? Or is it going to happen piecemeal?

Diane Coyle 21:05

One of my principles is that you need to make sure that the interests of those using the system and those affected by it are aligned with each other, and not rushed to use it in places where those interests are in conflict with each other. So I think in a funny way, it's being used most in areas where it should not be like the criminal justice system or like benefit decisions for people on low incomes, because it's so consequential for them if it goes wrong.

Let's make it work better first in areas where everybody wants the same outcome.

Rory Cellan-Jones 21:37

Such as?

Diane Coyle 21:38

I gave the example of fraud detection, medical imaging. We all, our doctors and patients themselves want that work better. And that's a really promising area for the use of AI. So my questions would be about these public policy areas where people are potentially being forced to

do things that are... they're going to not be happy with at all. And I'm trying to think through the implications, the profit motive and other areas where we would worry about ramping up the profit motive using systems in this way.

Rory Cellan-Jones 22:07

Who do you think can be in charge of this Daniel? There has been a debate amongst computer scientists, some might say quite a late debate. There's obviously lawyers, economists and so on. Is this a case for sort of international regulation, local regulation, what?

Daniel Chen 22:23

I think it's certainly something that requires a group consensus and coming to agreement as to what they think would be the general principles that we should be applying in these circumstances. I also agree with what Diane was saying that it's going to be a gradual process. That's the best way to get people to trust in a new system.

The gradual process can be in areas that are more innocuous and there's general agreement, but I guess I'll still quibble again with the idea that even in a criminal justice space where decisions are extremely consequential, that there's no room for decision support when we see that humans can, in fact, be making erroneous consequential decisions.

But yeah, I mean, groups, regulators, international, national. I mean, it's the way many other technological shifts have come to be accepted and regulated across other decades. So I wouldn't be surprised if that's what's going to be needed in this circumstance as well.

Rory Cellan-Jones 23:18

We're in a sort of space race, an AI space race with the United States and China trying to lead in this and huge amounts of money being poured in. Is there any real prospect of people stopping and saying hold on a minute, let's worry about the ethics before we worry about moving ahead too fast. And in particular practitioners in the area, looking at the large amounts of money being thrown at them? Are they really going to pause for thought, Jean-François?

Jean-François Bonnefon 23:43

Maybe the third player here is Europe, you know, and in this race between the US and China, Europe is sort of trying to be the reasonable person at the table, trying to offer perspectives about what should be done and what should not be done. I'm thinking, for example of the recent AI Act from the European Union that suggests that, for example, the use of AI in social credit scoring, which is machines watching you and giving you points, or taking away points to your from your profile, depending on what you do, that this, for example, seems so dangerous, and so potentially harmful, that it should not even be developed. So I mean, who knows what's going to happen, right, but I see here, at least, someone, which is Europe, taking a stance about hitting the pause button, or at least one application of the technology.

Rory Cellan-Jones 24:33

If we look at what people in the field have been doing? Is there evidence that they are beginning to look before they leap? You know, there was a great debate amongst nuclear scientists about the creation of the atom bomb didn't stop it happening. Are we seeing that debate amongst computer scientists about whether just because they can they should? Daniel?

Daniel Chen 27:46

Yeah, I mean, as evidenced by the movement towards embedded ethics, ES, where they're trying to teach philosophical principles whilst also teaching coding. And as that moves to younger and younger ages, that bodes well, for coders who might have a broader objective function than just targeting certain criteria, or maybe a monetary bottom line.

Just the fact that we're having this kind of debate and discussion right now suggests that, you know, we are at a moment where many people are thinking about should we be building this programme? And what are the pros and cons and consequences that you might not have thought about in advance if you developed that programme?

Rory Cellan-Jones 25:29

So if I were to ask each of you what you were most worried about, where you thought the ethical concerns were most urgent, what would that be? Diane?

Diane Coyle 25:38

So I do think it's in areas where there are consequential decisions, and people don't trust the organisation's deploying the AI systems. I mean I completely appreciate what Daniel says about the scope for improvement in criminal justice. But I think that requires people to have trust in the criminal justice system or trust in the government, if the governments are, or in the police if the police are using it.

And I think we're at a time where the AI systems are being rolled out. And yet, trust in many organisations that have power of people's lives is declining, and actually Post Office scandal won't make them any, any more trusting of big powerful computer systems. So it's a trust question. I think as much as it is a question for the ethics of computer scientists themselves, or people designing machine learning systems.

Rory Cellan-Jones 26:28

Jean-François?

Jean-François Bonnefon 26:29

I think it would be trying to predict things that people don't want to disclose, and I have a right not to disclose. The example I had in mind was this model trying to predict sexual preferences from pictures, things like this an algorithm that would try to predict whether someone was gay or not based on their picture? And the thing is that whether these things work or not, it's extremely dangerous. There are many, many places where people have to be very cautious about not disclosing that information, and developing a predictive model that could provide this answer to the authorities again, whether it works or not, that's irrelevant, is extremely dangerous.

Rory Cellan-Jones 27:07

And Daniel, Daniel Chen, what is your most urgent concern, your most urgent, ethical issue around AI that you want addressed?

Daniel Chen 27:16

I would be curious about the… the use of AI in education, just as we have started to see so much the social media debates and how algorithms can contribute to more polarisation or less polarisation. If teaching and education writ large, from a very, very abstract perspective, as

opposed to a sequence of homework exercises that you know that you engage in certain math concepts.

Rory Cellan-Jones 27:40

You're worried that you academics will be replaced by robots, and that will be a bad thing?

Daniel Chen 27:44

I was thinking more... more younger, younger age. As people's normative equipment are being shaped and formed, you know, what happens if an AI is teaching kids what is right or just thing to do?

I don't know... that would be something that society might want to think about how they want to do that or if they want to do that and so forth.

Jean-François Bonnefon 28:03

It's funny because Daniel's fear is actually my hope that someday you know, machines will be able to teach us what is right and what is wrong. And give us clarity about this.

Rory Cellan-Jones 28:16

Well, that seems like a good point on which to wrap up this episode. Hope versus fear in these machines. Thanks to our expert panel Jean-François Bonnefon and Daniel Chen from the IAST and Diane Coyle from the Bennett Institute.

Today, we've discussed the ways that AI can be made ethical, why it needs to be and who's responsible when it isn't.

Let us know what you think of this third edition of Crossing Channels. You can contact us via Twitter: the Bennett Institute is @BennettInst, the Institute for Advanced Study is @IASToulouse, and I am @Ruskin147.

If you enjoyed this episode, then do listen to our other Crossing Channels editions covering topics of Running Government and Incorporating Nature Into The Economy. And please join us next month when we'll have a new edition looking at whether digital technologies can help Africa overcome infrastructure barriers to development.

ENDS