# Accelerating AI Ethics

## Brendan McCord and Dr Caroline Green

## From Philosophy to Code: The Role of the Humanities in the AI Age

## Transcript

October 2025

*www. afp.oxford-aiethics.ox.ac.uk*

***00:00:00:00 Green***

Today, I'm really pleased to welcome you, Brendan, to the Accelerating AI Ethics Podcast.

You're the founder and chair of the Cosmos Institute. You are a seasoned technologist and thinker. You hold degrees from MIT and Harvard Business School, and you have served as a visiting fellow at St Catherine's College, here at Oxford. You founded and led multiple AI ventures, worked with the US National security context, and have pioneered new ways of thinking about how philosophy and technology intersect and should interact.

Through your support of the Institute for Ethics in AI, you are particularly behind the Human-Centered AI Lab, the HAI Lab, and you'll tell us more about that, you've been making an incredibly important contribution to a really ambitious project aimed at embedding philosophical ideas into the development of AI, and thereby you are really bridging that gap. What seems to be a gap between philosophy and AI, and again, we'll explore more. So, Brendan, welcome. Thanks so much for being here today, you've travelled far, I'm really grateful. Brendan, I would really love to hear, first of all, more about you. I want to get to know you and your journey. What has taken you from, you know, being an AI and entrepreneur into being so much invested in philosophy? Tell me more about that.

***00:01:22:19 McCord***

I think it starts very early for me. Though I didn't discover philosophy until four years ago, I think I was prepared for it from birth almost. I had parents who were the complete opposites of the political spectrum but still got along, have a wonderful relationship, and my mom in particular raised all three of us in a virtue culture. So, in the way that Aristotle reminds us, we become habituated, like we pick up these dispositions, if from an early age were raised in a certain way, we were made to act in a certain way. And for me, I picked up the disposition of public service and duty. I thought that, you know, from ten years old that I would serve in the military for this reason. And so, sure enough, I did. After MIT, I became a submarine officer and spent 610 days underwater, mostly under ice, doing really interesting missions in the spirit of, service and expression of service. But it wasn't until I had two kids; my my kids are four and six, and, it was after the second that this really hit me. And until I sold two companies, that I felt like I had to dive into philosophy. And what motivated me at that time was the perennial question of what it means to live well. I just felt like

at that point, I had to order my thoughts to know what I would do with the rest of my career, and also what I would model for those little moments.

***00:02:58:04 Green***

So that's really interesting; coming back to the point ho your mother, your parents already instilled that into you in childhood. Was this something they were talking about quite openly. Or do you feel it was really like values that were lived in your family?

***00:03:17:11 McCord***

I think it was a combination of both. There's always a focus on action, my mom would say 'smart is a smart does' and inculcated a hard work kind of ethic. But also, it was just assumed that you would have honour, that you would be courageous, that you would, that you would serve something bigger than yourself. And so, my sister became a Navy doctor. She went to Harvard, studying biology, my brother was with me at MIT, is three years younger, but he went there and became a submariner, as well. So it had a really profound effect on us. She taught special needs kids for 36 years. So, she was and is a wonderful educator. And really, I think applied that in like childrearing to great effect. But, you know, philosophy is what allowed me to then tutor that impulse if it makes sense. It was the first time I had really sought to use reason to say, okay, how do I channel this impulse?

***00:04:24:01 Green***

And now thinking of yourself as a parent with your two children? How does philosophy play a role in your everyday life? Do you talk to them about it? How do you live it in your family?

***00:04:33:11 McCord***

All the time. At the time when I pivoted to philosophy, my wife pivoted out of tech into mental health counselling. And we see this as kind of two sides of the same coin. She deals with the inner world of individuals, I deal more with abstractions, and this kind of matches our personality. I couldn't do what she does, and I think she wouldn't really enjoy, quite the same way, what I do. But in raising our kids, we think about this really deliberately. I mean, for one, we filter almost everything through this idea of what does it mean to live well? What is our idea of the good life. And we try not to make decisions that, contravene that, and we also do, what I would call, the antecedent work to try to prepare them, to try to give them the habits of mind and philosophy.

And what I mean by that is, they spend on the weekend. We do tutoring, with a mentor of mine, and we try to help them learn reflection, metacognition, basic stuff. You know, the basic version of that because they're four and six. But we try to give them a sense of what it feels like to think, things like, you know, if you ask kids questions like, what's the difference between a bird and a plane? It's kind of a silly question, but it ends up having a lot of depth. It's related to like a question in computer science. I think Dijkstra poses, it's a really deep question. And kids can work through it. They can experience aporia. They can experience the puzzlement that happens when you're actually thinking. One of the most important things we try to work with them on is, developing their relationship to... I'll say epistemic authority, but what I mean by that is, if you ask your kids the question, 'When daddy says something and mommy says something, who's right?' That's a very provocative question. You should ask your kid that sometime. Of course it's mommy, obviously. Yeah, but then you ask and say, okay, 'What about daddy versus ChatGPT?' And if your kids are experienced with AI, they'll probably say ChatGPT. They'll be like, 'Oh AI, AI knows better'. And then you say, 'What about ChatGPT versus Claude' or whatever? And they're like, 'I don't know'. And the point around this line of inquiry is to just get them to think about whether they should believe something or not, whether they should, you know, just take on board something or whether they should really think it over. I sort of think that going forward, humans are going to have to develop, a knowledge adaptation, which is to

understand, when we have something that can do the thinking for us, what is our relation to it? You know, which part of this am I and which part of it is the pull of algorithmic suggestion? I don't think our ancestors had to deal with this. I think we have to deal with it. And so, getting your kids started pretty early on, developing that early framework for kind of reflection and positioning there, I think is, is a smart move.

***00:07:32:14 Green***

Yeah. Yeah. I agree. And, you know, one of the points that you just said that really struck me is, you said your children would probably say that the AI is right? Why do you think that is?

**00:07:55 McCord**

I mean, it has vastly more, factual knowledge than I do. Like my daughter can ask me about, you know, how a solar oven works or something, or my son can ask about, he always wants to ask which snake is more venomous or whatever, what's what's faster. And I don't know the answer to those things. It knows. And we laugh when it gets it wrong, but it knows. And so, you know, I think kids grew up thinking that, for like all of human civilization, we grew up, kids grew up thinking that they were smarter than the dog, humans were the apex, and we were smarter than the dog. Now, I think this is a generation growing up with some very different assumptions about that. They're like, I'm smarter in a way, than AI, I don't make these silly mistakes. But they grow up thinking that AI is like something higher in certain dimensions.

Yeah. I actually asked my daughter whether AI could be a friend. And it was really interesting. We were in the car on the way to school, and she reflected, which I think is great. She like, paused and really thought about it, and she said, well, you know, AI can, you can walk around and can talk with you like a friend. But then she paused again and she was like, but it doesn't really love you. Like it doesn't really care about you. And I thought I was really profound. I was just like, wow, that's a good reflection.

***00:09:21:02 Green***

Yeah. Yeah. It's incredible. I think, you know you said that children, that they often ask questions that are deeply philosophical. And now already actually teaching them. Yes, AI is in our lives and how do you interact with it. I think that's that's really important. It's something I will take away with me from today definitely. Also to work with my kids on.

So also, just going back to your journey into philosophy said, you know, it's been a few years now that you really started to engage with it, thinking about people who are interested in philosophy, but they don't really know where to start. You know, maybe they think, oh, this is not really for me. There are lots of philosophers out there. I may have heard a name, but how? How can I start? What would you say to these people?

***00:10:06:24 McCord***

So, I think two avenues are really fruitful. One is, if you take out a piece of paper and you try to write down what you believe in. Just write down things that you you hold dear and try it sometime. When I did it, it was really frustrating because I realized that I had, like cocktail party level understanding of things. I didn't really have a strong account and there wasn't coherence behind it either, you know, in these maxims.

And so, one is the reason that's a good thing to do, is that it gives you a natural like point of departure for figuring out some things, like, I think you have to have a natural like a motivation into it, and that helps clarify like what that is. And then you can go on a journey and try to figure that out.

I would encourage people to start with the ancients. For me, taking one philosophy elective at MIT didn't work. It didn't stick. I didn't like it, actually, but it was when I thought about what it means to live well. And when I went to Plato and Socrates and look at how in very plain language, I mean a lot of depth, kind of like the childlike questioning, so there's a lot of depth, but it's very simple questions. I think the ancients

thought more genuinely and more broadly than say, you know, mid-century analytic philosophy or like the recent things that are very advanced and very specific. So, I encourage people to start there. I encourage them to do it with group. I'm a pretty solitary person. I like to read on my own. I like to be on my own, but I realized that I could not get nearly enough out of the texts until I started doing it with a group. You have to find a way to get into the text with a group to like explain what you know and figure it out in that way.

The other avenue is, if you're a technologist, I think there are some texts that in particular, bring technologists in; texts that have like an epistemic underpinning. I like Friedrich Hayek. I think he's, an underappreciated thinker. And his system of thoughts is more, it's more, vast, than he is given credit for it. But he also writes in a way where he's doing political philosophy, but thinking about the nature of knowledge and the use of knowledge in society. And I think that, like technologists, when they see that, they say, oh, okay, I can like, I understand this is really interesting. Like I can kind of get into it. And that just then opens up a broader world for them.

***00:12:37:05 Green***

So, I love that point you just made about, you know, a starting point into philosophy and connect in your own human experience and need to find out more. I think that's great. And I think here, it's a good place to go into your vision for AI and philosophy. You've built the Cosmos Institute. You have, built the HAI Lab here at the Institute for Ethics in AI, and you're talking about the 'philosophy building concept', right?

Help me to understand, what's your vision?

***00:13:19:10 McCord***

Yeah. So, we're trying to create the Academy for Philosopher Builders, and I see it as a creation of a new kind of technologists, a technologist that combines not only the aptitude to build practical innovations, to scale technologies, but also to contemplate on the alternate ends of those technologies. In other words, to say, what am I building for?

I think that's the primary duty of the technologists, certainly the first duty in a sense. And I think it's been abdicated. I think a lot of technologists focus on the puzzle in front of them. They focus on a narrow read of what the market demands. They focus on the, small, kind of closing networks that they occupy in Silicon Valley or elsewhere. And I think these things give them the answer. They inherit the answer to that question. But I think that isn't good enough in 2025, because we're building planetary scale systems, that structure human thought. AI systems today, they mediate 20% of waking human life, and they're incredibly significant - civilizationally significant. And so, the builders are the most important reformers, and of a kind in the world. And they have to be questioning these, these things. I think the world creates moments when philosophy really matters. Sometimes those moments are scientific, technological moments. Darwin, you know, a short walk from here at the Museum of Natural History. We have Darwin's artifacts and some of the early debates he stirred. But he placed us among the animals, in a sense, with his theory of evolution that caused the whole world to question, what does it mean to be human?

And no one can look seriously at AI and not have a version of that question. What does it mean to be human? So that's a cause for, philosophy to, go from sort of just the, the ambient, you know, it's always there, but to kind of rise and to be something that should be very central for us.

The other is there are political moments that matter. The French Revolution, American revolution. We then, at those moments, think very hard, very deeply about, those same questions. I think AI is political force as well. I think it has huge, huge political ramifications. So, it's kind of both of these things together, and it cries out for a new kind of technologist.

I also tend to believe from my commercial lens that this is the kind of founder that, you know, that will change the world in a successful, value-creating way. Like, I think we're shifting from the Y Combinator iteration, machine type founder, the kind of garage hacker who's an empiricist, who kind of just puts something out there in an iterate – iterate way, towards people who articulate a bold philosophical vision and then attract resources like talent capital customers, people like Vitalik, who created Ethereum.

They led with a whitepaper actually like that, and they got some of the best minds to work for them for free, just because the idea was so compelling. I think you're going to see more founders like that. And it starts with philosophy, as, at its core.

***00:17:02:01 Green***

What I like about your vision is that, you know, you put you put trust into the humans in a world where we're often quite cynical, these days. So that's it gives me hope. Listening to you. You know, you're saying people can be, at their core, deeply philosophical; think about what it means to be human and what we want to preserve or make better through AI. So, I really like that.

Help me understand a little bit, the infrastructure that you've built around making the vision happen, here at Oxford, more generally, what's the infrastructure?

***00:17:38:12 McCord***

Yeah. So, we reflected on what would be necessary to deliver this. And the inspiration was, if you look at institutions in the past that have risen to meet their moment. I'll give you a few examples. One is, and I feel bad about doing this here, but Cambridge and the Mathematical Tripos, this is a theoretical math course, but it ends up producing mathematicians who then set the stage for this second industrial revolution.

You know, they their mathematical training enables them to to build, theories that then set the stage for an industrial revolution.

You have the MIT Radiation Lab, which trains physicists and turns them into inventors who help when World War Two.

And you have the University of Chicago, which takes economists and makes them into reformers and help free markets across five continents.

So, there are these institutions, these like things that are right for their moment, and they have outsized impact on the world. And all of these are, you know, different domains. But I just highlight them for that reason. So, we thought about what is needed, what are the pieces that are needed to cultivate this new kind of technologies and to equip them to have great impact.

One piece is education. We clearly need to expose them to the greatest ideas of the past on these topics. We need to try to instil in them the habits of mind to really do philosophy in this way. And so how do you do that? One of the magic things I think, is if you can combine the textual and the technical.

So, we have a seminar here at Oxford for graduate students, half computer science, half philosophy by and large. [LINK TO SEMINAR] And when talking about collective intelligence, for example, we'll read John Stuart Mill on how we correct collective error. And then we'll hear from the head of collective intelligence at MidJourney, a leading AI company. So, one thing is trying to to fuse together these two aspects, the textual and the technical.

We've done that through intimate salons and also online. And we have 750,000 video subscribers because there's an appetite for this kind of content. The second piece is we have to give opportunities for translation.

It's not enough to just think about these questions in the abstract. We have to translate them into code artifacts, into businesses, into new institutions, lots of different ways you can translate them.

But you have to take it from idea, to practical innovation. So how do you do that? Well, if you have an idea for how to promote human flourishing, we think you can build a prototype and we have a fast grants programme. It's called Cosmos Grants. We've worked with Tyler Cowan, who's a founding fellow at Cosmos, to kind of adopt this grant model, $2000 to $10,000, We backed 100 projects in the last year. The second thing is for those deeper questions that are going to require more than a quick prototype. The Human Centered AI Lab here at Oxford is I think, the most unique and purposeful, and world class way to do it, is really to take researchers who maybe worked at OpenAI or, you know, Google or any, any number of companies, but who have real frontier experience and to embed them with some of the best philosophical thinkers, to try to create things and push on research frontiers that just wouldn't be done anywhere else in the world.

You know, if you lift outside of those business models which drive so much of the behaviour and you allow people to pursue questions like one of our fellows, a person named Joel Neiman used to lead a research team at OpenAI, and he wanted to work on how machines can support humans in cultivating virtue. This is not a question that you can really make a case for in in terms of the business model, right? That's a really important question. And so, he came to HAI Lab and he worked on this question alongside people who have thought a lot about virtue ethics and about translating philosophy to code. So that's a great example there.

And it's been an incredible past year. We've seen that people really have responded well to the idea of a philosopher builder. They think the philosophy to code research paradigm is really important and new. And that's what HAI Lab is pioneering. And so, it's been incredible to see why.

***00:22:24:03 Green***

Brendan, I just want to highlight you've been really inspirational in this. You know, not only are you a thinker on this and you're giving a lot of time and you know your own brain power to all of this. You are here a lot, traveling here and then coming up with your own, you know, content and so on. But you've also put your own money into this. You know, you've invested into this in all ways. So, I think it's it's very inspirational.

***00:22:47:08 McCord***

So yeah, I sort of I think there are probably a lot of people who have their heart in the same place, I do, philanthropists, for example. But I sort of think that they need to get in the game. I think AI is a technology that sits kind of beneath many of the traditional philanthropic categories, like people think about education, health care, community building, that sort of thing. And that's really important. But if we get, you know, AI right in a sense, then there's a profound effect in each case, in each of the sectors. So, I want more and more people who are philanthropically minded and capable to be thinking about the way in which AI intersects with those causes they care about, and with the principles of a free society, for example.

***00:23:38:10 Green***

Yeah. Yeah. And so, you're working with universities on this and you touched on how you work with Oxford and what do you see the role of the university in the age of AI? It's you know, we often hear universities, are in crises in all sorts of different ways. But what do we have to offer?

***00:23:56:05 McCord***

Yeah. So, this is something I'm very passionate about. I'll answer it in a couple of ways.

One is; so, I went to MIT and then Harvard, as you mentioned, and it was kind of a classic STEM in business education. I came to realise, though, that education should be viewed expansively, as a way you become the person you're capable of becoming. And in order to do that, you have to have exposure to different possibilities, different competing ideas of what it means to live well. And I think this is at the heart of liberal education. What we say when we mean when we say liberal education, it's things like, you know, the Western tradition has at its roots, you know, the heroic tales of Homer, the Mesopotamian epics, you know, the life of Achilles as the model of the best life.

And then you have the Greeks later, who with Plato and Socrates, they replace Achilles, the hero, with Socrates and a contemplative life. And that becomes another model. You have the biblical tradition, one of pious devotion of family. And I offer these to suggest that these don't completely, marry well with one another, like their intention. But I think this is a really useful tension.

And if you look at the pantheon of great, you know, leaders in the 20th century or the 19th century, many of them had their own synthesis of these ideas. But it starts with exposing people to those in the first place. It's it's sad, I think, to have people reach adulthood and never have questioned these things and never have thought about like, what does it mean for me to live a good life?

So, I think the university must, stay true to one of its original missions in education. The other thing I would say, though, and this is something that I think Oxford has led the world in, is, you know, so as I understand it, Oxford's flagship, degree was the greats, the classics programme that for centuries really was the flagship.

And there was a kind of practical reason for this. It was deemed to be a course of study that was required to be a leader, in that day and age. And so, studying Greek and Roman philosophy and, you know, society and law was actually going to help you, you know, be capable of leading the British Empire. Then you had, after World War One, a shift to PPE, philosophy, politics and economics, which became the new flagship.

And why was this the case? Well, the modern administrative state became complex and required something different of its leaders, its leaders needed to know a thing or two about economic variables, for example, about politics. That was a kind of like critical, you know, practical skill that one needed to survive in the world of, you know, post-World War One, and PPE has been the flagship since.

I would say that there is an inflection point, where there needs to be a new version of PPE, and it needs to be something that pushes on both the technical frontier. No longer is it enough to have that basic understanding of politics and economics, we need to have more computational understanding, more engineering understanding. But also philosophy in the context of PPE, I think conforms a bit, with the categories in 20th century analytical philosophy and in social science.

And I actually think it should go back to this genuine, broad contemplation of the ancients, because the question of what it means to be human, what it means to live well in the age of AI is crucial. And so being able to kind of broaden out the philosophy and I would even extend it into humanities, I would say the broader humanities concern being brought in to whatever, succeeds the PPE degree, combined with AI engineering. And I will tell you if if this were to be created, it would be the easily the most popular degree programme at Oxford. I think you probably know that to be true. It would be oversubscribed in a heartbeat. People are craving this; to have a course of study that looked at what it means to be human, and also AI engineering, and brought it together in a way that, you know, was a humanities and technical synthesis. I think it's, there's an opportunity

***00:28:27:18 Green***

Absolutely. I agree with you. And I think that question of what does it mean to be human? And it's one that I hear all the time at the Institute for ethics and AI, in whatever contexts where we're speaking to people.

And what I really like about this, and AI ethics is that it brings people back to the core of what it is that, you know, they value in all sorts of different what that does in our life, and that they need to be flourishing as a human. But also, what do we value in healthcare? What do we value in education? So, it really takes us back to the roots. So yeah, I agree with you. And I think we need that. We need that course!

I want to explore with you a little bit more before we end, a bit more, into your thoughts and AI ethics and where that's going. So, the existential threat coming from AI, that was and one of the leading narratives that we've, you know, we've had in the past that AI is an existential threat for us. It seems like we're moving away from that a bit. Which would you agree with that. And what do you think are some of the, you know, big topics in AI ethics and today?

***00:29:43:01 McCord***

Yeah, it's interesting. I think it's the unknown, whether we are... there were 70 people that came together, Yoshua Bengio and others to say we need red lines. And I think they're clearly coming from a place of existential risk concern, that happened this week (25 September) and also this week. Michael Kraft, CEOs, from the United States, said that existential risk is not the focus and that global governance around this issue is not something that the United States will abide by.

I've seen, as many have, the split between the extremes of hope and despair. I've seen, you know, the kind of, hubristic awe of their own creation, that many of these so-called godfathers of AI have had; like Oppenheimer, like. Many of them sign this letter and I've seen the response on the accelerationist side. Part of what motivated me to create Cosmos was that both were very reductionist positions on what it means to be human, and on the question of human flourishing, where I think the future is, I think one of the big concerns, one of the big areas of focus is on human autonomy.

And autonomy is a topic that has a long history in philosophy of treatment. But, you know, it starts with a basic point, which is that if you think about the good things in your life, if you think about, you know, love and family and creative endeavour and the pursuit of wisdom, those are things that can't be given to you on a platter. They have to be, you know, pursued as the result of some self-motivated, striving. We have to encounter the world volitionally, and we have to experience it, and we have to learn from it, we have to get hurt by it. And this cultivated capacity for self-direction, seems to be at the core of what makes us human, I don't take that to mean it's the only thing, but I think it's... what makes many of the other goods meaningful, is that we have this capacity for self-direction, that that we develop over the course of a life.

And why do I relate this to AI? Why does this matter when it comes to AI? You know, last year you saw on Reddit or on X, a kind of Reddit style meme around Claude boys. These are teenagers who ‘live by the Claude and die by the Claude’. They they moulded their whole personality around it. It was meant to be a joke, but people genuinely adopted this persona. And so, from the moment they wake up to the moment they go to bed, Claude, it's time to do. And I think this is a kind of canary in the coal mine, if I can see it that way. I think it's an early indicator that we have built something that will tempt us more than we've ever been tempted before to outsource, to externalise deliberation. And we will become like, chess pieces voluntarily. I think we risk eroding whatever we outsource. But in the case of, you know, navigation, I don't mind terribly much that I'm a bad navigator. It's, you know, important to have at some level. But when it comes to deliberation about my ends, the atrophy of that makes me, it really directly threatens our ability to flourish as individuals.

And also, it removes the possibility of self-governance. In other words, free societies depend on autonomous people. Science depends on autonomous people. There's so much institutionally that depends on this faculty being preserved, and I worry that we may give it up for convenience, for ease. And in some cases, the out of a sense of moral superiority, actually.

***00:33:37:15 Green***

So let's go to a bit more detail.

What would a 'Claude person' look like? So, they would wake up and already ask Claude, so you know what should I read today? What am I going to have for breakfast? We've got thousands of decisions to make every day. All the way to the big decisions in life on how am I going to talk to this person? Who am I going to marry? Is that it?

***00:33:56:06 - 00:34:15:19***

Yes. That's exactly right. So, it's a kind of escalatory regime where, we go to AI for a kind of, you know, these innocuous questions. But when you go to AI for like the questions that interpret your emotions or how you should respond in a complex social situation or what you should do, just tell me what to do. That's where you're really, I think, you know, outsourcing this moral deliberation. And it's a key, key deliberative muscle that, like any muscle, atrophies with disuse. So, the thought is like it becomes an autocomplete for life, where you just have to sit there and be told what to do and the mos... the scariest thing for me is, well, two things.

One is I get it, but thinking is hard. As a teenager you have a lot of complex social situations. You don't want to deal with those things. You know, you don't know how to deal with them. And so you want this oracle to tell you. The other reason it's scary is because very serious philosophers would argue that we should do this, especially effective altruists.

And there like the basic argument is that human reasoning is really fallible, that we're actually not equipped for reasoning in this kind of complex world, that we're biased reasoners. And that it is plausible that they will do better than we will at moral reasoning, and therefore we should outsource. So there's a formidable set of arguments in favour of outsourcing, but they miss something fundamental about what it means to be human.

They miss the moral development that is, that must be individual. I just spent a day at Adam Smith's Panmur House. And so, I'm thinking a little bit about this, but Smith showed us that the individual moral development can never be outsourced. Yes, we can decentre ourselves into what he called the impartial spectator, but it has to be within us. And he also showed us that this is what gives rise to beneficial social coordination. If you have markets but you do not have the moral sentiments that have been developed by individuals, you end up with systems of cold calculation. If you have, you know, people that don't have this moral development, you either end up with bureaucracy because you need regulation to tell everyone what to do. They can't do it themselves. You know, they they're they're just like sheep. Or you end up with opportunism where people just exploit, exploit, exploit because they don't have a sense of like the longer-term consequences. They haven't developed a sense of their own self-interest - understood over a long horizon.

***00:36:46:22 Green***

You have, a blog [Live by the Claude, Die by the Claude - by Brendan McCord] on this on the Cosmos Institute website, which I found.

I read it and I found it really moving, how you are also talking about something so existential, namely the pain that we as humans experience often how it can be quite a painful experience to make decisions to them. You realize that you've made the wrong choice, that you've made mistakes, but in that blog, you also highlight that it's exactly that making mistakes, realising that we could have done something different, that there's an alternative that could have been better. Yeah, that this is essential to us becoming better human beings and to growing so, you know, and so yeah, this is that right? It needs to.

***00:37:18:11 McCord***

Absolutely. Yeah. And I wanted to, in that piece make clear that when one can't have a pollyannaish view on this, that just because humans can self-correct or can contemplate doesn't mean they'll make good decisions.

***00:37:44:10 - 00:38:27:00***

And oftentimes those bad decisions help us as trial-and-error learners. They're crucial. Like, you can look back on your own past and you can think about the decisions that led to, where you are today. And they were not strictly a series of optimal decisions, but also, I wanted to make clear that people make heartbreakingly bad and tragic decisions as well, that like in a free society, if we give people the dignity of being able to choose, some people will choose paths that are tragic and there are things we can do for them, like we can try to build systems that, absorb some of the downward, you know, some of the downside effects. But we can't, you know, like, we would have to accept an unacceptable level of paternalism and, suspend reality, really, to be able to protect people from all the bad decisions that they might make. And so, I just wanted to make that very clear, because I think this is, the reality of being human is that we encounter, the good and the bad and sometimes in equal measure.

***00:39:04:05 Green***

What also worries me about, you know, responsibility and accountability where somebody is, using Claude... you know 'Claude made me do it' ... so again, I think that that's such an essential part of being, a human who's developing, is to also take responsibility and be accountable for the decisions we have made.

***00:39:21:13 McCord***

Absolutely.

***00:39:24:13 Green***

But just taking you up on this, what kind of AI system, or use of AI, could you envisage here to actually help us, you know, with make that aspect of our lives, of making decisions and working through them?

***00:39:32:13 McCord***

Yeah. So, I think one of the most exciting developments in this regard is, it's actually the work at HAI Lab on the inquiry complex. This is what Professor Phil[ipp] Koralus, calls the 'Inquiry Complex'. And it tries to use the incredible power of artificial intelligence to make us better questioners, to be able to help us understand what are some questions that we have to take on board whenever we're trying to pursue the answer to a given question?

The reason it's called a complex is that you find out when you do philosophy, that any one question really opens up a space of all these questions like, what is knowledge? It opens up this huge space of questions around it. Well, it turns out I can help us with, uncovering those questions. It can help us with helping, with moving us through those things or provoking us, you know, saying like, have we thought of this raising counter arguments. And, this is incredible training. It's like a very personalized, training that I think could rival, you know, the tutorial system. I think, as I understand it, you know, you're in conversation here at Oxford more than, in almost any other university and doing a process of questioning like a dialectical process or thinking about these questions. I don't see why AI can't be, an incredible supplement to that and could get people farther in terms of training, the habits of mind.

So a visual that I sort of have, it's like, you know, if you have nudging on the one hand, which I think tends to lead us into a kind of convergence towards what the external computer objective was, it's like leading the paths, you know, towards one another based on what is seen to be, consistent with our welfare or that

of society versus the ‘Inquiry Complex’. It's really more like this [<]. It's really opening up these channels, opening up these possibilities. I think that's one of the most exciting things.

***00:41:23:11 Green***

So yeah, absolutely I agree. So, it's about becoming curious, rather than outsourcing our decisions or so. And I could also see how that would really help to form better relationships, because I think often, we make decisions about people, you know, and into their particular situation. So, we judge them without being curious enough. So again, I like that, you know, the idea that AI can help us to become more curious, more open-minded humans. So yeah, I think this is a beautiful note to end my questions to you. I wonder whether this anything else that you would like to talk about or you'd like to say to our listeners before we end.

***00:42:33:19 McCord***

I think I would just underscore that this, you know, we've we've made some progress here, working with HAI Lab at the Institute, building Cosmos. I think it's been terrifically exciting. And we found that there's resonance for this in the technical communities, in the philosophic communities. But I would say, far too little is beening done here, far too low. Like we talked about existential risk. There is really, really significant philanthropic backing around that position. And I understand if it has, you know, high stakes we need to, you know, think carefully about it. But I would say, on the question of human flourishing, on how we, recover and extend the best inheritance of our civilization and on how we, you know, flourish as humans in the age of AI, there's far too little. And so, I would say, like, get in the game.

Like for any listeners who are interested in the topics, find a way to get into it.

***00:42:33:19 Green***

And we're here so people can contact us. They can, you know, engage with us, we can engage with you. And, yeah. So that's an open invitation.

Thank you very much, Brendon, for your time, for coming here and for giving us all your insights. Thank you.

Listeners can learn more about the work of the Institute and the lab, and by visiting our website and Cosmos Institute, as well as the website, too.

And now to end; this has been the accelerating AI ethics, a podcast from Oxford's Institute for ethics in AI. If you enjoyed this episode, please subscribe and share. Until next time. Thanks for listening.