AI and You

Transcript

Guest: Charles Radclyffe, part 1

Episode 57

First Aired: Monday, July 19, 2021

Welcome to episode 57! And before we get to today's guest, I'll just do a quick catch-up for new listeners, because the show is growing and we are getting more listeners all the time. You've heard our three guiding questions in the opening, but what does that mean? First, we're on a journey to understand just what AI is. It's not a journey I ever expect to complete, because AI covers so much ground, is such a fundamental shift in our collective experience, and is so reflective of the human condition, that we'll be picking apart the distinctions of AI for a century. I want you to get an idea of the breadth of what AI spans, to understand it from a kind of gestalt view from looking at it from so many angles. I love the challenges of coming up with metaphors and mental models that help you understand this, boiling the complexity down to something digestible. And I'm really grateful to all our terrific guests who give their time to help you understand AI.

Second, I want you to get why it's a big deal, why this is not just some fad that's grabbed headlines like Cabbage Patch Dolls or k-Pop. Andrew Ng says that AI is the "new electricity." Kevin Kelly says that the business model for the next 10,000 startups will be to "Take X and add AI." If those and many similar assertions are even close to accurate then we're at the outset of a shift in the human experience that will outstrip the Industrial Revolution in the rate of change. That's a shift that's worth getting ahead of, because getting behind could be very painful.

Third, we want to look at what do you do to get ahead instead of being left behind, and that applies in both your personal and professional lives. We've not done that very directly so far because solutions tend to be narrow and specific to a particular sector or profession, but we'll definitely get into it more, and in fact you'll hear some advice on deploying AI from today's guest. If you have particular questions about leveraging AI or being resilient to AI disruption at any level from individual to global, then send them in. I talked about this in general terms in my 2020 TEDx talk and believe me, it was very hard work to create a message that had relevance and usefulness to a completely general audience and fit in 13 minutes. Whereas when I'm talking to people in a particular company or in a particular profession, I can make that content highly targeted with a clear return on investment. But I really *enjoy* talking to general audiences like this. Just let me know you're out there from time to time, with a question or comment. And especially by rating the show.

Now to today's guest. I met Charles Radclyffe when we were both speaking – virtually – at British Telecom's AI Festival in February 2021. Until recently he was head of AI at Fidelity International in the UK and then founded EthicsGrade, which inhabits a space now known by the initials ESG – Environmental, Social, and Governance, and EthicsGrade scores companies on their AI governance. Just as an example, they give Skype an A and AirBnB a D. So I invited him to come and talk about the ethics of AI, including bias, the challenges that companies exploring AI face in that, and how they can handle ethics better. Here we go.

Charles, welcome to the show.

Thank you, Peter. Nice to meet you.

You have done a lot here with ethics, in particular of artificial intelligence, and that topic of the bias in artificial intelligence really fascinates me and it is so important and real right now. Do we even have a good handle on what bias resulting from artificial intelligence deployment is and what to do about it?

In a word, no. But I think you want probably a slightly longer answer than that… So, I mean, I think that there's good bias and there's bad bias and I guess what we're talking about here is unintended bias, because essentially a lot of the kind of fancy mathematics that goes into making machine learning so powerful, relies on bias. I mean it's the very nature of what we're doing is, we're exploring and exploiting the mathematical properties of bias. So, I think to talk about a bias free data set or a bias free model is a bit of a misnomer. So, I guess that's my kind of starting point, I guess the challenge really comes around unintended consequences and when we see racist chat bots so or the CV scanning tool that Amazon used, these things when they hit their headlines, they do so not because the project team or the ML engineers involved were inherently bad people or racist or biased or discriminatory in any way, though, they happened because the right engineering controls were not in place and I guess, where I come at this and it's probably an unfashionable statement to make and it probably will leave me in hot water after this show. But, I don't see bias and discrimination as ethical challenges; I guess that's my kind of starting point. I see much of the downsides that come as a result of these things being caused by engineering practice and I think those are kind of comforting to the machine learning community that they can solve for these problems with standard controls process, and maybe a bit more technology. That's not to say there's no ethics challenges with AI; but I don't think bias is necessarily the one that I would focus on.

Understood. We will get to what are the ethical challenges then, but I don't want to give up this question of bias so quickly because there's so much meat on the bone; and I agree with you that it's an issue of hygiene: clean data - well not just clean data but clean processes. You could have data that's 100% accurate and it's still not producing the result that you want and it could be completely representative. So an example right now would be how data or algorithms, when asked to describe a doctor, typically do it in terms of a male and asked to describe a nurse, typically do it in terms of a female, and we go… well that's biased, that's not reflecting what we want; but it is real because that is what the data is right now: what we want is something different. So, we are in this process of trying to move to a better, more equitable, fairer distribution of those roles but we don't have it yet. So, all our data is in a state that's inherently biased. Now what do you do in a case like that when the best, most accurate data you have, is still not going to do what you want?

Yeah. So, I think the two things in this question, one is that choice we're making, that objective we're optimizing for a more equal, more fair, more balanced society, that's the ethical choice

that we've made and I'm certainly somebody who would vehemently support those notions. But I think it's fair to say that those are not universally held ethical standards and so I think we need to be conscious that is the ethical choice. In terms of what you're describing, I mean this happens, I guess it's very uncomfortable, it happens too much still but I think the key thing is really around so engineering controls it's about understanding the data that you're training a model on. Understanding that data, the bias already in that data, which may be kind of human biases, you describe language models, typically have a lot of that sort of human bias because of preconceived ideas, data comes from a generation gone before, etc. And you need to understand that and you need to make sure that when you're training the model and deploying it in the field, you're not accelerating or amplifying those qualities in the data set and I guess that's the challenge, in machine learning or a wooden stick, both these pieces of technology are forms of leverage. And the downside to leverage, as we know from the financial markets, is when it goes wrong, it can go wrong quite spectacularly. The way you solve for this is you need to make sure that, there's the good news is I guess we've been aware of these things now for a number of years and they're really… I would say, two disciplines and the European, the new European regulations on AI, point to these two things. One is essentially quality management in terms of the lifecycle of your data and the life cycle of your models. And the second is risk management. Risk management matters because obviously there are some use cases where discriminatory algorithmic systems are very, very toxic and the perhaps other situations where it's not so bad and I think we have to recognize that not all aspects of discriminatory algorithms have such negative consequences, but the ones that do we want to make sure that those are tightly managed.

I've got a lot of questions here. First, when you talk about risk management, what type of risk are we managing?

So, the way this works in the financial markets and certainly my experience comes from looking at risk management from a trading perspective, from sort of the algorithmic systems that you use in trading. You are looking at what is the likelihood of the negative events. Firstly what is a negative event? What is the likelihood of it happening? And then what is the severity of that risk? Those are the kind of the first questions you need to look for, and then if the answers to those questions mean that you're above a certain risk threshold and you'd really expect that the mitigating actions are then identified and then followed through. So what's probably most disturbing, right now in 2021 is that a lot of organizations that are experimenting with machine learning capability either because they're buying it in or they're developing it in-house and they don't necessarily know where it is. That's kind of really the first step, they don't have that inventory in place, to really be able to say; this is the extent of machine learning capability in our organization and therefore they can't control those risk levers, they can't centrally manage that risk register and this is the key problem and so when things have gone wrong, like the amazon, CV scanning bot, or Microsoft word Tay etc, the principles that organizations have, the goals that an organization has in terms of the outcomes it wants to optimize towards, don't trickle down to what's happening on the ground and that's largely because there's a disconnect

between those two things. The way you solve for that is you put some basic disciplines in place, such as risk management. Quality management is more of an operational thing but risk management; I think is a key, the key leadership discipline that needs to be instilled first.

It sounds like in that kind of organization, it sounds more like either a cultural issue, or a governance issue, or a leadership issue that you don't have the two-way communication that you need, and, that's an interesting take on it. So, your risk management is risk in the broadest sense, any kind of consequences and what are the cost and the likelihood. So, some of those could be good, and you mentioned good bias, I don't know if that fits in the same category, but talking about unintended consequences here, then, are we pioneering here with AI, all kinds of unintended consequences?

I think that this is the point; I think that just taking a maybe a 40,000 foot view of this. The challenge to organizations getting this wrong, is really a challenge of reputation in many respects, when these things go on. I mean when we think about the kind of the big scandals in the tech industry over the last 10 years, we're probably… there's a list of five things that we're probably both agree on, as the most impactful and they ultimately will scar the reputations of the organizations involved for potentially a very long time, much in the way that people still talk about Ford and the Pinto, 40, 50 years later. So, what organizations need to think about are really how might this go wrong and then make sure that they've put in place the right mitigating controls and I think one of the simplest ways of doing there's two simple ways of doing this. I think first of all, the discipline of red teaming, which has been used quite extensively in the military. I think it's a fabulous methodology for thinking outside of the box in terms of unintended consequences and be happy to explore that further with you and I think the other thing is around stakeholder engagement and I think that's the other aspect of this which is… yeah that's where I would argue that the ethical conversation really lies, is all the stakeholders involved in a project consulted in terms of where they see the impact, where they see the consequences, is that feedback loop built back into the design of the system. And sometimes when that does happen, it happens at the end of the life cycle. It happens once the system has been designed, built, has a focus group and then it's deployed and I would argue it needs to happen, maybe a little bit earlier and maybe a little bit more throughout the life cycle, but I think the key thing is the red team, that's what we'll catch most.

And we'll get to that. Talking about the stakeholder engagement and you use the keyword lifecycle there, stakeholder engagement is an issue that we've been dealing with for decades, that's independent of AI and really agile methodologies that require or enforce this stakeholder engagement and close communication, are a key to the success of projects that have some risk, so that makes me think that, that's the same issue rearing its head again just in the context of AI. But, let's talk about red teaming. For those of our listeners who maybe are not familiar with the term. Can you explain that?

Yeah, absolutely! So, I think the term red teaming comes from… I think it's the sort of Prussian, one of the Prussian wars. So, it was talking like going back into themselves 200 years ago and I

think what they were doing was, they were doing some military scenario planning, and the Prussian army wore a blue uniform, and so in their kind of scenario planning the enemy was deemed to be red, and so what they did is, they divided their group into two and they said you are going to be the enemy, you can wear the red costume and that's the red team. So, I think that's where the origin of the term comes from. As I understand it was used in the Israeli military, in the sort of mid-20th century and then adopted by the United States, as really a sort of military scenario challenge mechanism. I think the way that a red team is most effective is when you've got a bunch of people who are tasked with challenging the assumptions that are put forward in any decision and considering how might this go wrong, where and in which situations may this play out in a bad way and I think it takes a special type of person to be able to do that. Yeah. I think you've got to be a bit of a lateral thinker, to be able to do that and I think importantly there needs to be a sense of working outside of the hierarchy, because one of the essential things that a red team needs to do is challenge the hierarchy, and to be able to have the empowerment to be able to say to the boss, "Your pet idea may not be right," and we need to challenge that. So, I think those are some of the important qualities of red teaming. I mean my firm, we run red teaming ourselves, we evaluate companies, we rate them, we take a view on companies, and we make sure that's a consensus held decision and we challenge, we play that sort of devil's advocate position. And really that's the kind of place where red teaming is known by a different name, is actually from the Catholic Church, where we're all familiar with this idea of Devil's Advocate. Essentially when a person is considered for sainthood, there's two people appointed; one person to argue the case: Why should this person be made a saint? And then the other person is known as the devil's advocate. Don't ask me the Latin name, I can't remember, and the devil's advocate is there to argue all the reasons why that person should not be a saint and it's basically a very traditional western kind of clash between one and the other, confrontation between those points of views. I guess to be the devil's advocate, for certain people might be quite hard, if they've led a very virtuous life but nevertheless that sort of prosecution service is very important and really that's the job of reading an AI context is to prosecute the idea and challenge where it might go wrong and make it stronger as a result.

And that's really like the antidote to groupthink and Yes Men and Women: people who feel that they don't have the freedom to contradict the bus or to rain on the parade of a project or say that there's something wrong with it because they fear of being accused of not being a team player or something like that. If you have a red team, their *job* is to say no, that's their job the more problems they find, the more they're doing their job. So of course, if the culture is already accepting of the idea of a red team in the first place, it's advanced beyond most of that mentality. So, as you say that is key to working out these kinds of issues in a corporate context, but let's get on to the ethical issues more now, and talk about then, if bias is not so much a deliberately ethical issue, which ones are?

So, my starting point to answer that would be to explain why conflating bias and discrimination and those questions with a view of ethics is dangerous and limited, because I think a lot of people, particularly from an engineering background are comforted by the idea that

discriminational ethics is biased, because those challenges can be overcome with process standards, tooling, methods, diligence which are good engineering qualities. I guess what makes engineers uncomfortable is the topic of ethics in its broadest sense, because really that discipline requires, I think more of a social sciences background than most engineering folk have. The reason I think it's important to have I would say, a maximalist view towards ethics, is because if you don't, then you're running the risk of leaving yourself, open to attack from a place where you haven't potentially mitigated yourself fully and so I was thinking, I've been thinking many years for this this question, I used to run a data analytics company, and I was finding that I was getting very excited about the opportunities that we were involving ourselves with; we had clients who had data, and we had engineers who were deeply skilled at being able to find the opportunities to answer questions with that data, and me and my love of technology was also very exciting but also concerned that sometimes we were crossing a creepy line, and what I think concerned me the most was that, I might have been the only person in the room that spotted that. And this troubled me and I came to realize that one of the reasons why I was identifying what I was uncomfortable with - I'm not saying that my creepy line is necessarily the same as yours or somebody else's - but the fact that I felt, I was able to articulate this and understand these things probably comes from my own background even though I've been in the technology industry in my life, I'm a lawyer by training and it was really legal philosophy, which was my favorite aspect of that. So, I've been thinking about this question of, and I've been slowly watching us collectively falling out of love with technology over the last decade and it's pained me because I've been to conferences where people have talked about the amazing capabilities of big data analytics, etc and then you have this kind of grenade thrown, at the stage to the speaker or the panel, what about the ethics? What about those ethical considerations? And I guess the dismay I've had is, seeing that momentum build, and seeing - even to now - those questions being brushed off, as a question of, oh there'll be regulatory answers to this, or, "We're fully compliant with GDPR," or "That's a really important question. Thank you for raising it! It's a really important question for somebody else, at a later point, once we've figured out the tech and the engineering," and for me that's really where the gap lies and I think the answer, this is going to sound very wishy-washy, but I think what you said earlier was very true that these are not new issues, they're just new to the tech industry. If you look at other contexts, this idea of stakeholder engagement operates quite effectively, in fact and really that's what's missing. So, to me ethics is really about that conversation, it's an activity, and it's about that conversation with stakeholders about understanding what their concerns are, helping them understand maybe where that creepy line is for themselves, and they're making sure that if you are going to cross that creepy line you do so in a very deliberate way, a very intentional way and in a very communicative way and I think most people would be comfortable with that and I think the instances, where we've seen big ethical challenges happen to us or imposed on us, I think the one, where we feel most uncomfortable or dismayed by those ethical choices are where we don't feel that we've been consulted. Not necessarily because we disagree with the outcome. I think most people would agree with outcomes which they don't agree with, but at least if they had a shot at being able to voice their view.

So, let's try and drill down into this a bit because ethics is such a big thing and I want to get a handle on some of it. First of all, let's just draw a distinction. In your view is ethics something that only applies to people? In other words if an AI is behaving unethically, should we use a different word for that or does it get us into trouble to talk about an AI being unethical. What's your viewpoint?

Yeah. So, I think this is where philosophy can be unhelpful and that's what we do I think in philosophy is, we play word games with each other, to borrow a line from 20th century philosopher Wittgenstein and so I think we do have to be… we do have to find some precision in our language, and it's important for philosophy as a discipline to do that, much in the same way it's important when you're writing computer code to be precise with your language and your grammar otherwise the system simply won't work. So, to me ethical choice is only something that we as humans - but humans that are experiencing the world - can truly make and the reason for that is because I'm talking to you, now this is the first time we've spoken, we've exchanged a few emails, this conversation gives me a sense of you as an individual and because I'm an individual, I can relate to what it might be for you, as somebody on the other end of this conversation. Obviously, I don't know you deeply, and may never do, and there's always a limit to how much you can know any person, but because I'm a person, I can understand certain things and therefore if something were to happen to you, if you were to experience a situation, I can place myself in that and I can appreciate it, and therefore I can really process the ethics of that. Of course! There are people who can't do that, and we call them psychopaths and I think if we just say that ethics is something that all humans can process it, it's true but for that special class of people, but for a machine I don't think it makes any sense to talk about a machine making an ethical choice or machine being unethical. Although that is lazy shorthand we make, it's shorthand I make. In fact, on my website I deliberately make that shorthand because I'm trying to be provocative but I think it's important that we know that we're doing that when we are.

And because we like to anthropomorphize our machines, particularly if we're proud of having created them. Let's look at some of the ways then that ethics shows up in the context of AI. And the one that immediately comes to mind is abuse of privacy. So here, unethical behavior would be someone in a company, or the company, using artificial intelligence to violate privacy of people. The poster child for this is Facebook, whether you agree that they've done that or not, they certainly are accused of it, and is that an issue that you work with clients on or what do you find are the tall tent poles, the big issues surrounding managing personal data, when AI can process it?

Yeah. I mean this is a really great question. So, I think… again I just want to say the context that I think, I very much take the view that ethics is a maximalist topic, it can be anything you want it to be, and I mean that in terms of it can be anything you want it to be and somebody else might have a conversation with us and say well it also means this other thing and I think we have to accept that if they are affected by that, if that's something which moves them as a person,

changes their intersubjective relationship with us, and then indeed it is an ethical choice. But, in many cases, we are talking about kind of fairly narrow bounds, such as data privacy and I guess where this gets blurred and confused and this is why quite often it gets conflated with regulatory compliance, is we are starting to see laws around what data can and cannot be used in what context and so, I think the question here really is, has an organization or an individual or a project team - I say organization in a sense that there is a sort of joined-up thinking from the senior leadership - but has an organization understood and thought through and considered the ethical impact of a particular project. I think there's sort of really two components to that to come back to what we were saying earlier, if you're not going out to a wide group of people and asking that question, then a red team can help you do that as a kind of shortcut, but really the better answer is to consider your stakeholders and to go and ask them, that can be a… I'll just say focus group because I think focus groups have a kind of reputation for being a bit airy-fairy, fluffy and meaning meaningless. But I think it's that sort of notion that you want in a controlled way, to understand a representative sample of the stakeholders that are involved and then understand how they feel about a certain thing. Now of course you mentioned Facebook and one of the interesting things about Facebook and I remember last summer, I was watching a talk from Noah Feldman, who was one of the principal architects of Facebook's oversight board, which is a quite an ambitious idea, a fabulous idea, in many respects and so limited in others. What he described was a sort of challenge that Facebook has as an organization, in that they have nearly three billion users, and he said the problem - also paraphrased, what I remember from his talk - he said the problem is if you can't go and ask three billion people what they feel about something and also you can't just ask anyone, these sort of questions and he says you, if you do, you end up with a sort of Boaty McBoatface type answers, and this referring to that ship that us Brits tried to name through a kind of crowdsourced answer and you ask you ask an open-ended question, you can get quite easily a silly answer and that answer can quite easily go viral and be uploaded by many others and you end up in a very strange place, you didn't set out to intend. You can't do these things in an open-ended way, you have to constrain the conversation, you have to limit it, and you have to get it to tend towards decision and outcome. But, I think just because you have three billion users, just because you've been amazingly successful as an organization, doesn't get you off the hook, and I think that's the challenge with some of the big organizations that are exploiting machine learning capabilities, AI more broadly, has had such an impact on our lives is that they are so big, that actually for them managing ethics in a reasonable and responsible way, isn't possible and I think for that reason alone, we need to rethink some of these structures we've got today.

Okay, that's the end of part 1 of the interview, we're breaking it into two parts to keep the episode sizes more manageable.

We were talking about red teaming there, and I think of that as a good example of how a learning organization expresses itself. "Learning organization" is what you get when an organization advances beyond the primitive reactive stages – not all of them do – then it looks at its own practices. It goes

meta, if you like. It asks the question, “How can we improve what we do and how we do it?” To set up a Red Team means that an organization has already realized that its normal processes may lead to blind spots, that people are naturally incentivized to go with the nominal plan and not rock the boat. To be a contrarian is risky, and a career limiting move in organizations that don’t understand this side of their culture and psychology. The ones that do are mature enough to realize that you can’t just declare that all opinions are going to be respected, that you have to set up a structure where contrary views are not just okay but required.

Why do I mention this? Because it’s one of the essential traits of any organization that wants to make it through the revolution in AI and other exponential technologies. I’ve just been preparing a training for executive coaches on what those traits are and how to work with an organization to grow them.

I also liked how the conversation about ethics got into the philosophical territory of the shared backgrounds that human conversation take place against, starting with each of us knowing that the other is human. That resonates deeply with my Neurolinguistic Programming training. And it makes me wonder to what extent AI has to understand shared backgrounds to have conversations with us. And what would we think the background of the AI is?

Charles also has a podcast about AI, called “Are You a Robot?” and I had a fascinating interview with their excellent host Demetrios Brinkman. It will be Season 9, Episode 5, and it will come out on August 30, 2021. There’s a link in the show notes and transcript.

In today’s news ripped from the headlines about AI, Agility Robotics has a robot called Digit that to me looks a bit like Tom Servo from Mystery Science Theater 3000, at least from the chest up, and it’s a bipedal robot for general purpose use. It has the backwards knees that are like a robotic goat, and it’s targeted as an apparent competition to Boston Dynamics. It’s a bit smaller than their Atlas robot, and it’s designed for things like picking up and delivering packages. Ford bought a couple of Digits for doing that, Agility is in Oregon, so on the other coast from Boston Dynamics. It can roll up in a hatchback, hop out of the back, and take a package to the front door. Digit will set you back a cool quarter of a million dollars, although this is supposed to drop to 70k at scale.

Next week we’ll finish the interview with Charles, talking more about ethics, the obligations of developers, and segueing to the future of jobs and Charles’ motivations in addressing that topic and his personal experience in putting that message out there in a TEDx. That’s next week on *AI and You.*

Until then, remember: no matter how much computers learn how to do, it’s how we come together as *humans* that matters.

http://aiandyou.net