DATA ANALYTICS MARKETING MAR 4, 2019

How to Build Artificial Intelligence that Everyone Can Trust

Experts from IBM Watson and Kellogg discuss how to remove bias and increase transparency in machine-learning algorithms.

BASED ON INSIGHTS FROM

Florian Zettelmeyer
Inhi Cho Suh

Artificial intelligence is here to stay. Machines are getting smarter, faster, and are poised to play ever-greater roles in our healthcare, our education, our decision-making, our businesses, our news, and our governments.

Humans stand to gain from AI in a number of ways. But AI also has the potential to replicate or exacerbate long-standing biases. As machine learning has matured beyond simpler task-based algorithms, it has come to rely more heavily on deep-learning architectures that pick up on relationships that no human could see or predict. These algorithms can be extraordinarily powerful, but they are also "black boxes" where the inputs and the outputs may be visible, but how exactly the two are related is not transparent.

Given the algorithms' very complexity, bias can creep into their outputs without their designers intending it to, or without them even knowing the bias is there. So perhaps it is unsurprising that many people are wary of the power vested in machine-learning algorithms.

Inhi Cho Suh, General Manager, IBM Watson Customer Engagement, and Florian Zettelmeyer, a professor of marketing at Kellogg and chair of the school's marketing department, are both invested in understanding how deep-learning algorithms can identify, account for, and reduce bias.

The pair discuss the social and ethical challenges machine learning poses, as well as the more general question of how developers and companies can go about building AI that is transparent, fair, and socially responsible.

This interview has been edited for length and clarity.

Florian ZETTELMEYER: So, let me kick it off with one example of bias in algorithms, which is the quality of face recognition. The subjects used to train the algorithm are vastly more likely to be nonminority than members of minorities. So as a result of that, the quality of facial recognition turns out to be better if you happen to look more conventionally Western than if you have some other ethnicity.

Inhi Cho SUH: Yes, that's one example of a bias because of a lack of data. Another really good example of this bias is in loan approval. If you look at the financial-services sector, there are fewer women-owned businesses. So therefore you may have loans being arbitrarily denied rather than approved because the lack of sufficient data adds too much uncertainty.

ZETTELMEYER: You don't want to approve a loan unless you have some level of certainty [in the accuracy of your algorithm], but a lack of data doesn't allow you to make your statistical inputs good enough.

What do you think of the Microsoft bot example on Twitter [where the bot quickly mirrored other users' sexist and racist language]? That's another source of bias: it seems to be a case where an algorithm gets led astray because the people it is learning from are not very nice.

SUH: There are some societal and cultural norms that are more acceptable than others. For each of us as a person, we know and we learn the difference between what is and isn't acceptable through experience. For an AI system, that's going to require a tremendous amount of thoughtful training. Otherwise, it won't pick up on the sarcasm. It'll pick up on the wrong context in the wrong situation.

ZETTELMEYER: That's right. In some sense, we face this with our children: they live in a world that is full of profanity, but we would like them to not use that language. It's very difficult. They need a set of value instructions—they can't just be picking up everything from what's around them.

SUH: Absolutely. And Western culture is very different than Eastern culture, or Middle Eastern culture. So culture must be considered, and the value code [that the algorithm is trained with] has to be intentionally designed. And you do that by bringing in policymakers, academics, designers, and researchers who understand the user's values in various contexts.

ZETTELMEYER: I think there's actually a larger point here that goes even beyond the notion of bias.

I'm trained as an economist, and very often economics has not done a particularly good job at incorporating the notion of "values" into economic analysis. There's this very strong sense of wanting to strive for efficiency, and as long as things are efficient, you can avoid considering whether the outcomes are beneficial to society.

What I find interesting is that in this entire space of AI and analytics, the discussion around values is supercharged. I think it has to do with the fact that analytics and AI are very powerful weapons that can be used in very strategic, very targeted ways. And as a result of this, it seems absolutely crucial for an organization that chooses to implement these techniques to have a code of conduct or a set of values that governs these techniques. Right? I mean, just because you can do something doesn't mean that you actually ought to do it.

Where you have these very powerful tools available that can really move things, you have an obligation to understand the larger impact.

SUH: Accountability is one of the five areas that we are focusing on for creating trust in AI.

Many businesses are applying AI to not just create better experiences for consumers, but to monetize for profit. They may be doing it in ways where, say, data rights may not be balanced appropriately with the return on economic value, or efficiency. So it's an important discussion: Who's accountable when there are risks in addition to benefits?

ZETTELMEYER: Do you think this is new?

SUH: I do a little bit, because in previous scenarios, business programs and applications were programmable. You had to put in the logic and rules [explicitly]. When you get into machine learning, you're not going to have direct human intervention at every step. So then, what are the design principles that you intended?

ZETTELMEYER: So a fair way of saying this is, in essence, we've always had this issue of ownership, except with machine learning, you can potentially get away with thinking you don't need it.

But you're saying that that's a fallacy, because you do need accountability at the end of the day when something blows up.

SUH: Exactly. And this goes back to [training an algorithm to have] a fundamental understanding of right and wrong in a wide range of contexts. You can't just put the chat bot into the public sphere and say, "Here, just go learn," without understanding the implications of how that system actually learns and the subsequent consequences.

ZETTELMEYER: Okay, accountability. What's your second focus area to build trust in AI?

SUH: It's a focus on values. What are the norms for a common set of core principles that you operate under? And depending on different cultural norms, whom do you bring into the process [of creating these principles]?

There's a third focus area around data rights and data privacy, mostly in terms of consumer protection—because there are companies that offer an exchange of data for a free service of some sort, and the consumer might not realize that they're actually giving permission, not just for that one instance, but for perpetuity.

ZETTELMEYER: Do you think it is realistic today to think of consumers still having some degree of ownership over their data?

SUH: I do think there's a way to solve for this. I don't think we've solved it yet, but I do think there's a possibility of enabling individuals to understand what information is being used by whom and when.

Part of that is a burden on the institutions around explainability. That's number four—being able to explain your algorithm: explain the data sets that were used, explain the approach holistically, be able to detect

where you might have biases. This is why explainability and fairness—that's number five—go hand in hand.

ZETTELMEYER: In an academic context, I refer to this as transparency of execution.

I actually thought you were going to say something slightly different, that we need to move to a place where some of the more flexible algorithms like neural networks or deep learning can be interpreted.

It's a hard problem because, in some sense, precisely what makes these algorithms work so well is what makes them so hard to explain. In other words, the problem with these algorithms isn't that you can't write them down. You can always write them down. The problem is that it's very difficult to create some easily understandable association between inputs and outputs, because everything depends on everything else.

But I think the point you were making is: okay, even if we do have a so-called "black box" algorithm, a lot of the biases arise, not necessarily from the algorithm per se, but from the fact that we're applying this algorithm to a particular setting and data set, yet it's just not clear to people how it's being implemented.

SUH: That's right.

When and for what purpose are we actually applying AI? What are the major sources of that data? And how are we working to, if not eliminate bias, maybe mitigate it?

ZETTELMEYER: I think a lot of the trust problems that have occurred in the tech industry—and particularly in advertising—over the last years are directly related to a lack of transparency of that type. I'm always amazed that when you go to the big advertising platforms, and you approached them purely as a consumer, and then you approach them as a client, it feels like you're dealing with two different universes. As consumer, I'm not sure you have the same sense of exactly what's happening behind the scenes as you do if you happen to be an advertiser, and you have exposure to all the digital tools that you can use for targeting.

I think transparency, the way you're talking about it, is not particularly well implemented in many tech companies.

SUH: No. And there's not a common language for talking about it either, in terms of explicitly saying, "We only use data that we have access and rights to, and this is how we collect it, and you've given us permission for it." The standards around the language itself are still being developed.

ZETTELMEYER: What are you doing about all this at IBM?

SUH: We actually developed a 360 degrees fairness kit as part of our broader AI OpenScale initiative. AI OpenScale is an open-technology platform that enables your business with visibility, control, and the ability to improve AI deployments, helps explain AI outcomes, and scales AI usage with automated neural-network design and deployment, all within a unified management console. It includes open-source toolkits to check for unwanted biases in data sets and machine-learning modules. It checks for biases like explainability around your data sets to provide feedback on different aspects of your models.

It's the first open-platform and open-source toolkit to even begin to get developers thinking about bias proactively.

FEATURED FACULTY

Florian Zettelmeyer

Nancy L. Ertle Professor of Marketing; Faculty Director, Program on Data Analytics at Kellogg; Chair of Marketing Department