[MUSIC PLAYING]

CATHERINE D'IGNAZIO: It's the human behavior that is making the brain of the machine. That is how you make the machine intelligent.

SARAH HANSEN: Today on Chalk Radio, how the future of artificial intelligence and machine learning education might look a lot more human.

JACOB ANDREAS: What's the difference between saying, you know, a restaurant review expresses positive sentiment, which is a very complicated social phenomenon, and saying that, you know, a picture of the number 9 is a picture of the number 9, which is much less complicated as a social phenomenon.

SARAH HANSEN: I'm your host, Sarah Hansen. This week, we're talking with three interdisciplinary collaborators about their mission to foster a new kind of approach to computing technologies. They've created an assignment that challenges students to take a critical approach as they build the technologies of the future.

CATHERINE D'IGNAZIO: I am Catherine D'Ignazio. I'm an Assistant Professor of Urban Science and Planning.

JACOB ANDREAS: My name is Jacob Andreas. I'm an Assistant Professor in the Department of Electrical Engineering and Computer Science and also the Computer Science and Artificial Intelligence Laboratory.

HARINI SURESH: My name is Harini Suresh. I'm a fifth year PhD student in Computer Science.

SARAH HANSEN: In spring 2021, Catherine, Jacob, and Harini were brought together as part of a special initiative called SERC from the MIT Stephen A. Schwarzman College of Computing. SERC stands for the Social and Ethical Responsibilities of Computing.

CATHERINE D'IGNAZIO: So the mission broadly is thinking about how do we cultivate responsible creators of computational tools and technologies for the folks who are going to go out and be building the tools of the future. You know, in a lot of cases, you don't know what the ethical implications of something are until, you know, a kind of more abstract tool or technology or algorithm is kind of plugged into the human context, right-- like that place where, you know, the machine learning and people meet.

JACOB ANDREAS: There have been lots of high profile incidents involving things like face recognition software, people trying to deploy machine learning systems in the context of things like sentencing or recidivism prediction, and in many cases, having sort of seriously harmful effects.

SARAH HANSEN: For Jacob, exploring the social consequences of AI got him to start thinking a little differently in his own survey course-- 6.864 Natural Language Processing, or NLP for short.

JACOB ANDREAS: The way we have traditionally taught machine learning classes, it's always, OK, you know, here's a data set of photographs of digits. Classify this picture as whether it contains a 0 or a 1 or a 2 or a 3. And you know, here's a data set of restaurant reviews. And they just happen to already have assigned to them labels for whether this is a positive restaurant review or a negative restaurant review. You know, go train the machine learning model.

And at no point do you stop and ask, OK, but where did these restaurant reviews come from? And what's the difference between saying a restaurant review expresses positive sentiment, which is a very complicated social phenomenon, and saying that, you know, a picture of the number 9 is a picture of the number 9, which is much less complicated as a social phenomenon.

SARAH HANSEN: So he teamed up with Catherine and Harini to show students that the machine part of machine learning is very much influenced by humans.

HARINI SURESH: So the way that machine learning systems work in many cases is something called supervised learning, where the data set contains examples as well as labels for those examples. So for example, in content moderation, you might have-- in your data set-- comments from a message board as well as annotations that say this is a toxic comment or this is not a toxic comment. So then what the machine learning system is trying to do is learn from that data What. Makes up a toxic comment, what are characteristics of toxicity, and it uses the annotations to figure that out.

And those annotations typically come from somewhere. So they might be automatically generated by looking at historically what sorts of comments have been moderated. Or they might be generated by people.

You might crowdsource this or get specific groups of people to annotate comments as to whether they think they're toxic or non-toxic. And those would become the labels in the data set that the machine learning system would learn from.

CATHERINE D'IGNAZIO: We need to train it basically. And the way to train it is you look either, like Harini said, at historical data or you make your own data set where you and your team or you and a group of people or you hire people to say, this is toxic, this is not toxic. Or let's keep this one, let's not keep that one.

And that-- that annotation-- that is how you make the machine intelligent. And so if we back up a little bit, it's the human behavior that is making the brain of the machine. That's why, you know, that step is really important in there, that human step. It's also the step that we were trying to tune people into as not just being just like this objective thing that the machine makes up by itself based on completely objective parameters or whatever.

JACOB ANDREAS: People take these classes and then they go out in the real world and find themselves building detectors for even more complicated things, like these toxicity detectors. And that was what really felt like the sort gap between the way we were training people and the way these tools were getting deployed in practice.

SARAH HANSEN: Together, they designed a brand new assignment for the course, one they hoped would get students thinking about the human element of machine learning. First, students wrote instructions for annotating data sets, and then they tried to follow each other's instructions.

HARINI SURESH: The overall goal of the assignment was to try and get students to go from thinking about data as this pre-existing, objective, ground truth to thinking about it as the product of a long and complex process that involves many steps and is driven by human judgments and values. The goal of this assignment is not to say that data is not useful or that it's bad but rather to help students critically think about data sets when they receive them or use them or hear about them being used and to help them sort of ask those questions of who was this created by, how was it created, what are its capabilities, and what are its limitations.

SARAH HANSEN: When Catherine, Harini, and Jacob started reading the students' responses, they realized that the assignment helped students think differently about the instructions they were giving annotators. But it also did something none of them had expected.

JACOB ANDREAS: The thing that surprised me most was the number of students who said, I've never done an assignment like this in my whole undergraduate or graduate training, right? And this is an advanced class. These are people who are seniors in college or in their first or second year of graduate school. And for many people, it was really the first time they'd actually been asked to sort of think about the process by which these data sets that they've been seeing since, you know, their sophomore year were actually being generated.

HARINI SURESH: Initially, we had designed this to focus on subjectivity that annotators might have-- so like subjectivity in labels in data sets. But there were a bunch of other things that people learned about the entire data set pipeline. So for example, the categorizations that people came up with for the same problem were drastically different in some cases.

People were sort of surprised by how much personal judgment they had to use to decide these things. Everyone was like, wow, I was very surprised by the amount that I wasn't sure and the amount that I had to sort of rely on my own biases or judgments to decide what I actually thought about this.

SARAH HANSEN: And these students' reflections actually pointed to much bigger questions within the field as a whole.

JACOB ANDREAS: Of the numerous ways that Harini and Catherine mentioned of constructing these data sets, one that has become I think particularly important within the machine learning research community these days is crowdsourcing, where there's some online portal where people can log in and sign up for a teeny little labeling task-- like just look at this one picture and tell me whether or not this is a picture of a cat. Or look at this one comment and tell me whether it's a toxic comment or not. And then you get $0.05 or $0.10 in exchange for doing this little micro task.

So there's two sorts of things to think about when using these kinds of platforms. One is just thinking about the well-being of the annotators themselves, that it's very easy to miss calibrate the amount of time that it takes to do one of these micro tasks and wind up not actually paying people a living wage for doing them. And there are actually people all over the world who rely on these kinds of crowdsourcing platforms as their primary source of income.

And another thing-- once you, again, start to think about tasks like toxic comment detection or even more sensitive things like recognizing images of pornography or images of violence or whatever, it is relatively easy-- without you as the sort of system builder yourself having had to look at any of this data-- to dump just an enormous amount of like really traumatizing content on people who you're paying $0.10 a pop to label. And there's all kinds of studies showing that people really experience post-traumatic stress disorder, various other kinds of mental health issues when subjected to these kinds of things.

SARAH HANSEN: In our conversation, this notion of context kept coming up. The context in which data are created, pulled from, and annotated is incredibly important when thinking about how to improve machine learning systems.

HARINI SURESH: One thing that comes to mind is Desmond Patton's lab at Columbia. He works on context aware annotations of social media data. So specifically, the work of his that I read is around Twitter-- so looking at tweets and specifically tweets in inner city Chicago.

And the task that they're trying to do is analyze tweets from gang-involved youth in Chicago. And if you look at some of these tweets, they're trying to annotate them with things like whether they're violent or whether they indicate that there's a violent event going on. And if you were to just read them without any awareness of the context and try to annotate them, something that seems like it's super violent, if you actually were in the context and were part of that community, you might know that it's like a lyric from a local rapper or something that requires a lot of community specific expertise.

So what they did in this project was actually get experts from the community to look through this data and do context aware annotation. And they found that they were able to do a much better analysis of this data that was much more accurate and grounded in the actual context that it was a part of. So that's, I think, one example where, if you were to just apply generic tools, you would really fail at this task.

CATHERINE D'IGNAZIO: There are problems pieces that are slightly less culturally fraught, where there's less room for stereotypes, bias, pre-existing structural inequalities to enter into, like if we're training a system to recognize numbers in an image-- you know what I mean? That is just-- there's less room. I mean, obviously there's still going to always be space for interpretation, but there's less room because we don't have pre-existing conceptions of number 9's unworthiness or something like this, right?

But it's more when we enter into data about using human language or that is being used for decision-making systems that have real-life consequences for human beings. So if we're training resume screeners, for example-- so like systems where a large company would do an automated system to screen resumes and then only put the top ones up to the humans or whatever.

We're baking in a lot of biases in the process, and that has to do with where the data are coming from. And it also has to do with who are labeling the data and then who are developing the technologies.

And it's not because they're evil, you know. Like, it's not because there's bad people at all stages of this pipeline. It's that we haven't sufficiently kind of trained people along the way.

There are tools to deal with bias, stereotypes, structural inequalities. But they just-- they come from other disciplines. So how do we bring those things together to ultimately develop a more robust system that works better for everybody?

But just because the data are not the subject of ground truth doesn't mean we just throw up our hands or, like, forget about it, we can never do anything. It just means we have to enter with more caution and more transparency and reflexivity around what are the boundaries and the applications of the knowledge that we're producing.

SARAH HANSEN: I asked Jacob, Harini, and Catherine what they'd like to hear from you, our listeners, about how to help students take a critical approach to computing.

HARINI SURESH: One question that I have that I've been thinking about is, what's the right format to introduce students to these concerns? So in this case, we did an assignment within a machine learning class. And I wonder, how does that compare to sort of having it be a small portion of every assignment instead of just one assignment? Or how does that compare to having a class that's more primarily dedicated to social and ethical concerns?

CATHERINE D'IGNAZIO: This is my eternal thing with these classes there's like how do you weave in the critical while still giving people-- empowering people with tools to ultimately change practices and also to change the tools eventually too. Because the tools don't in themselves also work perfectly. They all have their own politics.

So if people have ideas about how do you both build skills but also have them interrogating tools, interrogating politics, the critical context in which the tools are used, I would love feedback or ideas around that.

SARAH HANSEN: If you have insights to share about engaging students in thinking about the social and ethical responsibilities in computing, please get in touch with us at the link in our show notes. And when you do, you'll be joining Catherine, Harini, and Jacob in spotlighting just how human the digital world is. If you're interested in learning from their open and free teaching materials or remixing them in your own teaching, you can find them on our MIT OpenCourseWare website.

Thank you so much for listening. Until next time, signing off from Cambridge, Massachusetts, I'm your host, Sarah Hansen from MIT OpenCourseWare.

Chalk Radio's producers include myself, Brett Paci, and Dave Lishansky, scriptwriting assistance from Aubrey Calaway. Show notes for this episode were written by Peter Chipman. The SERC resource site on OCW was built by Cathleen Nalezyty. We're funded by MIT Open Learning and supporters like you.