College of Business
LEHIGH UNIVERSITY

ilLUminate Blog Transcript: Kofi Arhin on Using AI to Moderate Content

Recorded September 12, 2025. Listen to it here.

ANNOUNCER: 00:01	This podcast is brought to you by ilLUminate, the Lehigh Business blog. To learn more, please visit us at business.lehigh.edu/news.
STEPHANIE VETO: 00:13	Welcome to ilLUminate, the podcast for Lehigh University's College of Business. I'm your host, Stephanie Veto. Today is September 12th, 2025, and we're talking with Dr. Kofi Arhin about using artificial intelligence for content moderation. Dr. Arhin is an assistant professor in the Decision and Technical Analytics Department. His research interests include artificial intelligence design and implementation, information security, ethical issues in IS, human-computer interaction, and web technologies. Hi, Kofi. Welcome to the show.
KOFI ARHIN: 00:49	Thank you, Stephanie. I really appreciate you having me.
VETO: 00:52	How did you get into studying AI and ethics?
ARHIN: 00:55	Back in 2018, I was a new PhD student looking for experience on research projects. And my advisor-- he wasn't my advisor then, but a professor in the College of Business gave me the opportunity to work on a project that involved using AI to predict patients that would require critical care sometime in the future. So we are talking three months, six months, one year down the line. And in that project, I understand how powerful AI could be. So it got me interested in looking at how we could use AI for social good. And that got me on the path of all of these research projects that I'm working on, that are focused on making sure that AI works for everybody, not just a few people.
VETO: 01:42	Can you discuss the difference between machine learning and AI?
ARHIN: 01:47	Yeah, that's a question I get very often. People have different takes on it. My take is that you can think of AI as the broad umbrella. So when we talk about AI, we are looking at building systems or machines that make decisions like human beings. So these systems should exhibit some human intelligence, whether it's decision making or solving problems, right? And then machine learning is a subset of that focused on creating systems that make decisions based on a lot of data or the amount of data that we feed it. So for example, if I want a computer to be able to distinguish between a cat and a dog, which is like one of the popular examples out there, I'm going to feed it with a lot of cat pictures and a lot of dog pictures. And then I'll leave the machine to figure out where the differences are. And so those are different branches of machine learning. And when you hear supervised, unsupervised, reinforcement learning, all of them do different things, but they are all under the umbrella of AI. I don't know if that's clear enough, but you can ask questions and then I'll give you more detail.
VETO: 03:02	No. I think that's really clear. No, I understand it. And when I was researching, just sort of Google searching ideas about what you're working on and ideas for questions, one thing that came up was this term natural language processing or NLP. Can you describe what that is?

ARHIN: 03:23

When we say natural language processing in the context of AI, we are looking at creating representations of textual data that we can feed to machines. So again, going back to machine learning, when you are training these models to be able to make these decisions, the data we feed them are in numerical format. Although we are giving them information, it has to be presented in tables and columns with figures that represent different things. And so in natural language processing, we are converting the text into numbers and then feeding those numbers to the models to be trained. So you can think of natural language processing as an entire pipeline of taking some text, for example, reviews of products, converting those reviews into numbers, and then training models on those numeric data sets to be able to make decisions. For example, which product is not doing well, which product is getting good reviews and so on and so forth.

VETO: 04:29

Can you break down how AI is trained to do certain jobs like moderate hate speech online? Because that seems so specific, but also so huge at the same time.

ARHIN: 04:40

Yeah, I hinted on that a bit when I was explaining what natural language processing is. So when we are training AI models, our goal is to get it to make decisions like humans. What are some decisions that humans make? For example, should I hire this candidate or should I publish this marketing promo? When is the best time to put out this product, and so on? And so when it comes to natural content moderation, we are training models to decide whether this content should be made public. Should we allow other people to see this content? Should we approve the publication of this content? And so on and so forth. Content moderation is a bit broad. It can be anything from deleting posts that are not appropriate to content that is hateful to preventing and blocking accounts or preventing people from seeing a particular post or social media content. And so the goal in content moderation for social media on social media platforms is to safeguard the sanity of community members. I think that social media is a great platform if it's used for the right things, right? And so with the benefits, there are also challenges that come with it. If you are giving everybody access to these platforms, and think of it as a marketplace, you have good products in there, you're going to have bad products in there, and content moderation is focused on taking those bad products from the marketplace so that everyone is safe.

VETO: 06:24

I mean, you always hear a lot about ethics and AI, and I feel like this is a huge one here as well. And also, does it often depend on the person or company behind that AI model to decide if something should be published or not, is hate speech or not? What if that moderator, the human person behind the AI or whatever's being trained, thinks that this isn't bad?

ARHIN: 06:57

No, that's a great question. And I think that your previous question was trying to get at that, and I didn't fully complete the explanation. So the way we train the content moderation, or we used to train them in the past was we curate a large number of social media content online. And then we hire these people, we call them annotated, right? We hire these people and ask them to label these thousands of social media content. We ask them to label whether this is hateful, this is not, is this offensive? And so in the end, you have a large data set of social media content and human labels. Hopefully the content captures all the variety of posts we have on these platforms. And then when you have these labels, you now train your models to detect-- so you are feeding them with all of this content to be able to detect whether

something is hateful or not. And so the model behaves according or very similar to the knowledge base or the data set that you use to train it. And so if you have people called annotators making these decisions who are not putting in the effort or there's no integrity in the labeling process, then you have AI systems that are going to behave in a similar manner.

ARHIN: 08:17

And in content migration, there are issues around consistency, accuracy, and also the silencing of minority voices. For example, if I label content as hate speech that is not really hate speech because I do not understand the culture or the context of what was said, you create an ecosystem where these models now are looking for phrases or text or similar text samples to take off the platform and so on. So that results in the silencing of people's voices. And so for some platforms, what they do is they have human reviewers who review edge cases. So when I say edge cases, I mean cases that do not easily meet the threshold of what is hateful or what is not hateful. And so sometimes you have human reviewers making this decision. But look, even if we left content moderation to humans alone, people still make mistakes, right? And so there are challenges that are being addressed. You don't have a perfect system, but at least you want to make sure that the system is working for everybody and not against a few people or a few underrepresented groups.

VETO: 09:38

And if it's generative, it's learning like with everything, right? And so I'm curious about if it's there or if it's going to be there when it comes to things that are coded, like a phrase that's on the surface doesn't seem like hate speech, but it's become a popular term among a certain group and used as something hateful. Do you think AI will get to a point where it catches that or it does that now? What's that like?

ARHIN: 10:11

That's a great question. So in the past, it was very difficult for machine learning models to catch-- and these are our core edge cases. Things that do not seem hateful. There are no hateful words being used. But when you read the entire post or content, you know that it has been-- for example, biased toward a certain group, right? And so in the past, it was very difficult for models to catch these. But generated systems are showing promise because they have been trained on a large amount of data. It's not just social media content. It's news articles. It's published journals. It's websites all over the world. And so what happens is they begin to understand context, they begin to connect the dots, and they can make more intelligent decisions, at least in the context of content moderation. I am seeing generative AI showing a lot of promise in terms of addressing the challenges that the predictive AI systems had.

VETO: 11:24

This sort of relates to that too, where it's generative AI. And is it learning that something like hate speech isn't objective, right? So is it doing that?

ARHIN: 11:40

Yes. Like I said, it's showing a lot of promise. The reason why I feel like generative AI systems have an advantage is they are not just learning the text and the representations. They're also learning the context, right? And so you can show two similar phrases to a generative AI model, and then it has to assume the content and then give you some output. I think that that is where traditional or predictive AI systems fall short. And so because generative AI systems have that capability to learn contexts and representations, it's easier for them to capture. It's just based on the information they are trained on and how they work. I would say that, yes-- again, back to my-- I know I'm using this a lot, but they are showing a lot of promise. They

are far better than traditional predictive AI systems in detecting coded language, but there's also room for improvement.

VETO: 12:40 That's incredible. It's really incredible when you think about it that it can do something that isn't objective, and it'll get there to know that these two sentences mean completely different things in different contexts. This has nothing to do with your articles. [laughter]

ARHIN: 13:01 So frankly, even if you put humans in charge, right, and you said, "Hey, I'm going to put humans in charge to make all of these decisions," one, humans also make mistakes, right? And so, I mean, there's no way to get around that. People are going to make mistakes. These are not malicious to say-- they are not malicious in the context that people are not deliberately labeling something that is hateful and not hateful, but you need context, you need a lot of information to make these decisions. And who has the capacity to be able to connect millions of dots in a few seconds, AI systems. And so that's why I'm claiming that there's a lot of promise in that area, and we are seeing that bear fruit. There are companies that are currently using generative AI systems for content moderation.

VETO: 13:58 Absolutely. And I love that you described it as it's connecting millions of dots in an instant. And I use this example only because it just happened the other night, and I think about it. And it has, again, nothing to do with your research, but my daughter's 10 and she's interested in AI now. And she's growing up with it. And she's like, "Do you have ChatGPT?" And I was like, "Yeah." She's like, "Let me see." And so I showed her and she just wanted me to type in a random word. She's like, "Type in cucumber." And I was like, "Okay." And then it answered and it said, "A cucumber is this, this, this, it's a fruit." And she goes, "Well, wait a second. Isn't it a vegetable?" And so I typed in, "Isn't a cucumber a vegetable?" And it was like, "Yes." And it gave this entire thesis on why it justified calling it a fruit. And then it asked if we wanted to generate a graph of vegetables that are often called fruits and vice versa. And it blew my mind.

ARHIN: 14:58 Yeah. You could never think of that. That's a fascinating thing about, especially, generative AI system and the ability to connect dots, give you context, and give you output that you would never have thought of. So yeah. Yeah. The other thing I was going to add to what I said previously was that when you have millions of people posting online on social media platforms, you want systems or a process that will be able to detect these efficiently or as fast as possible. And you cannot rely on humans for that. It will take way too much time for us to read content and give feedback before allowing people to post. And so I think with that, AI systems are in the best position to help with that, just based on the sheer amount of volume that people produce on social media platforms these days.

VETO: 15:57 You're working on a paper with several colleagues, and it's about AI and content moderation. Can you give us a rundown of the study?

ARHIN: 16:05 Sure. So before I go on, I'd like to give a shout out to my co-author. So I'm working on this with Dominic Packer in the psychology department, Haiyan Jia in the journalism and communication department. And then Karleigh Groves is also a PhD student in the psychology department. We have a number of projects in this area. Our focus first is to try to understand how generative AI systems make content moderation decisions. We are interested in first, seeing if they exhibit the same inconsistencies

and inaccuracies that we see in human labeling of hate speech. I refer to these humans as annotators initially. So you want to see if they have comparable outcomes in terms of labeling hate speech, one. And then we want to know if we can nudge the models to be more strict or more permissive using a theory we call the regulatory focus theory. And so what we do in this paper is we curate some hateful speech and non hateful speech all based on some publicly available data. And then we give these models three sets of instructions.

ARHIN: 17:41

In one instruction, we tell them to make sure they do not miss labeling any hate speech. In one instance, we tell them, "Hey, make sure that you do not falsely claim something is hateful when it's not." Or trying to nudge them to use the construct in the regulatory focus theory. And then we find variations in how these models respond to the instructions. And so for one set of instructions where we ask them to be very strict, we are seeing that there's very little variations in their output. And then in one set of instructions where we tell them to make sure they do not wrongly label things as hate speech, we are finding that there's a wide variation there just because we are giving them a lot of room. Now, someone will say that these findings are obvious, but we were just interested to see if they would respond to these stimuli like people do, one, and also to see if there were opportunities for us to explain biases and inaccuracies in labeling hate speech online. And so we have very important findings for [inaudible] practice. And this paper is currently under review, so I can only share so many details, but ask me anything about it. I'm happy to talk, provide more information if you need it.

VETO: 19:14

Okay. Yes. Well, I got a sneak peek of your research. And one thing that I did read that I would love a little bit more information on is you discuss how generative AI systems adjust the moral thresholds in relation to motivational cues with drift most pronounced under conditions of linguistic ambiguity. Can you talk about what that means, especially like the term drift?

ARHIN: 19:41

Okay, that's a great question. So in the paper, we proposed this theory that we give it a title. We call it something drift, and then we say-- the purpose of that is to try to explain the mechanisms that lead to generative AI model variations in labeling hate speech. And so in the experiment, we prompt a number of generative AI models, and then we compare how they agree based on the set of instructions that we give them. And so we refer to drift as the differences in their labeling behavior, where they agree, where they do not agree, and how they are reacting to the stimuli. And then in the end, we try to explain some mechanisms for why this is happening, why they are not consistent in one case, but very consistent in the other case and so on.

VETO: 20:39

Can you talk about you and your colleagues' takeaways from this research? I know you said some of it would be obvious, but you are learning a lot from the process of these three steps.

ARHIN: 20:49

Yeah. I think that what we did not expect going in was the level of significance that we would see in the model's response to these different stimuli. All of these are instructions, but just changing a few words in the instructions to nudge it towards different perspectives resulted in outcomes. Although we hypothesized these outcomes, the results amazed us. But also, it was important for us to highlight why certain stimuli might be best or better in certain situations. For example, if you have a platform where you want to ensure that every single voice is heard, you want to be

promotion-focused, for example, because when you make these models strict, they tend to make mistakes in the-- they make a lot of mistakes when they are labeling content that is not offensive. So you're asking them to be strict. They are seeing something that's not offensive. They would rather err on the side of labeling something hate when it's not hate speech. And then when you're asking it to be permissive and promotion-focused, they obviously let a lot of hateful speech go. But the difference is in the level of ambiguity of the hate. So you talked about edge cases, right? When something is hate, but it's not always like subtle hate speech, right, those ones also create a lot of variation in the models. And we put all of that in one bucket and we call it moral drift in the end.

VETO: 22:46 And so you're seeing AI for content moderation in a lot of companies now. And I know that an article came out a few weeks ago that Meta is using AI for most of its content moderation. What are your thoughts on that?

ARHIN: 23:03 So I think I alluded to this earlier on that we need AI for content moderation. We need more of that. Just because there's so much content being posted online these days, it's very difficult for you to hire a million people to just be focusing on finding out what's hateful and not. So for that, I think that the only way out is to use AI. And so content moderation has been going on for a long time. It didn't just start. I think that what companies are doing now is using generative AI systems to support that. So another reason why generated AI might be better than predictive AI systems is because for predictive AI systems, like I said, you give people some data, you ask them to label it, then you train a model on the label data, and then you try to deploy that model. For generative AI systems, like in our study, all you do is give some instructions, modify it. So you skip that stage or that step of giving people some data to label and then train models on their label data. That is where generative AI systems are showing a lot of promise. You give them instructions, show them a few examples, and then they perform a very decent job. And so I'm not surprised that companies like Meta are using it. I think it's the way to go. We just need to make sure that it works for everybody. And that has been the focus of a lot of the studies that I'm working on. AI is good. Let's make it work for everyone.

VETO: 24:40 I love that. Well, I mean, it wasn't in the not-too-distant past where it was a division of people in a room literally moderating content, looking at stuff that's posted, things that are flagged, things that are questionable, which can be absolutely horrible. I couldn’t imagine. I know there's a movie on it now where this lady sees like something bad happen and whatever. But I mean, I feel like using a generative AI system to take that and be able to stop stuff from getting so far is a lot easier and I think really important too.

ARHIN: 25:21 Yeah. And I think a lot of the skepticism comes from people saying, "Oh, the AI is going to make mistakes." I mean, we can't have it both ways. We can't have a perfect AI system without allowing the AI system to make mistake, right? It becomes a challenge if you are discovering mistakes by the AI system and then we are overlooking them. But the only way to build equitable AI systems is to allow us to deploy these systems and then make corrections or make adjustments as they do the work that we've deployed them to, right. And so sometimes a very initial elevation of these systems will be deployed. They might not be perfect, but it will help us build the systems that we are looking for. We have to start from some point. And I think that

people should be more open to allow AI systems, especially generative AI systems to support some of the processes in organizations and so on.

VETO: 26:24 I think it could even get to a point where it's on a small scale, whether it's a small business or a communications department or something like that where there's comments available on posts and AI can flag it automatically, and then it could be reviewed by a human a lot easier than something sneaking through the cracks and it's there for the world to see.

ARHIN: 26:46 Exactly. Yeah. I'll trust an AI system to be able to flag that compared to a human who has to be focused 24 hours a day looking out for-- yeah, I don't think that's-- I don't think that's an efficiency of anybody's time.

VETO: 27:06 Is there anything else that you're working on that you'd like to bring up or got any other research projects in the works?

ARHIN: 27:12 Yeah. So generative AI system is really broad. Like I said, it's a really broad area. Like I said, my research is focusing on using AI for social goods. So I'm looking at trying to explain how we can make generative AI systems companions in moderating, for example, inappropriate content for kids, inappropriate images like pornography and things like that that these platforms are plagued with. Some of my work is on using generative AI system to transform interview responses from candidates so that the review process is tailored or it's uniform across all candidates and so on and so forth. So really just focusing on how we can use all these AI systems that exist for social good. That'll be my focus.

VETO: 28:08 Well, Kofi, thank you so much for taking the time to talk with me today. It was really great hearing about all this and everything that you're working on.

ARHIN: 28:17 Thank you so much, Stephanie. I sincerely appreciate your time. And let's talk when this article is published and we have more insights to share.

VETO: 28:29 That was Dr. Kofi Arhin speaking with us about artificial intelligence and content moderation. This podcast is brought to you by ilLUminate, the Lehigh Business blog. To hear more podcasts featuring Lehigh Business Thought Leaders or to follow us on social media, please visit business.lehigh.edu/news. This is Stephanie Veto, host of the ilLUminate podcast. Thanks for listening.