TechXplore

Insider Q&A: Trust and safety exec talks about AI and content moderation

April 23 2024, by Barbara Ortutay

Credit: AP Illustration/Jenni Sohn

Alex Popken was a longtime trust and safety executive at Twitter focusing on content moderation before leaving in 2023. She was the first employee there dedicated to moderating Twitter's advertising business when she started in 2013.

Now, she's vice president of trust and safety at WebPurify, a content moderation service provider that works with businesses to help ensure the content people post on their sites follows the rules.

Social media platforms are not the only ones that need policing. Any consumer-facing company—from retailers to dating apps to news sites—needs someone to weed out unwanted content, whether that's hate speech, harassment or anything illegal. Companies are increasingly using artificial intelligence in their efforts, but Popken notes that humans remain essential to the process.

Popken spoke recently with The Associated Press. The conversation has been edited for clarity and length.

QUESTION: How did you see content moderation change in that decade you were at Twitter?

ANSWER: When I joined Twitter, content moderation was in its nascent stages. I think even trust and safety was this concept that people were just starting to understand and grapple with. The need for content moderation escalated as we, as platforms saw them be weaponized in new ways. I can sort of recall some key milestones of my tenure at Twitter. For example, Russian interference in the 2016 U.S. presidential election, where we realized for the first time, realized in a meaningful way, that without content moderation we can have bad actors undermining democracy. The necessity for investing in this area became ever more important.

Q: A lot of companies, the bigger social media companies are leaning on AI for content moderation. Do you think that AI is in a place yet where it's possible to rely on it?

A: Effective content moderation is a combination of humans and machines. AI, which has been used in moderation for years, solves for scale. And so you have machine learning models that are trained on different policies and can detect content. But ultimately, let's say you have a machine learning model that is detecting the word 'Nazi.' There are a lot of posts that might be criticizing Nazis or providing educational material about Nazis versus like, white supremacy. And so it cannot solve for nuance and context. And that's really where a human layer comes in.

I do think that we're starting to see really important advancements that are going to make a human's job easier. And I think generative AI is a great example of that, where, unlike traditional. AI models, it can understand context and nuance much more so than its predecessor. But even still, we have entirely new use cases for our human moderators now around moderating generative AI outputs. And so the need for human moderation will remain for the foreseeable future, in my opinion.

Q: Can you talk a little bit about the non-social media companies that you work with and what kind of content moderation they use?

A: I mean, everything from like retail product customization, you know, imagine that you are allowing people to customize T-shirts, right? Obviously, you want to avoid use cases in which people abuse that and put harmful, hateful things on the T-shirt.

Really, anything that has user-generated content, all the way to online dating—there, you're looking for things like catfishing and scams and ensuring that people are who they say they are and preventing people from uploading inappropriate photos for example. It does span multiple industries.

Q: What about the issues that you're moderating, does that change?

A: Content moderation is an ever-evolving landscape. And it's influenced by what's happening in the world. It's influenced by new and evolving technologies. It's influenced by bad actors who will attempt to get on these platforms in new and innovative ways. And so as a content moderation team, you're trying to stay one step ahead and anticipate new risks.

I think that there's a little bit of catastrophic thinking in this role where you think about like, what are the worst case scenarios that can happen here. And certainly they evolve. I think misinformation is a great example where there's so many facets to misinformation and it's such a hard thing to moderate. It's like boiling the ocean. I mean, you cannot fact check every single thing that someone says, right? And so typically platforms need to focus on misinformation not to cause the most real world harm. And that's also always evolving.

Q: In terms of generative AI there's some doomsday thinking that it will ruin the internet, that it will just be, you know, fake AI stuff on it. Do you feel like that might be happening?

A: I have concerns around AI-generated misinformation, especially during what is an extremely important election season globally. You know, we actively are seeing more deepfakes and harmful synthetic and manipulated media online, which is concerning because I think the average person probably has a hard time. discerning accurate versus not.

I think medium to long term, if I can be properly regulated and if there are appropriate guardrails around it, I also think that it can create an opportunity for our trust and safety practitioners. I do. Imagine a world in which AI is an important tool in the tool belt of content moderation, for things like threat intelligence. You know, I think that it's going to be extremely helpful tool, but it's also going to be misused. And we're we're already seeing that.

Citation: Insider Q&A: Trust and safety exec talks about AI and content moderation (2024, April 23) retrieved 4 May 2024 from https://techxplore.com/news/2024-04-insider-qa-safety-exec-ai.html