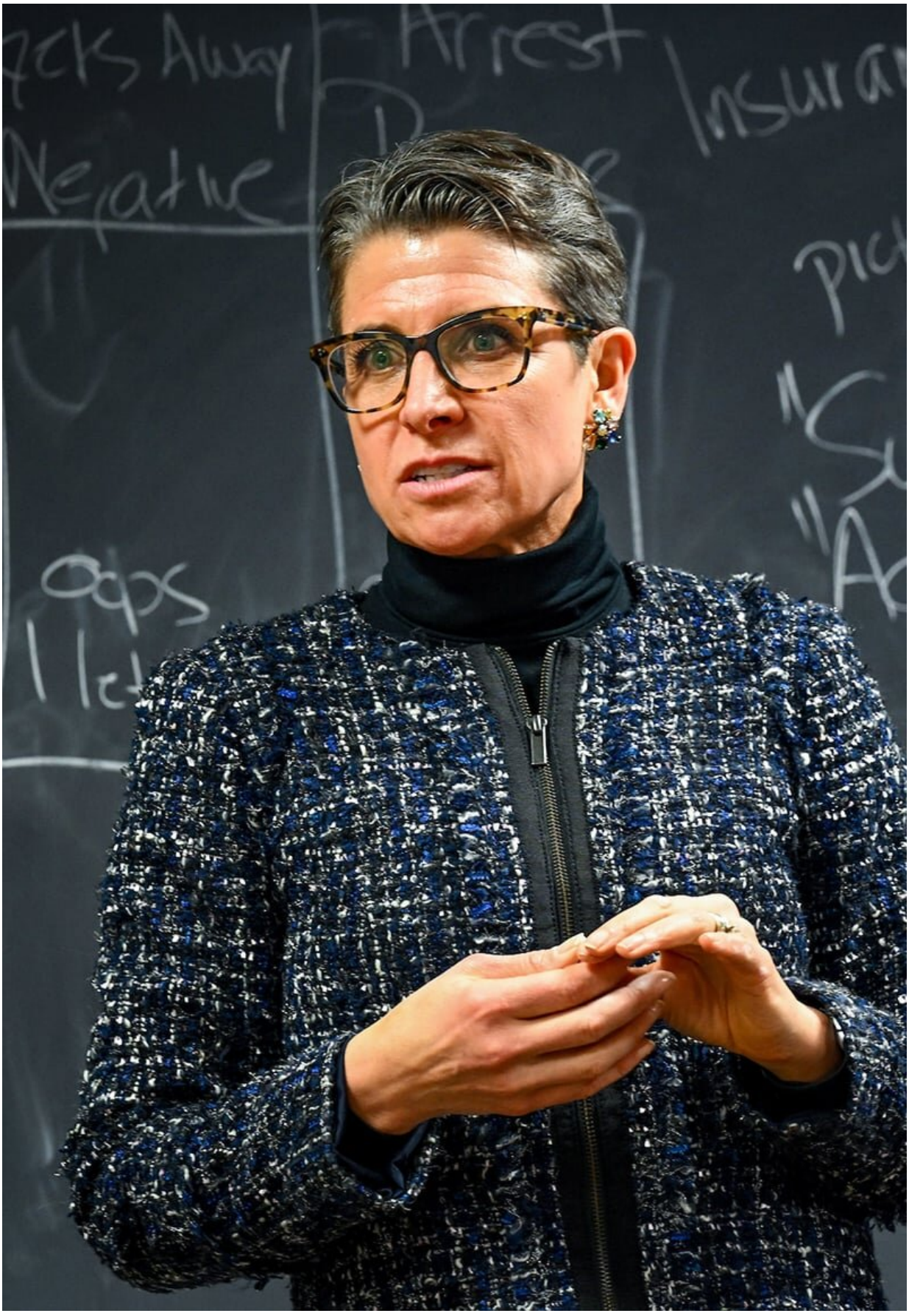
Expert promotes balance of moderation and engagement in technology ethics

February 15 2023


Arrest

Kirsten Martin, director of the Notre Dame Technology Ethics Center, teaches a class in January. ND TEC offers a 15-credit undergraduate minor in tech ethics that is open to all Notre Dame undergraduates, regardless of major. Credit: University of Notre Dame

While social media companies court criticism with who they choose to ban, tech ethics experts say the more important function these companies control happens behind the scenes in what they recommend.

Kirsten Martin, director of the Notre Dame Technology Ethics Center (ND TEC), argues that optimizing recommendations based on a single factor—engagement—is an inherently value-laden decision.

Human nature may be fascinated by and drawn to the most polarizing content—we can't look away from a train wreck. But there are still limits. Social media platforms like Facebook and Twitter constantly struggle to find the right balance between free speech and moderation, she says.

"There is a point where people leave the platform," Martin says. "Totally unmoderated content, where you can say as awful material as you want, there's a reason why people don't flock to it. Because while it seems like a train wreck when we see it, we don't want to be inundated with it all the time. I think there is a natural pushback."

Elon Musk's recent changes at Twitter have transformed this debate from an academic exercise into a real-time test case. Musk may have thought the question of whether to ban Donald Trump was central,

Martin says. A single executive can decide a ban, but choosing what to recommend takes technology like algorithms and artificial intelligence—and people to design and run it.

"The thing that's different right now with Twitter is getting rid of all the people that actually did that," Martin says. "The content moderation algorithm is only as good as the people that labeled it. If you change the people that are making those decisions or if you get rid of them, then your content moderation algorithm is going to go stale, and fairly quickly."

Martin, an expert in privacy, technology and business ethics and the William P. and Hazel B. White Center Professor of Technology Ethics in the Mendoza College of Business, has closely analyzed content promotion. Wary of criticism over online misinformation before the 2016 presidential election, she says, social media companies put up new guardrails on what content and groups to recommend in the runup to the 2020 election.

Facebook and Twitter were consciously proactive in content moderation but stopped after the polls closed. Martin says Facebook "thought the election was over" and knew its algorithms were recommending hate groups but didn't stop because "that type of material got so much engagement." With more than 1 billion users, the impact was profound.

Martin wrote an article about this topic in a case study textbook ("Ethics of Data and Analytics") she edited, published in 2022. In "Recommending an Insurrection: Facebook and Recommendation Algorithms," she argues that Facebook made conscious decisions to prioritize engagement because that was their chosen metric for success.

"While the takedown of a single account may make headlines, the subtle promotion and recommendation of content drove user engagement," she

wrote. "And, as Facebook and other platforms found out, user engagement did not always correspond with the best content." Facebook's own self-analysis found that its technology led to misinformation and radicalization. In April 2021, an internal report at Facebook found that "Facebook failed to stop an influential movement from using its platform to delegitimize the election, encourage violence, and help incite the Capitol riot."

A central question is whether the problem is the fault of the platform or platform users. Martin says this debate within the philosophy of technology resembles the conflict over guns, where some people blame the guns and others the people who use them for harm. "Either the technology is a neutral blank slate, or on the other end of the spectrum, technology determines everything and almost evolves on its own," she says. "Either way, the company that's either shepherding this deterministic technology or blaming it on the users, the company that actually designs it has actually no responsibility whatsoever.

"That's what I mean by companies hiding behind this, almost saying, 'Both the process by which the decisions are made and also the decision itself are so black boxed or very neutral that I'm not responsible for any of its design or outcome.'" Martin rejects both claims.

An example that illustrates her conviction is Facebook's promotion of super users, people who post material constantly. The company amplified super users because that drove engagement, even if these users tended to include more hate speech. Think Russian troll farms. Computer engineers discovered this trend and proposed solving it by tweaking the algorithm. Leaked documents have shown that the company's policy shop overruled the engineers because they feared a hit on engagement. Also, they feared being accused of political bias because far-right groups were often super users.

Another example in Martin's textbook features an Amazon driver fired after four years of delivering packages around Phoenix. He received an automated email because the algorithms tracking his performance "decided he wasn't doing his job properly."

The company was aware that delegating the firing decision to machines could lead to mistakes and damaging headlines, "but decided it was cheaper to trust the algorithms than to pay people to investigate mistaken firings so long as the drivers could be replaced easily." Martin instead argues that acknowledging the "value-laden biases of technology" is necessary to preserve the ability of humans to control the design, development and deployment of that technology.

The debate over online content moderation traces back to Section 230 of the Communications Decency Act. In 1995, an online platform was sued for defamation over a user's post on its bulletin board. The suit was successful in part because the platform attempted to remove harmful content, implying that moderation led to full responsibility, which discouraged any attempts.

In response, Congress passed Section 230 to protect the platform business model and encourage self-moderation. "The idea is even if you get in there and try to moderate it, we're not going to treat you like a newspaper," Martin says. "You're not going to be held responsible for the content that's on your site."

Social media companies have been successful and proactive about some types of moderation. When there is a consensus against a type of content, such as child porn or copyrighted material, they are quick to remove it. The problems start when there is widespread debate over what is harmful.

In tough cases, the companies sometimes use Section 230 to take a hands-

off approach, claiming that only more content can overwhelm lies. Martin says they argue that there is no way to do any better: "If you regulate us, we're going to go out of business."

She compares this to General Motors in the 1970s asserting that they couldn't put seat belts in cars to make them safer, or steel companies claiming that they couldn't avoid pollution.

"It's kind of this normal evolution of an industry growing really quickly without too much regulation or too much thought," Martin says. "People push back and say, 'Hey, we would like something different.' They'll come around. GM added seat belts. Eventually, we'll have better social media companies."

Martin may be optimistic because she's a fan of technology.

"It's not as though the people that critique social media want it to go away," she says. "They want it to fulfill everything that it can be. They like the positive side."

Social media can connect kindred souls with a specialized interest or people who feel lonely. That was especially valuable during the pandemic. People living under totalitarian regimes can communicate quickly and easily, avoiding government control of older technology.

In the early 2010s, Facebook was seen as a powerful force of freedom during the Arab Spring, when protesters organized online and toppled authoritarian leaders in North Africa and the Middle East.

Brought home, that power can also have a downside, depending on perspective. "If I can communicate with someone else who's just like me, that means if I want to plan an insurrection, I can find someone else that wants to plan an insurrection," Martin says.

"A lot of times, the people that are loud with hate speech have followers that see it as a call to arms," Martin says. "It's not just one person, like Alex Jones saying to Sandy Hook victims, 'You don't exist and that never happened.' It's that there's thousands of people that will then also target those people online and offline. I think it's misunderstanding the asymmetry of the bullies."

Recent phenomena, from Facebook's COVID-19 disinformation to Instagram's effect on teenagers' body image, may have soured some people on social media since its early promise. Martin identifies Gamergate in 2014 as the impetus of change in many people's perception. A loosely organized online harassment campaign targeted feminism and diversity in video game culture, spawning many of the worst behaviors that have followed.

Martin says social media users must demand moderation when hate speech goes too far. Advertisers can be another powerful force. "Brands don't want the new electric Cadillac to be right next to a white supremacist post," she says.

Musk may care more about speech he wants to promote than targeted groups right now, but that choice could open the door to a competitor like Mastodon. For her part, Martin will continue to identify the problems and trust that technology can correct course to provide a social benefit.

"I always think of tech ethics as having two prongs," she says. "One is a critical evaluation examination of the technology. But the other side is to help people figure out how to design and develop better technology. The skill set to do both aren't necessarily in the same person.

"You need people that are calling out what's wrong and then you need other people to say, 'Oh, I could fix that and this is how I would do it

differently.'"

Provided by University of Notre Dame