MIT Technology Review

MIT Technology Review

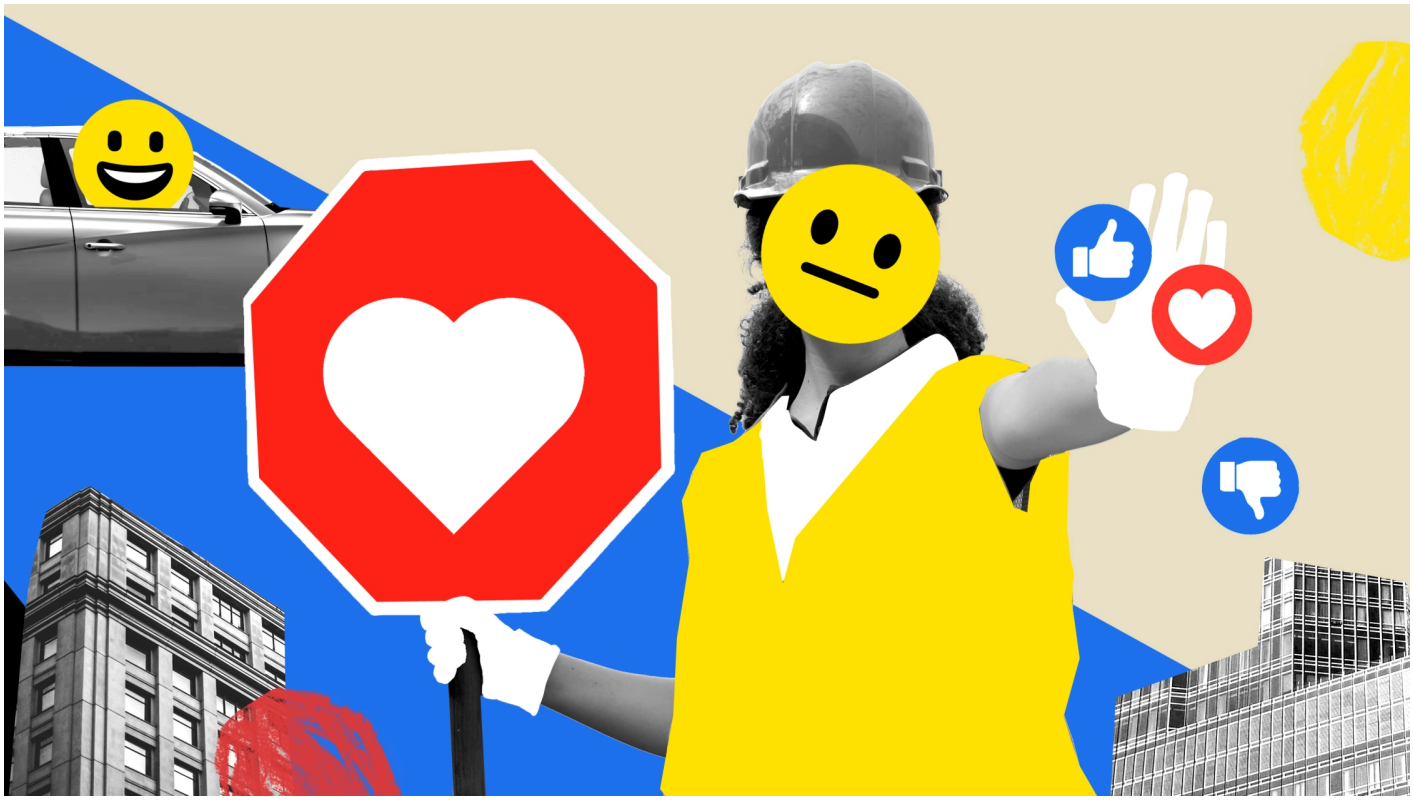
SILICON VALLEY

How to save our social media by treating it like a city

We need to make our online spaces more similar to our offline ones to limit the reach of bad actors and keep people safe.

By Sahar Massachi

December 20, 2021



ANDREA DAQUINO

Being on social media can feel a bit like living in a new kind of city. It's the greatest city in the world. Millions of people can do things their parents never dreamed of. They can live together, play together, learn together. The city is a marvel.

My job used to be to protect the city. I was a member of the Facebook Civic Integrity team. My coworkers and I researched and fixed integrity problems—abuses of the platform to spread hoaxes, hate speech, harassment, calls to violence, and so on. Over time, we became experts, thanks to all the people, hours, and data thrown at the problem. As in any community of experts, we all had at least slightly different ways of looking at the problem. For my part, I started to think like an urban planner. The city needs to be designed correctly from the beginning. It needs neighborhoods that are built so that people, societies, and democracies can thrive.

This is a different approach, one that is emerging in companies across the social media landscape: integrity design. Integrity workers like me try to defend a system from attackers who have found and learned to abuse bugs or loopholes in its rules or design. Our job is to systematically stop the online harms that users inflict on each other. We don't (often) get into the muck of trying to make decisions about any specific post or person. Instead, we think about incentives, information ecosystems, and systems in general. Social media companies need to prioritize integrity design over content moderation, and the public needs to hold them accountable about whether they do so.

First, let's take a step back: if social media is a new city, why is it so hard to govern? Why don't real cities see millions of citizens fall into cults in a manner of months? How can they have conferences without (Gamergate-scale) harassment, or clubs that don't turn people into propaganda-spewing automatons? Why don't they have waves of Nazi recruitment? What does the physical city have that the virtual one doesn't?

Physics. That is, physical limits.

As a society, we've evolved a combination of rules, norms, and design patterns that work, more or less, to rein in some kinds of terrible behavior. Those rules assume that we haven't developed superpowers. Online, however, people do indeed have powers like cloning (bot armies), teleportation (ability to post in many places simultaneously), disguise (sock puppets), and so on. In a physical city, any single propagandist is limited by vocal stamina or wallet capacity. In the online

work. In the city of social media, it requires a two-minute signup process to make a new account. The physical city is populated by human beings. In the city of social media, you could be talking at any time to someone who is secretly a robot. In a physical city, travel takes time. In the city of social media, it's trivial for Macedonian teenagers to assume the identities of thousands of people in a different hemisphere.

In a system where the worse your behavior is, the more you're incentivized to do it, after-the-fact punishment is doomed to fail. Luckily, we have other approaches. After all, the physical city also doesn't solve problems by surveilling and arresting everybody. Public health campaigns and social workers can help people before it's too late. We build public spaces like farmers' markets and libraries to create a sense of community.

If we are urban planners, then content moderators on the platforms are cops, judges, and juries rolled together. These people, who are typically underpaid and over-traumatized contract workers, have been given the impossible task of reviewing millions of potentially problematic posts, and—in a matter of seconds, without critical context—determining whether they violate the ever-changing city laws so that the proper sanctions can be applied. They are forced to decide high-stakes cases with minimal evidence and no trial. Whichever way they rule, people are furious—and the city never seems to get safer. Content moderation cannot fix the systemic problems plaguing social media any more than traffic cops could safeguard roads with no lane markings, speed limits, signs, or traffic lights. We'll never arrest or censor our way out of this problem, and we shouldn't try.

Related Story

How Facebook and Google fund global misinformation

The tech giants are paying millions of dollars to the operators of clickbait pages, bankrolling the deterioration of information ecosystems around the world.

The work of integrity teams provides a different solution. We may be in the spotlight now, but we have a long history in the industry. We've learned a lot from approaches to fighting spam in email or search engines, and we borrow a lot of concepts from computer security.

One of the best strategies for integrity we've found is to bring some real-world friction back into online interactions. I'll focus on two examples to help explain this, but there are many more such mechanisms, like limits on group size, a karma or reputation system (like Google's PageRank), a "neighborhood you're from" indicator, structures for good conversation, and a less powerful share button. For now, let's talk about two ideas that integrity workers have developed: we'll call them driving exams and speed bumps.

Imagine if it was impossible to tell whether you were talking to a bunch of people or one person rapidly changing disguises. This lack of trust is no good. At the same time, we need to remember that pseudonymous accounts aren't always bad. Perhaps the person behind the pseudonym is a gay teen who is not out to family, or a human rights activist living under a repressive regime. We don't need to ban all fake accounts. But we can make their costs higher.

One solution is analogous to the way, in many countries, you can't drive a car until you've learned how to operate it under supervision and passed a driving exam. Similarly, new accounts should not get immediate access to all the features on an app. To unlock the features that are more abusable (to spam, harass, etc.), perhaps an account should need to pay some costs in time and effort. Maybe it just needs time to "ripen." Maybe it needs to have enough goodwill accrued in some karma system. Maybe it needs to do a few things that are hard to automate. Only once the account has qualified through this "driving exam" will it be trusted with access to the rest of the app.

Spammers could, of course, jump through those hoops. In fact, we expect them to. After all, we don't want to make it too hard for the legitimate users of fake accounts. By requiring some effort to create a new "disguise," however, we're reintroducing some physics back into the equation. Three fake accounts could be manageable. But hundreds or thousands would become too difficult to pull off.

Online, the worst harms almost always come from the power users. This is pretty intuitive to understand—social apps generally encourage their members to post as much as possible. Power users can do this much more often, and to different audiences, and more simultaneously, than is possible in real life. In legacy cities, the cost of one person doing harm is bounded by the physical need for any given person to be in one place or speak to one audience at a time. This is not true online.

Online, some actions are perfectly reasonable if done in moderation, but they become suspicious when done in volume. Think of creating two dozen groups at once, or commenting on a thousand videos an hour, or posting every minute for a whole day. When we see people using a feature too much, we think they're probably doing something akin to driving at an unsafe speed. We have a solution: the speed bump. Lock them out from doing that thing for a while. There's no value judgment here—it's not a punishment, it's a safety feature. Such measures would be an easy way to make things safer for everyone while inconveniencing only a small fraction of people.

These ideas draw from the same principle: friction. It's another way of talking about imposing escalating costs on actions when they start being physically impossible. With driving exams, the action is account creation; with speed bumps, the action is posting or commenting. Costs don't have to be literal money, though. They could involve inserting a bit of lag before the action is completed, penalizing the ranking of the user's content in a feed, or changing the way something looks to make it less attractive. In essence: past a certain threshold, the *n*th post/comment/retweet/action should cost more than the previous one.

In fact, from this, we can get a nice definition of spam. Spam is what happens when people cheat by taking advantage of the lack of physics online. The solution seems to be simple—fix the physics to fight the spammers. After all, as an industry, we know how to fight spam.

Integrity design can solve some thorny problems, like the continuous war over "censorship." It can relieve pressure on an overburdened content moderation system. It can create a system that is robust to both intentional "attacks" from intelligence services and the "organic" problems of people learning how to abuse the rules. But integrity design alone can't stop companies from making bad decisions.

The stakeholder trap can ruin the work of any integrity team. A large part of fighting spam comes from figuring out loopholes in your system and then closing them. It's impossible to crack down on these spam-enabling loopholes without also hurting the people who currently benefit from them. When you do that, they get mad. They might complain, loudly. Remember this: their current reach comes from spam. They don't "own" it. It's theft, not an entitlement. If you let them cow you, you will fail.

Often, social companies fail that test.

Related Story

For too long, integrity work has been a public service trapped within private entities. We've been identifying ways to build better cities and fighting to get them implemented, but if our proposals cut against other company goals, they might not see the light of day. We've also been alone: if a team in one company

"This is not normal. This is not healthy."

we just launched the Integrity Institute, a first-of-its kind nonprofit think tank run by a community of integrity professionals. We will advance the theory and practice of integrity work, and share that knowledge with the public, companies, and each other. The world deserves to hear independent, honest explanations of how these new cities work, and how we can indeed build them better.

We know what companies need to do. They need to prioritize integrity design, with concepts like physical-inspired limits and spam-fighting strategies. They need to keep some content moderation, but focus more on investigating attacks instead of metaphorically arresting people for littering. They need to hold strong and actually enforce their own policies. We also know what the public needs to do—either pressure the companies to do the right thing or persuade their workers, advertisers, or lawmakers to force them. Just quitting won't work: your uncle will fall for violent conspiracy theories no matter how often you do or don't log on to YouTube.

Integrity design is already happening at some companies, but it needs support. Too often, companies block these teams from doing their work to its fullest when it conflicts with other company priorities, such as boosting engagement.

Over the last few years, I've had the honor of seeing a Cambrian explosion in the theory and practice of integrity (thanks, in part, to hiring sprees at companies like my former employer). I've seen the rigorous work, the deep sense of professionalism, and the high level of skill that my coworkers bring to the job. They know, in detail, how to build the social media cities that we all yearn for. Let's applaud them, let's learn from them, and let's help them do their jobs.

Despite everything, I still believe in the internet. The social city is too precious to let rot.

Sahar Massachi is the cofounder and executive director of the Integrity Institute, and a former civic integrity engineer at Facebook

by Sahar Massachi