Mobile

If Only AI Could Save Us from Ourselves

Google has an ambitious plan to use artificial intelligence to weed out abusive comments and defang online mobs. The technology isn't up to that challenge—but it will help the Internet's best-behaving communities function better.

by David Auerbach December 13, 2016

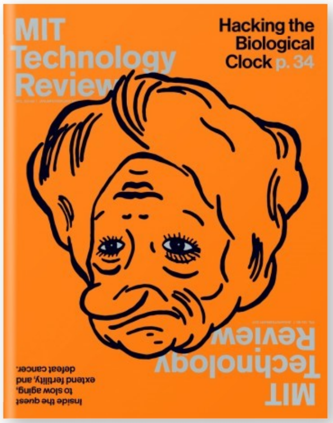
Humans have broken the Internet. Cyberbullying, harassment, social shaming, and sheer unpleasantness plague such sites as Twitter and Reddit, especially if you happen to attract the wrong sort of attention.

Consider the way *Ghostbusters* star Leslie Jones and public relations executive Justine Sacco became targets for mass abuse.

The companies that run online services are typically squeezed between charges of indifference to harassment and suppression of free speech. But now Google thinks it can use artificial intelligence to lessen this tragedy of the digital commons. (Disclosure: I worked for Google in the 2000s.) A technology incubator in the company, called Jigsaw —formerly known as Google Ideas—says it intends to spot and remove digital harassment with an automated program called Conversation AI. As Jigsaw's president, Jared Cohen, told *Wired*, "I want to use the best technology we have at our disposal to begin to take on trolling and other nefarious tactics that give hostile voices disproportionate weight, [and] to do everything we can to level the playing field."

measured by who and how?!? see Ccls

It's gutsy for Google to take this on, and it's different from some of Jigsaw's previous work. That has included Project Shield, which protects news sites and sites promoting freedom of expression against denial-of-service attacks. Another Jigsaw effort, Password Alert, is a Chrome extension that guards against phishing attacks. Those were primarily technical challenges. But fighting trolls and online mobs is also a sociological problem.



This story is part of our January/February 2017 Issue

See the rest of the issue

Conversation AI is an offshoot of one of the most successful of Google's

Thing reviewed

Conversation AI
from Google's Jigsaw

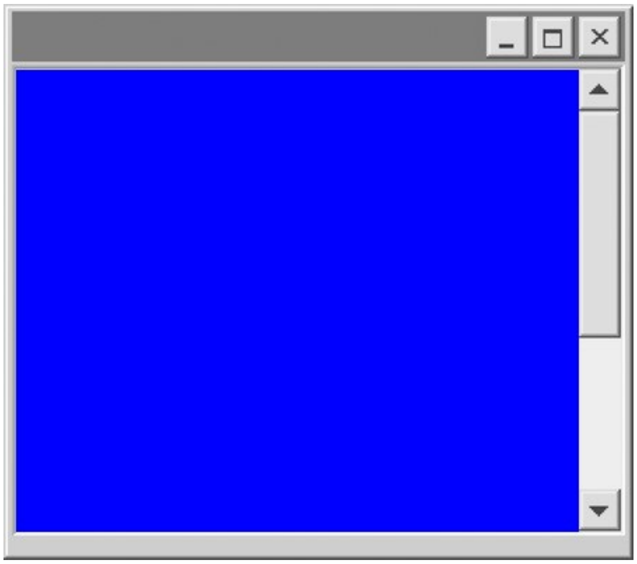
"moonshot" projects, Google Brain. It has helped revolutionize the field of machine learning through large-scale neural networks, and given Google advantages such as software that is more skillful than humans at recognizing images. But Conversation AI won't be able to defeat online abuse. Though Jigsaw's stated goal is to "fight the rise of online mobs," the program itself is a far more modest—and therefore more plausible—project. Conversation AI will primarily streamline the community moderation that is today performed by humans. So even if it is unable to neutralize the worst behavior online, it might foster more and better discourse on some sites.

Allusion detection

Jigsaw is starting Conversation AI at the *New York Times*, where it will be rolled out in a few months to help the company manage its online comments. Human moderators currently review nearly every comment published on the site. Right now, Conversation AI is reading 18 million of them, learning to detect each individual category of comments that get rejected—insubstantial, off-topic, spam, incoherent, inflammatory, obscene, attack on commenter, attack on author, attack on publisher.

The *Times*'s goal is not necessarily to reduce abuse in its comments, a problem it already considers under control. Instead, it hopes to reduce the human moderators' workload. "We don't ever expect to have a system that's fully automated," Erica Greene, engineering manager of the *New York Times* community team, told me. *Times* community editor Bassey Etim estimates that somewhere between 50 and 80 percent of comments could eventually be auto-moderated, freeing up employees to devote their efforts to creating more compelling content *from* the paper's comment sections.

The *New York Times* site poses very different problems from the real-time free-for-all of Twitter and Reddit. And given the limitations of machine learning—as it exists today—Conversation AI cannot possibly fight abuse in the Internet's wide-open spaces. For all the dazzling achievements of machine learning, it still hasn't cracked human language, where patterns like the ones it can find in Go or images prove diabolically elusive.



Illustrations by Erik Carter

The linguistic problem in abuse detection is context. Conversation AI's comment analysis doesn't model the entire flow of a discussion; it matches individual comments against learned models of what constitute good or bad comments. For example, comments on the *New York Times* site might be deemed acceptable if they tend to include common words, phrases, and other features. But Greene says Google's system frequently flagged comments on articles about Donald Trump as abusive because they quoted him using words that would get a comment rejected if they came from a reader. For these sorts of articles, the *Times* will simply turn off automatic moderation.

It's impossible, then, to see Conversation AI faring well on a wide-open site like Twitter. How would it detect the Holocaust allusions in abusive tweets sent to the Jewish journalist Marc Daalder: "This is you if Trump wins," with a picture of a lamp shade, and "You belong here," with a picture of a toaster oven? Detecting the abusiveness relies on historical knowledge and cultural context that a machine-learning algorithm could detect only if it had been trained on very similar examples. Even then, how would it be able to differentiate between abuse and the same picture with "This is what I'm buying if Trump wins"? The level of semantic and practical knowledge required is beyond what machine learning currently even aims at.

Consequently, a dedicated Twitter troll will no doubt find a novel way of expressing abuse that evades a system like Conversation AI. By blocking some comments, machine learning could do a decent job of getting commenters to stop casually calling each other “fags” and “homos,” if that’s the goal. But machine learning will not be able to foil a person hell-bent on insinuating that someone is queer.

In other words, Conversation AI will enable moderation tasks to be executed more efficiently in communities that already tend to be pretty well behaved. It is incapable of rooting out the worst of the abuse we hear about, which frequently shows up on sites with minimal moderation standards. Policing abuse on Twitter and Reddit is impossible without fundamentally altering the nature of those platforms.

Gated communities

Facebook’s success is a reminder that most people, and certainly most companies, prefer a relatively sheltered and controlled environment to one where strangers can intrude into others’ business and start fights. So if Conversation AI or similar tools make it easier and more efficient to exercise such control, it’s a reminder that “solving” the abuse problem, whether through human or automated means, requires moving away from maximal inclusivity as the highest ideal online. Even seemingly “open” communities such as StackExchange and MetaFilter require constant moderator intervention and community policing. Truly anarchic communities, such as Twitter, 4chan, and some channels on Reddit, prove to be the exceptions online, not the rule. Nor are anarchic communities moneymakers. Twitter has had trouble attracting a buyer, partly because of its reputation for abusive content, while Reddit has had a high degree of staff turnover and difficulties monetizing. The Wild West nature of those sites will become only more apparent if tools like Conversation AI make moderated sites function even better.

Policing abuse on Twitter and Reddit is impossible without fundamentally altering the nature of those platforms.

Can online discourse get better?
Weigh in.

It’s worth noting one big potential downside. Because Conversation AI is being trained to approve content that hews to certain lexical, grammatical, and stylistic guidelines, it won’t just filter out abusive content. It could also tend to strike *diverse* content. That raises questions of what censorship-minded governments could do with it. Just as the *Times* curates its communities, so too can the governments of Turkey and China curate theirs. While Jigsaw efforts like Project Shield aim to provide defenses for politically sensitive websites, Conversation AI makes it easier to filter out unwanted speech—but the question is, unwanted by whom? There is no label on the box that says, “Use only to prevent abuse.”

!!!!!!!
!!!!!!!
!!!!!!!
!!!!!!!

David Auerbach is writing a book on human and computer languages and their convergence, to be published by Pantheon. He worked for 11 years as a software engineer at Google and Microsoft, primarily in server infrastructure.