Toda Peace Institute

Policy Brief No. 45 **July 2019**

Polarisation and Peacebuilding Strategy on Digital Media Platforms:

Current Strategies and their Discontents

Lydia Laurenson

Abstract

This is the second of two policy briefs on polarisation. The first policy brief, "Polarisation and Peacebuilding Strategy on Digital Media Platforms: The Current Research," reviewed the research, and concluded by recommending directions for future research. This brief describes interventions: (a) interventions currently being attempted by NGOs and other peacebuilders using digital platforms as their medium, and (b) interventions that the platforms themselves have tested and/or put into action. The conclusion of this brief sorts interventions into categories, and provides recommendations for digital media platforms.

Author Note:

Any opinions, biases, and/or mistakes in this policy brief are my own. I would also like to credit some of the people I spoke to, in alphabetical order: Zahed Amanullah of the Institute on Strategic Dialogue; Alisha Bhagat of Forum for the Future; Kelly Born of the Hewlett Foundation; George Davis; Renee DiResta; Shauna Gordon-McKeon, independent researcher; Helena Puig Larrauri of Build Up; Rachel Lazerus, independent researcher; An Xiao Mina; Jonathan Stray of Columbia University; Aviv Ovadya; and others who asked not to be named.

Interventions by Independent Organisations Operating on top of Digital Platforms

Many organisations and individuals are working on digital peacebuilding, while a larger number explore broad societal impacts of digital media and the research questions identified in the first brief. This policy brief focuses on peacebuilding — but "peacebuilding" touches on larger concerns. In the second part of this brief, the overview of recent platform changes takes a broader societal perspective.

Two of the initiatives below are working specifically on "violent extremism" (though not all of them are). In that context, it is worth noting that many Countering Violent Extremism (CVE) programmes have attracted criticism for racism and suppression of free speech. The

Brennan Center for Justice, a US-based law and policy institute, has a resource page that describes racism against Muslim communities as an ongoing problem in US CVE programmes (despite the fact that right-wing and anti-government extremists kill far more people than jihadists in the US, as summarised in this 2015 article from the New York Times). Lisa Schirch of the Toda Peace Institute published a book last year, *The Ecology of Violent Extremism,* which explores the negative impact of counterterrorism and CVE programmes and offers peacebuilding approaches.

The two relevant programmes described below make an effort to work on the threat of white supremacy as well as other extremist movements, and also do not use race as a criterion for identifying threats.

1. The Commons: Depolarising Conversation on Facebook and Twitter

This 2018 Medium post is a good place to start learning about The Commons, which is a project of Build Up, an international nonprofit based in the USA. First, they tested a number of strategies for peacebuilding on social media, and now they are scaling the most successful one. They use automation to surface polarised conversations on social media, and they have developed a methodology for human moderators to step into those conversations and depolarise them. Their success metric was whether they could get people to "reflect on the way they were engaging on social media."

This approach requires a great deal of human effort, of course, and that is a potential barrier to scaling the intervention. But at least using automation to find polarised conversations is efficient. It is hard to imagine how a similar approach might have worked before digital media.

Build Up's Helena Puig Larrauri told me that she and her team had a "steeper learning curve" engaging the political right than the left. The organisation did not come out of a politically conservative community, so had to put in extra effort to find facilitators who could engage with conservatives. Larrauri also made the point that the language of "polarisation" is associated with liberal values, and said that talking about "civility" helped with more right-oriented people. As they scale their process, The Commons is running a series of local workshops where they ask about language and wording.

The Commons has tested interventions on both Facebook and Twitter, though Larrauri suggests that neither platform is a great container for dialogue. This, of course, is part of why facilitators are needed: They model a type of conversation that the platforms' affordances do not offer.

On the other hand, there are scale advantages to these public conversations, especially on Facebook, where conversations are usually focused around one post with lots of people commenting beneath it. Given this architecture, Larrauri says, lots of people will see the conversation, and so facilitators can shift the entire conversation by setting an example for the group, particularly when engaging with very aggressive people.

2. The Institute on Strategic Dialogue: Reaching Violent Extremists with Video

The Institute on Strategic Dialogue is a UK think tank that has been testing videos designed to convince violent extremists to leave their movements. Sometimes, these videos link extremists with emotional support afterwards.

For example, in partnership with an organisation called ExitUSA, ISD helped former white supremacists in the US create videos about their disenchantment and post them on Youtube, with contact info at the end of each video. When white supremacists contacted ISD, each was put in contact with a former white supremacist who encouraged them to leave the movement. ISD summarised this pilot study and others in a 2016 white paper; one of their most interesting indicators of success was that eight white supremacists reached out to ExitUSA, asking for help leaving the movement.

Not all ISD video initiatives include conversation with a human, but the ones that do are clearly similar to efforts by The Commons — both use digital scale to identify specific people with whom a meaningful intervention is possible. However, other initiatives simply create and distribute anti-extremist video.

I spoke to ISD's Zahed Amanullah, who said that in their outreach, they looked as carefully as they could and found only 1-2% of the population are in the desired audience for these anti-extremist videos. Of those, Amanullah said, only an extremely small number would be a threat. When targeting an audience, ISD narrows their scope based on people's use of memes, terminology, etc. (Amanullah noted that they generally ignore race and "hope to be colorblind," in the sense that other information about a person — like how the person uses memes — is more useful for identifying potentially violent extremists than their race.)

ISD has now been tasked with figuring out a definition of extremism for the UK, because the country has a new commission for violent extremism. But, as Amanullah pointed out, "This is like answering the question of what is an acceptable level of environmental protection — it is arguably a question that society at large must answer, not any individual or any state. Perhaps at some point, there will be an equilibrium of discourse, where crossing the line is obviously extreme. But right now, everyone is testing that boundary — speaking in coded language that has meaning to some people, and so on."

Amanullah went on, "The scale and speed of the problem are overwhelming for everyone trying to address it. In the US, there is free speech as a modifier [to the culture]. In places with hate speech laws, such as Germany, where the law says that Facebook and Google have to take down extremist content when informed, those companies are trying to grasp the scale of the problem and make good decisions about what gets taken down. But everything is really fast-moving and there is not an official definition of extremist content."

3. Moonshot CVE (and Google's Redirect Method): Redirecting People who Search for Extremist Material

Jigsaw, an R&D organisation within Alphabet (Google's parent company), developed an open source methodology called Redirect Method after hosting a 2011 conference about violent

extremism. Redirect Method identifies pre-existing videos that could sway violent extremists, and then puts those videos in front of people who search Google for keywords that lead to extremist content. In other words, it attempts to redirect users to credible voices whose videos challenge the arguments of violent extremists.

The methodology Jigsaw developed was eventually adopted by Google. A startup called Moonshot CVE then partnered with Google and is deploying Redirect with independent funding. This 2018 case study from RAND Corporation puts Redirect Method in context of other digital efforts to oppose violent extremism, including ISDs.

The most important takeaway from the RAND case study is that many peacebuilding initiatives on social media measure their impact largely (or solely) by counting the number of people who saw and/or reacted to the posts, tweets, videos, etc., and that more in-depth evaluation measures are needed.

For Redirect Method specifically, the case study offers suggestions for more in-depth evaluation strategies, like administering surveys to extremists exposed to Redirect Method to see if they are less radicalised as a result. Many Redirect Method efforts seem to perform well when measured by social media reactions, but it is hard to know what those reactions translate into off the internet.

4. Other Relevant Efforts

There is quite a bit of relevant work happening among well-regarded researchers and nonprofits. Most of these efforts are not specifically about depolarisation or peacebuilding, but here is a sample of relevant stuff:

- The Electronic Frontier Foundation has advocated for privacy and free speech on the Internet since its founding in 1990. In the context of peacebuilding, the EFF offers an excellent blog post about non-censoring approaches to these problems, "Private Censorship Is Not the Best Way to Fight Hate or Defend Democracy: Here Are Some Better Ideas," by Corynne McSherry, Jillian C. York, and Cindy Cohn. They make critically important points about power and privilege — how and why we *must* preserve free speech as we design interventions.

- Another organisation in the space is Data & Society, which creates reports about the dynamics of different digital media platforms. Their 2018 white paper "The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators," by Whitney Phillips, reviews the digital influence landscape and offers tips to journalists and other thought leaders reporting on extremism, in order to give minimal support to extremism via exposure.

- Julia Kamin and J. Nathan Matias oversee research at the US nonprofit CivilServant, which "works directly with online communities to test ideas in moderation and evaluate the impact of the tech industry in our social lives." Explore their research here.

- Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker co-founded the Social Media and Political Participation Lab at New York University, which "studies the effects of social media on politics." Their research is detailed here.

- At MIT Media Lab, Ethan Zuckerman and his team are building a social media filtering platform called Gobo, which enables users to pull in social media updates from many platforms and gives them very fine-grained control over their filters (an example he gives is that a user could choose to see only posts from non-men). He has a blog post about that here. Since users have such fine-grained control over what filters get applied to their experience on this platform, it could potentially be used to run experiments on how filters shape the experience of digital media users.

- At the Center for Media Engagement at UT Austin, Natalie (Talia) Jomini Stroud and her team have conducted tests relevant to platform affordance design, such as the 2013 experiment described in the policy brief on polarisation research that found a "Respect" button could depolarise digital discourse when compared to the more common "Like" button. (That experiment was published several years before Facebook switched from the "Like" button to six different "emoji reactions," in 2016.) The rest of the Center's research is here, though it is more focused on digital journalism than platform design.

Strategies Attempted By Digital Platforms Themselves

Although no major tech companies list "peacebuilding" among their goals, some are interested in depolarisation. Of course, they are all driven primarily by economic incentives, but some have other goals too — and their values can be genuinely different from each other, which may drive different behaviour.

Additionally, whether or not they are interested in peacebuilding, the moral panic of the last few years has driven significant change at the major platforms. Here are some of the things they have tried.

1. Facebook

Facebook is the most-discussed company of recent years. They have long had a reputation of "moving fast and breaking things," even compared to the rest of Silicon Valley. (In fact, the "move fast and break things" slogan was originally a Facebook slogan, though it has become associated with the tech industry in general.) But now that the company routinely affects large-scale civil society and politics, they seem to be working to slow down and show foresight.

As with all the internet giants, many people at Facebook are genuinely trying to do good. Many of the problems are genuinely difficult to understand and solve, and many of their lessons could be useful to other (or future) platforms. This policy brief details Facebook's efforts more closely than those of other companies, because everyone asks about Facebook. While this brief includes some criticisms, these problems are not unique to Facebook.

Real Name policy

Facebook has long required real names on the platform, sometimes claiming that this policy increases civility. Indeed, for a while, it was fashionable for pundits to insist that all internet platforms ought to do this — yet there has long been opposing evidence. This 2012 TechCrunch article by Greg Ferenstein points out, "there's surprisingly good evidence from South Korea that real name policies fail at cleaning up comments. In 2007, South Korea temporarily mandated that all websites with over 100,000 viewers require real names, but scrapped it after it was found to be ineffective at cleaning up abusive and malicious comments."

In other words, there is no solid evidence base that suggests that real name policies actually improve online discourse (but real name policies may serve other purposes, such as improving the company's ability to build data-rich ad targeting profiles). This may be one of the least well-supported interventions any major platform has tried.

Ad transparency

In 2018, after coming under fire for hosting non-transparent advertising (which is especially worrying in a world drowning in propaganda and false news), Facebook made it possible to track all the ads a given Page is running when you look at the Page. This was hailed as a good first step, though many commentators felt it was not nearly enough. Ben Thompson at Stratechery, a comparatively tech-friendly blog, was critical when he wrote that "These ads can still only be seen by going to the actual pages, which are impossible to know about unless you are shown an ad; the company should have a central, searchable, repository of all those hundreds of millions of ads."

"Community Standards" transparency

For years, it was difficult for people outside Facebook to learn about the company's "Community Standards," the internal guidelines that moderators follow to decide which content and accounts to remove. This non-transparency did not go over well with the public, so in 2018, Facebook made the Community Standards public (including the standards of what counts as sexual content, fake accounts, child exploitation, inappropriate violence, and other stuff that the company will censor). In early 2019, Facebook announced that it is building an outside "Oversight Board" to help decide its most challenging content moderation cases, but it is not yet clear what this Board will look like.

Facebook also works with outside organisations to stay on top of fast-moving areas such as hate speech, where many people use coded language — i.e., the words that count as hate speech often change. These partnerships seem good in the sense that they add transparency and accountability to the process, but some critics say the platform is asking civil society partners to do work that Facebook should be taking on itself.

Content moderation processes, at the implementation level

Aside from the transparency (or lack thereof) of their guidelines, Facebook has also come under fire for outsourcing most of the content removal to consulting companies whose employees are far less privileged than Facebook's core employees, as detailed in a 2019 Verge article by Casey Newton and 2018 documentary "The Cleaners" by Riesewieck and Block. These content moderators report both significant personal trauma (from looking

nonstop at horrifying images) and gaps in the moderation process (gaps that are hard to report to Facebook because the moderators are not "real" Facebook employees). But excitingly, in mid-2019, Kate Klonick reported in Slate that Facebook has taken these concerns to heart and is setting a new bar for how content moderators are treated.

Removal from Facebook ("deplatforming")

Facebook has been removing fake accounts from Facebook for quite a while and has enforced its "real name policy" as well (i.e., even if an account is not fake, it can still get taken down if it is not attached to the user's "real name"). Recently, it has come under more pressure — and has shown more willingness — to remove or "deplatform" real accounts. For example, in early 2019, The Guardian published a piece by Julia Carrie Wong on how Facebook has chosen to "ban four ethnic armed Myanmar-based groups from its site," all of which oppose the Myanmar government that is currently in place. (Indeed, although the industry historically was extremely reluctant to censor anything, Silicon Valley has been tilting more towards censorship than it once did — sometimes in problematic ways, as the Electronic Frontier Foundation points out.)

While some observers praise Facebook for banning distasteful figures like American white supremacists, it is not clear how (or whether) Facebook will create a consistent, transparent, and just policy for who gets deplatformed.

Removal of creator income ("demonetisation")

In 2018, Facebook began rolling out tools to enable users to send money to their favourite creators. (For example, a user could send money to an artist who posts videos on Facebook.) While rolling out those tools, Facebook also announced a policy about which content is eligible for monetisation. This means that a creator who earns income from Facebook can have that income taken away if they violate Facebook's monetisation rules: they can be "demonetised."

Facebook demonetisation seems uncontroversial so far, perhaps because Youtube has already been monetising and demonetising content for years. Advertisers have also insisted on not showing their ads next to controversial content since time immemorial. Both these factors mean that Facebook did not have to develop brand-new policy or procedure for this. (On the other hand, it took several years for Youtube's demonetisation policies to occasion blowback, as described by Peter Kafka in Recode back in 2016, so there may be blowback yet to come.)

Political labeling

In 2017 and 2018, Facebook created tools called Town Hall and Candidate Info. These are clearly delineated spaces on the platform where candidates are labeled as such, and users exposed to a candidate message have the option to look at the opposing candidate's message. (The labeling is reminiscent of Reddit's "flair" concept, though it appears to be less driven by human moderators.) These spaces are intended to make political maneuvering more transparet and have attracted little criticism so far — although it will be hard for Facebook to decide who gets designated as "legitimate" politicians in less stable countries like Myanmar or Cambodia.**Fact-checking partnerships**

Facebook works with outside partners to fact-check news on the platform, and this effort

has been widely criticised. As with the outside partners on Community Standards, some critics say that Facebook should be doing (and paying for) more of the work internally, not forming partnerships to do it. There are also specific details of the programme's construction that lead to tensions; all the fact-checking partners are news outlets, which means that a partisan news outlet may find itself fact-checking another news outlet (i.e., a competitor) with opposing political views, as described by Laura Hazard Owen in 2018 in the journalism trade publication Nieman Lab.

Overall, it is hard to tell if there is enough scale and speed in these fact-checking efforts to matter, as there is little external information on how well the programme is working. Still, there is more information about the programme in Mike Ananny's 2018 Columbia Journalism Review article.

Labeling and ranking publisher content

Facebook has also experimented with labeling publishers who promote their content on the platform, including journalism companies. In 2018, Facebook introduced a Context button, which makes it easy for users to learn more about a publisher when they see a piece of news on Facebook. (At the time, the design team that developed the Context button wrote an admirably transparent Medium post about the process behind the feature.)

Facebook has also tested labels for forwarded messages on WhatsApp, a messaging service, as reported by India telecommunications blog Medianama. Specifically, they are trying to use labels to make WhatsApp users more aware that messages forwarded to them are not actually from the individual who forwarded the message.

All these labels seem relatively uncontroversial. In contrast, another Facebook initiative to create "trust scores" for publishers caused controversy, because publishers with higher trust scores reportedly get more traffic to their articles than publishers with lower trust scores, and (like everything else about Facebook's Newsfeed algorithm) the method by which trust scores are calculated is not public.

Focus on relationships

Partly in response to criticism about spreading fake news, Facebook announced in 2018 that it would decrease the prevalence of articles in the Facebook Newsfeed, and increase content posted by friends. Journalists criticised this move, and it is not clear that anyone else noticed.

Viral slowdowns

WhatsApp, a messaging product owned by Facebook, was an early testing ground for viral slowdowns — i.e., slowing the speed on viral messages by limiting the number of people that a message can be forwarded to. Previously, a single message could have been forwarded to hundreds of people by one person; now it can be forwarded to only five.

In 2019, Jasmine Garsd reported in NPR that the feature, which was tested previously in India, is now going global on WhatsApp. In theory, this could help with problems where people might behave better if they had time to reflect — like slowing down a lynch mob. On the other hand, it could also slow down a message about a genuine disaster — like people at sea warning people on shore about an incoming tsunami.

It is not clear whether Facebook has implemented a version of viral slowdown on Facebook itself. The dynamics of Facebook's Newsfeed are less visible because of its black-box algorithm, so they may have tested or implemented viral slowdowns there, or they may not.

Affordance redesigns

In 2016, Facebook officially switched from the "Like" button to six different "emoji reactions." Reportedly, the feature was driven by users who wanted to be able to express emotions like "Love" or "Sadness" when a Facebook friend posted about a death in the family, as opposed to a "Like." However, in a 2017 Forbes article, Amit Chowdhry reported that the different emoji reactions are weighted differently in the Facebook Newsfeed, meaning that different reactions affect which conversations are visible to users. And outside research by Stroud et al. has shown that changing the available reaction buttons can affect the polarisation of a digital conversation.

It is almost impossible for outside observers to know what is happening with Facebook's affordance experiments because almost none of the data is public. But we do know that every major tech company runs (literally) thousands of experiments on how affordances affect user behaviour. If polarisation or peacebuilding are priorities at Facebook right now, then they are now experimenting to figure out how affordances can depolarise or build peace.

2. Google and Youtube (owned by Alphabet)

In recent years, Google and Youtube's impact on society have been less examined and debated than Facebook's. But — at least in the US — both Google and Youtube are the only websites on the open Internet more visited than Facebook.

They are also far more trusted than Facebook. A 2018 survey by data company Jebbit, profiled in the marketing trade publication Adweek, found that of 100 major consumer-facing brands, Facebook was *least* trusted in the US while Google was the third *most* trusted (after Amazon, second most trusted, and Visa). This got backed up by a separate survey conducted by Georgetown and New York University, described in Vox — Facebook was still last. (That survey also found that Google is less trusted by Republicans than Democrats.)

Google's original slogan was "Don't be evil" (though the company stopped emphasising that slogan a few years ago). Some critics say that consumers still largely trust the company due to an internal culture of foresight and care. One article by Lily Hay Newman in Wired describes an early "privacy summit" that took place at Google in 2009, during which Google employees worked to anticipate future privacy issues and to build better tools to honour users' privacy. Newman then documents the summit's impact and the company's privacy-honouring culture over the next ten years. (On the other hand, Newman points out that Google's business model — which, like Facebook, is advertising — has still pushed them towards privacy missteps.)

Have these different company cultures resulted in different societal impacts? Google appears to be ahead of Facebook on some problems; for instance, while Youtube has many of the same content moderation issues as Facebook, Youtube's moderation guidelines were always public, whereas Facebook only made its public last year after significant public

pressure. An article by Russell Brandom in The Verge notes that in 2018, a coalition of non-profit groups came together to create the "Santa Clara Principles on Transparency and Accountability in Content Moderation," and Youtube's long-standing policies were closer to complying with the principles than Facebook's.

Examples like these may argue for the impact of different company cultures. Again, though, the companies have similar business models, so it could also be argued that the tech giants are more similar than they are different. Youtube certainly has some of the same problems and has taken many of the same actions as Facebook, including deplatforming and demonetisation. Along with other "recommendation engines" like Pinterest, Youtube has been particularly criticised for failing to moderate what is going viral in its automated recommendation systems. A 2018 article (and mantra), "Free Speech Is Not The Same as Free Reach," is one of many relevant pieces by Renee DiResta in Wired magazine. Around the same time, Zeynep Tufceki dubbed Youtube "The Great Radicalizer" in a New York Times op-ed. Youtube recently announced that they are changing their recommendation algorithm in response to these criticisms.

In some sense Google, too, can "deplatform" websites from its search results, although the word means something different with Google because the search engine almost never hosts content. If Google chooses not to offer a search result, that website still exists and can be accessed directly, whereas when Youtube or Facebook deplatforms content, the content cannot be accessed directly (though it might be re-uploaded by someone else). Hence, a better verb for Google removing websites from search results might be "deindex."

Google uses multiple forms of deindexing. Firstly, it can fully deindex a result: For instance, Google normally removes porn from search results entirely (unless the user turns off Google's SafeSearch feature), and it also removes search results if it receives notice that those results violate copyright.

Secondly, it can deindex the result from its Autocomplete feature. Autocomplete normally activates when users are typing a search query, but Google's policy states that it does not show predictive text related to hate speech or violence (including self-harm).

Thirdly, Google can impose penalties. For instance, a "-50 search penalty" moves a website 50 slots in search results — so if it normally would be the third result in a given search, instead it becomes the 53rd result. (The marketing website SearchEngineLand offers a guide to these sorts of penalties and how they are levied.) Usually, these penalties are issued due to technical problems with the website, or because the website's creators tried to game Google's search algorithm in violation of Google's policies.

Google penalties seem rarely used, if at all, for concerns related to the content itself, like extremism or false news. However, there has been an uptick in full deindexing on other platforms based on these concerns. In early 2019, Christina Caron at the New York Times reported that Pinterest has deindexed vaccine-related content in response to anti-vaccine activists. Although Google has not taken the same approach, it is worth understanding that Google has these capacities, and also has pre-existing policies for using them.

To return to the Facebook comparison, a bigger difference between Google and Facebook is

that Facebook's primary business interest is keeping users engaged with Facebook, whereas Google's primary business interest is locating the best resource the user is searching for on the open internet. This difference means that Google is incentivised to support the open internet. Thus, the company has developed many tools intended to keep the open Internet healthy, including free technical improvements for websites operated by outside organisations, and other tools that are open to the public.

In 2013, for example, Jigsaw launched Project Shield, a free service to help human rights organisations, journalism organisations, and others protect their websites from malevolent attacks. In 2018, Google launched a search engine for fact checks, which got written up in the journalism trade publication Poynter by Daniel Funke. Funke quoted a Google research scientist stating that "it's possible that Google will add more signals to its algorithm to surface more fact checks in search" once the fact-checking engine is out of beta — so Google might start actively fact-checking outside websites when users turn them up in Google search. But for now, the new project simply helps users to fact-check any topic they are interested in, if they have the idea to fact-check it on their own. (Like Facebook's fact-checking efforts, Google's appear to be reliant on outside civil society partners.)

If Google starts surfacing fact-checks in Google search, it may solidify a new direction for the company. Like Facebook and all the other major platforms, Google spent many years seeking to avoid the direct arbitration of truth on its platform. But it is unclear that such arbitration can be dodged.

For instance, Google's "featured snippets" — text that appears directly in Google search — are pulled directly from outside websites, which means the snippets are often wrong. Yet those snippets bear the authority and trust of Google's brand because they appear directly in Google search. Journalists like Sarah Perez at TechCrunch wasted no time criticising this feature, and Google's first response was to introduce "multi-faceted" featured snippets that show multiple answers from many websites at once (which, of course, could all be incorrect). If Google does add fact-checking to search, that will be a more direct intervention, and a more direct acknowledgment of responsibility.

Another example from 2017-2018 is a project called PerspectiveAPI, built by researchers at Jigsaw, an R&D organisation that used to be part of Google and is now owned directly by Alphabet. PerspectiveAPI is an attempt to use artificial intelligence to make Internet comments less toxic.

PerspectiveAPI started with the development of automated "toxicity scores" written up in this 2017 Medium post by Jigsaw researchers. Then, in 2018, Jigsaw partnered with Oxford's Rhodes Artificial Intelligence Lab to experiment with integrating those scores gently into comment sections. In another Medium post, the scholars talk about adding "speedbumps" to toxic comments — perhaps similar in spirit to the "viral slowdowns" Facebook has been experimenting with. These speedbumps operate by intervening *before* a user posts a toxic comment and asking the user to tone it down, perhaps with a question like: "Your comment is likely to be perceived as condescending by other users, and will be assessed by a moderator. Do you wish to rephrase it?"

Much of this is publicly documented, and PerspectiveAPI has also been implemented

publicly by outside partners. One partner, the Coral Project, is an open source comment moderation project that originated as a collaboration between the New York Times and the Washington Post.

Redirect Method (mentioned already when discussing MoonshotCVE) followed a similar trajectory. It is a methodology originally developed at Jigsaw, after they helped convene a 2011 conference about violent extremism, which has now been implemented by outside partners such as MoonshotCVE. Redirect Method attempts to redirect people who search for extremist content to credible voices whose videos challenge the extremists' arguments, and was assessed in depth by RAND Corporation in 2018.

3. Brief Notes on Other Platforms: Twitter, Amazon, and Reddit

Like Facebook and Youtube, Twitter has deplatformed users, often after public outcry, though in some high-profile cases they have resisted doing so longer than the other two companies. The company used to have an internal slogan saying they were the "free speech wing of the free speech party." They also got a lot of praise for — and drew a lot of encouragement from — supporting the Arab Spring soon after the company was founded, so it is no surprise that Twitter employees feel honour-bound to support controversial voices.

Yet although Twitter has generally considered itself pro-free-speech, there has always been content that violated its rules and required deletion. Twitter has long faced the technical challenge of moderating content at unprecedented scale. Their specific process is detailed in a 2018 Logic Magazine piece by Tarleton Gillespie, and has many of the same problems as Facebook's.

The main difference between Twitter and other social media platforms is that it has a much bigger harassment problem, especially for high-profile users and marginalised users. The human rights organisation Amnesty International is the latest of many to release a 2018 report about Twitter being toxic for women. Because the harassment problem is a constant burden on high-profile users, high-profile users have spent years asking for more effective harassment management tools. A 2018 Bloomberg op-ed by Noah Smith is a good example of an article that not only describes the problem, but asks Twitter to work on very specific features in response. For instance, Smith suggests that "users should be able to lock individual tweets, closing them to replies." (It is currently possible to lock tweets, but only by locking the entire account.)

Smith, who is based in San Francisco and writes about technology, also notes that when he talks to Twitter employees around town, "what frightens them most is the idea that Twitter might be used to create echo chambers, where like-minded people are not exposed to contrary viewpoints." But the idea of social media echo chambers is widely believed to be debunked — although it is possible that we do not know enough about Twitter, specifically, to reach conclusions about its particular dynamics.

Early in 2018, Twitter put out a request for research proposals about "conversational health metrics." The two proposals they accepted focus on (1) examining echo chambers, and (2) bridging gaps between communities, which supports Smith's claim.

Despite the use of the debunked echo chamber argument, the researchers are trying to include harassment in their assessment. The description for the first proposal notes that the group already "found that while incivility, which breaks norms of politeness, can be problematic, it can also serve important functions in political dialog. In contrast, intolerant discourse — such as hate speech, racism, and xenophobia — is inherently threatening to democracy. The team will therefore work on developing algorithms that distinguish between these two behaviours." It is true that Twitter serves a unique function in modern public discourse, so hopefully this research will help them do that better.

In mid-2019, Twitter also released an app, "twttr," where it publicly experiments with different affordances. Twitter gave Buzzfeed News reporter Nicole Nguyen access to the twttr team, and she wrote up their process. In doing this, Twitter is arguably following in the footsteps of Reddit, which gave extensive behind-the-scenes access to another reporter in 2018 and thereby raised the bar on platform transparency (more about this below).

Regarding other platforms: Amazon is worth a mention because the public trusts it enormously, and it controls the movement of both information and goods. The survey that found Google to be less trusted by Republicans than Democrats also found that Amazon is one of the most trusted across the US, by both parties. Yet Amazon can and does remotely delete books from users' Kindles, and experiments with which books it allows users to see. And Amazon's empire does not end at consumer sales — it owns the server space that powers almost half the internet, as documented by Vox. That places Amazon in an infrastructure role. It is worth thinking about how much the public should trust it.

Reddit, too, is worth a mention. It is firmly among the most-used websites on the Internet. As of this writing in March 2019, it is the sixth most-visited website in the U.S. according to Alexa's real-time rankings — that is after Google, Youtube, Facebook, Amazon, and Wikipedia, and *before* Twitter. Yet Reddit attracts remarkably little attention despite its impact.

An important difference between Reddit and the other major platforms is that its structure depends on volunteer moderators, who shoulder an immense amount of work to manage conversation on Reddit. There is more about the platform's history and values in this 2019 video interview with CEO Steve Huffman, which includes commentary on Reddit policies like the platform's continued proud support for pseudonyms.

In 2018, Reddit took the unprecedented step of giving New Yorker writer Andrew Marantz access to its internal moderation meetings, including all the company snacks and awkward decision-making. No other platform company has been willing to risk such transparency. (As Marantz wrote in the article, "I asked a few social-media executives to talk to me about all this. I didn't expect definitive answers, I told them; I just wanted to hear them think through the questions. Unsurprisingly, no one jumped at the chance. Twitter mostly ignored my e-mails. Snapchat's P.R. representatives had breakfast with me once, then ignored my e-mails. Facebook's representatives talked to me for weeks, asking precise, intelligent questions, before they started to ignore my e-mails.")

Reddit's transparency was both admirable and, in some sense, risky: The reporter was present when Reddit employees learned that, although they had just banned a forum called

SexWithDogs, they missed another forum called DogSex. Undeniably, transparency will reveal that sort of thing. Equally, such embarrassments are trivial given what is at stake.

Recommendations: Intervention Categories and Their Costs and Benefits

All the interventions discussed above can be sorted into broad categories. The list below recommends a set of categories, sorted by how aggressive they feel to their subjects, from least invasive to most invasive.

1. Outside Support

The platform creates or sponsors an intervention that does not affect user experience on its own platform but is relevant to its ecosystem. Examples include Google's creation of the separate fact-checking search engine (which may be integrated into its search results but has not yet), or Jigsaw's creation of PerspectiveAPI, which helps content-creation organisations outside Google moderate comments on their own websites.

2. Custom Interventions

After investing in careful research, a platform may devise an intervention that is not easily described in general terms. If you are looking for more to read when you finish this brief, try Sarah Jeong's 2015 book *The Internet of Garbage;* she reviews these issues in-depth and concludes that the best interventions are highly specific to the platform and are developed after deep research, often in partnership with outside researchers or users. For example, Jeong tells the story of an online game that had a significant harassment problem, which largely took place over direct messages. The game changed direct messages so that users had to consent to receive them from other users, and its harassment problem largely disappeared — an excellent example of a custom intervention that operated without invading users' space or silencing public speech. (This intervention required specialised knowledge to implement because direct messages worked differently in the game, and had unusual dynamics, compared to other platforms. For example, most harassment on Twitter does not take place in direct messages, so a similar intervention would not work there.)

3. Label

The platform adds clear descriptions or context labels to items on the platform, like content, usernames, or profiles. Examples include Facebook's Context button on news items, or Facebook's labels that describe a political candidate's party and candidate status. As long as labels are developed respectfully, they are a good way to improve users' awareness of the broader context without restricting speech.

4. Re-Design

The platform changes affordances to achieve a certain result. Examples include the Stroud experiment of trying a "Respect" button rather than a "Like" button, which yielded less polarised conversation. Many of these examples can change discourse, sometimes significantly, so it is important to have a sense of guiding values while making those changes: For example, would we prefer a less polarised discourse?

5. Moderate the Conversation

Careful, humane, personal conversation and moderation may be the most important strategy available for peacebuilding purposes. Examples of conversation moderation from

this brief include the conversational interventions designed by Build Up and the Institute for Strategic Dialogue. However, moderation is not usually perceived as a technical intervention — or an easily scalable one. It is also highly variable, and its results depend a lot on the personalities and goals of the moderators. The major platforms invest very little in conversational moderation, with the exception of Reddit, whose structure depends on empowering thousands of community moderators.

6. Derecommend

The platform continues to host the content, but restricts or slows down its reach. Examples include Facebook's viral slowdowns, or Youtube's changes to its recommendation algorithm such that conspiracy theories are recommended less. Renee DiResta's concept that "Free Speech Is Not The Same as Free Reach" is useful here: Promoting content by recommending it or enabling it to go viral is arguably a stronger action than merely allowing the content to exist. On the platform level, recommendations are perhaps the clearest analogy to the "editorial judgment" that an editor at a newspaper or magazine would exercise. In this sense, it seems that recommendation and derecommendation should be left largely to a given platform company's preference, just as editorial judgment is left to editors.

7. Demonetise

The platform continues to host the content, but does not allow it to make money on the platform. Facebook and Youtube both do this to content that violates their monetisation guidelines, and it seems best considered as a different form of "derecommending" — or editorial judgement — even if we do not agree with all the content they decide to recommend or monetise (or derecommend or demonetise).

8. Debunk

The platform continues to host the content, but shows users an alternative narrative. Examples include Jigsaw's Redirect Method. This is relatively uncontroversial when applied to topics like showing anti-extremist videos when users search for extremist keywords. But it would be far more controversial and problematic if Google employees decided to debunk, say, the concept of God in Google search.

9. Deindex

The platform makes it hard or impossible to search for the content, but continues to host the content (assuming it hosts content — Facebook, Pinterest, and Youtube all host content, for example, but Google generally does not). Examples include Google restricting autocomplete on queries that could lead to hate speech, or any platform making it impossible to search for certain queries. Deindexing and banning (below) start to edge into truly uncomfortable territory for a democracy.

10. Ban

The platform takes down the content. All content-hosting platforms have ways to do this, but not all their policies are the same. While, to some extent, this could be considered similar to editorial judgment, it becomes questionable once a platform is so widely used that it is practically a utility. For example, if a trans person's Facebook profile is taken down because she is not using her birth name on her profile (and it is therefore in violation of Facebook's "Real Name policy"), then she could plausibly lose her chance at getting a job — employers are increasingly asking to see Facebook accounts as a condition of employment, as the

entrepreneur and political organiser Maciej Czeglowski pointed out in his excellent 2019 Senate testimony about digital privacy.

In general, while banning or deindexing may be tempting to "keep the peace," suppressing speech entirely is unlikely to build peace. Banning and deindexing should be treated as extreme options, used with enormous care and transparency. (The Electronic Frontier Foundation has a good, short guide about this.) Unfortunately, they are not being handled carefully right now, which is both inflaming tensions across the globe and opening avenues for potential future institutional abuse.

Authoritarian governments are proving adept at mastering digital platforms, censoring them where desired, and using them for surveillance and propaganda. This may create "peace" in some sense, but at catastrophic cost to human rights. Democracies hoping to build peace must develop strategies that respect both individual expression, and rights such as privacy. The rapid scaling of digital media platforms has become an uncontrolled social experiment, but there is still time to inject a healthy dose of transparency and accountability.

The Author

Lydia Laurenson has spent her career working on media, technology, community-building, and social good. She has done in-depth research and digital product development at both social media and digital journalism companies. She has also written extensively about tech, media, and culture: about digital pseudonyms for The Atlantic; about digital media business models for Harvard Business Review; about San Francisco's very own secret society startup for Vice's tech publication Motherboard; and more. Find her on Twitter @lydialaurenson.

Toda Peace Institute

The **Toda Peace Institute** is an independent, nonpartisan institute committed to advancing a more just and peaceful world through policy-oriented peace research and practice. The Institute commissions evidence-based research, convenes multi-track and multi-disciplinary problem-solving workshops and seminars, and promotes dialogue across ethnic, cultural, religious and political divides. It catalyses practical, policy-oriented conversations between theoretical experts, practitioners, policymakers and civil society leaders in order to discern innovative and creative solutions to the major problems confronting the world in the twenty-first century (see www.toda.org for more information).

Contact Us
Toda Peace Institute
Samon Eleven Bldg. 5th Floor
3-1 Samon-cho, Shinjuku-ku, Tokyo 160-0017, Japan
Email: contact@toda.org