Countering online hate speech with methods that promote tolerance

European Council on Tolerance and Reconciliation Round Table Monte Carlo

5-7 March 2018

Prof. Susan Benesch

Faculty Associate, Berkman Klein Center for Internet & Society

Executive Director, Dangerous Speech Project

The internet can seem like a place polluted by hatred, where one is vulnerable to sudden harassment or attack,[1] and where other people are exposed to content that can inspire them to hate or kill, or simply teach them how to kill more efficiently.[2]

Online hateful and harmful messages are indeed so widespread that the problem cannot be laid at the feet of any particular culture or country, nor can such content be easily classified with terms like 'hate speech' or 'extremism' – it is too varied. Similarly, the people who produce harmful content (and their motivations) are too diverse to fit a stereotype.

Daunting though this problem is, there are opportunities to diminish it and to build norms of tolerance that have been largely overlooked so far. This paper thus offers a set of specific and contrarian ideas for better understanding hate speech and other harmful speech that proliferates online, and for reducing the damage such content causes, while limiting the risk of other harms.

1. Prevent hateful speech online

Most efforts to diminish hateful content online have so far been of one kind: trying to remove it. Deleting content or in internet parlance, "takedown," is essential for some types of egregious and clearly illegal content such as child sexual abuse material, but in general it is only

[1] See, e.g. European Commission, Directorate-General for Communication, *Special Eurobarometer 452, Media Pluralism and Democracy*, November 2016, 17 (reporting that "A large majority of those who follow or participate in debates has heard, read, seen or themselves experienced cases where abuse, hate speech or threats are directed at journalists/bloggers/people active on social media (75%); National Society for the Prevention of Cruelty to Children, "Online abuse: facts and statistics," accessed February 15, 2018 https://www.nspcc.org.uk/preventing-abuse/child-abuse-and-neglect/online-abuse/facts-statistics/; Maeve Duggan, *Online Harassment 2017*, Pew Research Center, July 2017, http://www.pewinternet.org/2017/07/11/online-harassment-2017/ (reporting a survey in which 62% of U.S. respondents regarded online harassment as a major problem and 40% had experienced it themselves).

[2] Jacob Asland Ravndal, "The Online Life of a Modern Terrorist: Anders Behring Breivik's Use of the Internet," *VOX Pol*, Oct. 24, 2014, http://www.voxpol.eu/the-online-life-of-a-modern-terrorist-anders-behring-breiviks-use-of-the-internet/.

a stopgap, and a losing game, since new content is posted at a staggering rate. On YouTube alone, more than 300 hours of video are uploaded each minute.[3] The problem of hateful and harmful speech online should be seen not simply as a law enforcement matter, but as a new challenge to public safety that requires social change, and new methods for building norms of tolerance and civility.

As a means of protecting public welfare, trying to clean up the Internet by removing content after it is posted is reactive, not preventive. It is roughly like pursuing food safety by removing harmful food from the market, without preventing new cases of adulteration or poisoning[4] – or recalling dangerous cars after their engines have exploded, without finding effective incentives or methods to make new cars safer. [5]

Deleting content *before* it is posted is a tempting alternative, but hate speech cannot be automatically detected without a large margin of error.[6] If hateful or harmful content were detected by algorithms and removed automatically, this would amount to an enormous and secret system of censorship. Already, internet companies delete millions of posts to enforce law and, mainly, their own internal rules which are largely unknown to users, and should be publicized, as discussed below.

Deletion is not a sustainable way to regulate the internet, in other words, especially on its own. Durable improvement requires methods for decreasing the rate at which harmful content is posted, in the first place.

Some internet companies and researchers have begun to test and study such methods. These efforts rely on an important, overlooked insight: that those who post hateful content are not all incorrigible extremists or 'trolls'. Some are only occasional offenders, and may be more susceptible to persuasion to stop. There have also been promising experiments in making the rules or norms of a community more visible, which tends to improve behavior since most people

---

[3] "YouTube Company Statistics," Statistic Brain Research Institute, accessed March 1, 2018. https://www.statisticbrain.com/youtube-statistics/.

[4] J. Nathan Matias, "A toxic web: what the Victorians can teach us about online abuse," *Guardian,* April 18, 2016, https://www.theguardian.com/technology/2016/apr/18/a-toxic-web-what-the-victorians-can-teach-us-about-online-abuse.

[5] William H. Shaw & Vincent Barry, "Case: The Ford Pinto," in *Moral Issues in Business,* 8th ed. (Belmont, California: Wadsworth, 2001), 83-86.

[6] Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces," *First Workshop on Text Analytics for Cybersecurity and Online Safety* (2016), http://www.ta-cos.org/sites/ta-cos.org/files/tacos2016_SaleemDillionBeneschRuths.pdf; see also Davey Alba, "Defining 'Hate Speech' Online is an Imperfect Act," *Wired*, August 22, 2017, https://www.wired.com/story/defining-hate-speech-online-is-imperfect-art-as-much-as-science/.

obey social norms of which they are aware. This has been demonstrated in a variety of ways, from an experiment to post the rules of a forum on the discussion website Reddit,[7] to Parlio, a site created to foster civil discussion even between people who disagree intensely.[8] Still, preventative work has thus far received minimal attention, and deserves more.

Widespread alarm about online hate speech should be channeled into new opportunities to define and reinforce norms of discourse, and to learn, by means of rigorous research, how to favorably influence behavior online. This is not unrealistic: public concern has helped to drive major behavior change to decrease harm of many other kinds, such as the wearing of seat belts, vaccination, or the decline in smoking. Even though some people continue to transgress such norms, majorities of people have become compliant, and therefore safer. Online, it would not be necessary to eliminate hateful content entirely, in order to increase tolerance and civility. It would be enough for such norms to be embraced by a critical mass of people.

2. Identify and understand harmful content

Hate speech online and related forms of harmful content should be understood not as one problem but as many, best tackled with diverse and customized tools. To find the right tools, forms of harmful speech should be identified as clearly as possible, and distinguished from one another. At present, some categories are unclear, and some harmful content does not fit into any category.

Although some hate speech is explicit and all too easy to identify as such, as a category it is large and contested, with blurry boundaries. The term 'hate speech' is in widespread use, including for decisions about what content should be permitted online, yet we are lacking consensus on how to define it in law,[9] scholarly literature, common parlance, and even in the rules under which Internet companies prohibit some content – and permit the rest.[10] This has led

[7] J. Nathan Matias, *Posting Rules in Online Discussions Prevents Problems and Increases Participation*, Civil Servant, Oct. 8, 2016.
[8] *Parlio is a global online community discussing what matters*. Matter, available at <https://matter.vc/portfolio/parlio/>
[9] For details on the variety of definitions for hate speech, s*ee* Susan Benesch, *Defining and Diminishing Hate Speech*, in State of the World's Minorities and Indigenous Peoples 2014, Minority Rights Group, p. 20. *See also Interview with Kenan Malik*, in P. Herz and P. Molnar (eds.), The Content and Context of Hate Speech, Cambridge University Press, 2012, p. 81.
[10] *See, e.g.* Facebook Community Standards, Hate Speech https://www.facebook.com/communitystandards#hate-speech; The Twitter Rules, https://help.twitter.com/en/rules-and-policies/twitter-rules; YouTube Community Guidelines, Hate Speech https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176

to many cases of mistaken deletion, and failure to delete hateful content, online.[11]

One thing that's clear, paradoxically, is that "I hate you," no matter how vehemently or sincerely expressed, is not hate speech, [12] since a common thread among definitions is that hate speech denigrates or attacks a person or people *due to some characteristic or identity that they share* with other people, such as race, religion, nationality, sexual orientation, gender, or disability. Most definitions list some but not all of these characteristics, which has generated disagreement over which kinds of groups 'count' as targets of hate speech.

To add uncertainty, the term 'hate speech' defines a category not by its substance (what it contains), but entirely by a state of mind, and it isn't clear whose mind matters for this purpose. For example, if asked whether a drawing of the Prophet Mohammed constitutes hate speech, should one consider the intention of the person who made the drawing, or of someone else who disseminated it (often to a new audience), or its effect on some or all of the audience, i.e. people who see it or hear about it?[13] Even once this question is resolved, the state of another person's mind is not always easy to discover, especially when its expression is found online.

Definitional problems arise with other categories as well. What may be called extremist content, for example, can be thus identified because it was produced by extremists, or because it depicts and endorses gore, or because – to the contrary – it is designed to recruit for extremist groups, by falsely describing a safe and satisfying life within them or simply by criticizing life outside those groups in ways that are compelling and convincing to certain audiences, such as lonely, frustrated youth. Though it might not be wrong to label all such content 'extremist,' it would be a mistake to use the same method to try to protect people from all of it.

To find the most effective ways to diminish harm, it is better to categorize speech online by the danger or harm that it presents, not only by its substance or by a state of mind that it expresses or stimulates. I proposed the term 'harmful speech'[14] to embody this idea. From there, it is possible to distinguish harms, and develop responses accordingly. For example, videos of

---

[11] Davey Alba, *Defining 'Hate Speech' Online is an Imperfect Act*, Wired, August 22, 2017. https://www.wired.com/story/defining-hate-speech-online-is-imperfect-art-as-much-as-science/

[12] European Commission Fact sheet: Countering illegal hate speech online, Jan 19, 2018.

[13] For key relevant ideas e.g. on the distinction between giving offense and taking offense, *see* Cherian George, *Hate Spin: The Manufacture of Religious Offense and Its Threat to Democracy*, MIT Press, 2016. For description of the overlooked role of 'malevolent bridge figures,' or people who transmit content from one normative community in which it is offensive or controversial, to another in which it is highly inflammatory, see Susan Benesch, *Charlie the Freethinker: Religion, Blasphemy, and Decent Controversy*, Religion and Human Rights, Dec. 15, 2015.

[14] *See, e.g. Perspectives on Harmful Speech Online*, Berkman Klein Center for Internet & Society, Aug. 14, 2017.

ISIS beheading captives are horrifying, but they may not be very useful as recruiting tools.

Another reason to classify content by the harm that it may cause is that not all harms should be eliminated, even if it were possible. A significant degree of offense, for example, should be tolerated to protect freedom of expression, especially political speech.

I have also coined the term Dangerous Speech,[15] to describe the narrower class of content that increases the risk that people will commit or condone violence against other groups of people. This idea is useful since there is strong consensus that mass violence is a severe harm to be prevented, of course, and since research indicates that the category exists, i.e. certain particular forms of expression seem to lower normal social and psychological barriers against violence between groups. There are striking similarities, or patterns, in the rhetoric that malevolent leaders use to turn their followers violently against others.[16] Lowering barriers to violence is most often accomplished by persuading people to fear another group, and especially to perceive it or its members as a serious threat. For driving groups of people apart to the point of violence, fear is very likely a more powerful and relevant force than hatred.

Other forms of harm that online content can produce, then, include driving people to hate, fear, discriminate, or even condone or commit violence against one another. In Myanmar, for example, Facebook is rife with posts that describe the country's small minority of Rohingya Muslims – or Muslims in general – as a mortal threat to the majority population of Buddhists, and to the continued existence of Myanmar as a Buddhist country.[17] Violent hatred of Jews, Muslims, people of color, and immigrants is also easy to find on websites in the United States, in many languages, since neo-Nazi and white supremacist groups from many countries post their vitriol online there, where it is not a crime. Most hate speech is not only not illegal in the United States, it is protected speech under the First Amendment of the U.S. Constitution.[18] U.S.-based internet companies can choose to remove such content, of course, but are sometimes slow and/or

[15] *See* The Dangerous Speech Project, www.dangerousspeech.org/faq
[16] Susan Benesch, Dangerous Speech: A Proposal to Prevent Group Violence, February 23, 2013. *See also* Jonathan Leader Maynard and Susan Benesch, *Dangerous Speech and Dangerous Ideology, An Integrated Model for Monitoring and Prevention*. Genocide Studies and Prevention, Feb. 26, 2018.
[17] Kevin Roose, *Forget Washington. Facebook's Problems Abroad Are Far More Disturbing,* New York Times, October 29, 2017. https://www.nytimes.com/2017/10/29/business/facebook-misinformation-abroad.html
[18] A small subset of hate speech is illegal in the United States, when it is also incitement to violence and when there is a likelihood of 'imminent lawless action' in response to it. *See* Brandenburg v Ohio, 395 U.S. 444 (1969), the U.S. Supreme Court case that established this legal standard.

reluctant to do so.[19]

Two other distinctions among forms of harmful content are useful: first, between forms that cause direct and indirect harm, and second, between content aimed at individuals, and content aimed at groups. They are briefly described here.

Though inciting one group to violence against another may lead to the greatest harm, this process is *indirect*: speaker A persuades group B to wish (or carry out) ill on group C. Online content can also harm *directly*, when speaker A addresses group C in denigrating, hostile, or frightening terms. In addition to harming members of group C personally, this can also damage the society in which they live, by limiting their participation in its civic and political life, as Jeremy Waldron has pointed out.[20]

Direct harm often comes from online attacks on specific individuals, which are legion. Many of these do not qualify as hate speech, however, since they make no reference to any group whose characteristics the target shares.

For example, individuals are routinely attacked online for an action they have taken or failed to take offline, such as shooting a lion,[21] failing to clean up after a dog,[22] failing to prevent a child from falling into a zoo enclosure,[23] offering to teach humane slaughter,[24] or attending an extremist demonstration.[25] Frequently in such cases, a group of outraged people incorrectly identifies the source of their indignation, attacking someone who had nothing to do with the act in question, and merely shares the name or the physical likeness of the person who did. Either way, the target is often subjected to abuse and threats from hundreds or even thousands of people, in what is known as a 'dogpile.' Just as often, individuals are attacked not for offline behavior but for what they have done online: comments they have posted, or simply "liking" a comment made by someone else.[26]

---

[19] *See, e.g. YouTube Removes U.S. neo-Nazi group Atomwaffen Division's channel*, Al Jazeera. March 1, 2018, https://www.aljazeera.com/news/2018/03/youtube-removes-neo-nazi-group-atomwaffen-division-channel-180301115414748.html.
[20] Jeremy Waldron, *The Harm in Hate Speech*, Harvard University Press, 2014.
[21] BBC Trending, *How the internet descended on the man who killed Cecil the lion*, July 29, 2015.
[22] Jonathan Krim, *Subway fracas escalates into test of the internet's power to shame*, Washington Post, July 7, 2005.
[23] Rozina Sini, *Justice for Harambe: Mother harassed online after gorilla shot dead*, BBC News, May 31, 2016.
[24] Natalie Bogwalker, *Wild Abundance Attacked by Vegan Extremists*, Nov. 16. 2016; Nancy Matsumoto, *Sustainable Meat Supporters and Vegan Activists Both Claim Bullying,* Civil Eats, Jan. 10, 2017.
[25] Laura Sydell, *Some are Troubled by Online Shaming of Charlottesville Rally Participants*, All Tech Considered, NPR, Aug. 15, 2017.
[26] *Shreya Singhal v Union of India*, 24 March, 2015 (an Indian Supreme Court case arising from two girls' sharing and 'liking' a Facebook post, to express their dismay that much of their city was closed down for the funeral of Bal

Finally, people have been attacked without reference to any real act, and without mistaken identity. A heartbreaking example of this are tweets sent from two Twitter accounts to the daughter of the actor Robin Williams just after he committed suicide, with images made to look like his corpse and text blaming her for his death. Zelda Williams was devastated, and chose to leave Twitter.[27] Thousands of Twitter users expressed their outrage over the attack against her, and she eventually returned.

A frequent source of online content that demeans and insults individuals – but that usually does not constitute hate speech – is U.S. President Donald Trump. Twitter, his social media platform of choice, has never deleted one of his tweets, since they are neither illegal in the United States, nor do they violate the more restrictive but still permissive rules of Twitter.[28]

Some Internet companies have revised their rules and policies to focus on content attacking individuals, or what Twitter calls "targeted" attacks, to distinguish them from attacks on un-named groups of people. Deleting such content often comes too late, however, to prevent or undo the damage. Prior censorship is not a solution, since personal attacks take such a variety of forms, and are often so dependent on context for their meaning, that they are impossible to detect reliably. Automated detection can be safely used for narrow classes of content that is obviously illegal and/or very harmful and not difficult to detect, such as images of child pornography that have previously circulated. In most other cases, automatic deletion is so inaccurate that it produces many false positives (it detects content that is not in fact harmful) and false negatives (it misses much content that is harmful). The only way to avoid harm is to prevent the attacks from occurring in the first place.

2. Durable solutions: behavior change and discourse norms

Efforts to reduce harmful content online, especially on large social media platforms, have so far focused on the content itself, not on the people who produce it. This approach fails to tackle the source of the problem, or even, usually, to prevent the harm immediately caused.

Thackeray, a thuggish and violence-inciting leader of poor Hindus, as if to honor him. The two girls were prosecuted; they and their families were also attacked by Thackeray supporters). Accessed Feb 15 2018 at <https://indiankanoon.org/doc/110813550/>

[27] Lauren O'Neil, *Robin Williams' daughter Zelda leaves Twitter due to 'cruel' comments*, CBC News, August 13, 2014.

[28] Jonah Engel Bromwich and Johanna Barr, *Twitter's Shifting Positions on Trump's Tweets*, New York Times, January 3, 2018. https://www.nytimes.com/2018/01/03/technology/twitter-trump-rules.html

Policymakers have tried to diminish the content by means of law,[29] their familiar tool, and have relied mainly on one direct, muscular way to deal with it: attempting to erase it.

States cannot do this themselves without building their own massive systems of censorship (such as China's) so most governments instead pressure internet companies to detect and remove illegal (and sometimes generally hateful) content, as much and as quickly as possible.[30] Large platforms such as Facebook, YouTube, and Twitter have recently increased their rates of removing hate speech, especially in Europe,[31] and this has been met with relief, and also concern that pressure to remove content that *might* be illegal is causing overbroad censorship.[32] This has already happened in many cases[33] under Germany's new Network Enforcement Law, which requires internet companies to delete content that is 'evidently illegal' under German penal law within 24 hours, or face the possibility of fines up to 50 million Euros.

Criminal law and procedure can be valuable, of course, but prosecutions for speech should be undertaken with caution since they often have the unintended consequence of garnering more attention, and more followers, for the defendant.[34] To make more significant and durable progress in diminishing hateful and harmful content online, Internet platforms and online communities must instead learn to motivate their users to produce less of it.

There have been reports of successful online behavior modification, for example by Facebook, to teach users to resolve grievances successfully with one another,[35] by the online gaming company Riot Games, to decrease 'toxic' comments made by players of its game League

---

[29] Cases have been brought against internet companies and also against people who have posted harmful content. Prosecuting all of the latter would be virtually impossible, and it sometimes draws new support for their causes, by publicizing them.

[30] European Commission, *Communication on Tackling Illegal Content Online – Towards an Enhanced Responsibility of Online Platforms*, Sept 28, 2017. *See also* Germany's Netzwerkdurchsetzungsgesetz, also known by the abbreviation NetzDG or as Network Enforcement Law in English, which was passed in July 2017 and took full effect at the beginning of 2018, requiring internet companies to delete content that is 'evidently illegal' under the German penal code, within 24 hours, or face the possibility of fines up to 50 million Euros.

[31] European Commission Fact sheet: Countering illegal hate speech online, Jan 19, 2018 (reporting that Facebook, YouTube, Twitter, Microsoft, and Instagram removed an average of 70% of content reported to them as illegal hate speech).

[32] Emma Llansó and Laura Blanco, EC Recommendation on Tackling Illegal Content Online Doubles Down on Push for Privatized Law Enforcement, Center for Democracy and Technology, March 1, 2018.

[33] See, e.g. Agence France-Presse, *Berlin to evaluate online hate law as minister falls victim*, Jan. 8, 2018; Jefferson Chase, *Facebook slammed for censoring German street artist*, Deutsche Welle, Jan. 15, 2018.

[34] Susan Benesch, *Words as Weapons*, World Policy Journal, March 26, 2012 (describing supporters of South African youth leader Julius Malema lustily singing a hateful song outside the courthouse in which he had just been convicted for singing it); Gabriel Samuels, *Dutch anti-Islam politican Geert Wilders rises in polls after hate crime conviction*, The Independent, Dec. 13, 2016.

[35] Jason Marsh, *Can Science Make Facebook More Compassionate*, Greater Good Magazine, July 25, 2012. https://greatergood.berkeley.edu/article/item/can_science_make_facebook_more_compassionate

of Legends, which is played by millions around the world,[36] and even, as far back as the 1990s, by the Massachusetts Institute of Technology's then-director of academic computing, to reduce online harassment of students.[37] This is a trove of information, but thus far findings have not been published in enough detail to permit replication or statistical analysis. It is essential to build up an accessible, rigorous body of knowledge on how to diminish harmful online behavior.

Since online 'bad actors' are highly varied, in identity, motivation, and objectives, we can expect that they will respond differently to diverse interventions. To improve the chances of success in diminishing harmful content from bad actors, some distinctions should be made among them, so that the effects of interventions on particular categories can be studied.

Already, observations from my own research and others' suggest some categories. Bad actors may be distinguished according to their output or rate of posting bad content, which can be quite easily detected. Some can be called chronic, since they steadily produce large amounts of harmful content. It seems to be widely assumed that such chronic bad actors are responsible for most of the bad or offensive content online: "common wisdom holds that the bulk of the cruelty on the internet comes from a sliver of its inhabitants – the trolls."[38]

However, some evidence suggests that this is not the case, and that in some contexts, the opposite is true. Jeffrey Lin reported that only about 1% of players of the game League of Legends were consistently producing what Lin called toxic content, and that they produced less than 5% of such content on the platform. "The vast majority was from the average person just having a bad day," Lin said. These people can be called intermittent producers of harmful content, to distinguish them from the 'chronic' producers. Using priming techniques including such seemingly minor modifications as changing font colors, Lin said, induced such 'average' bad actors to post much less toxic content, even on their bad days. Lin and his team also experimented with "player tribunals" – in which players who produced toxic content were judged and sanctioned by other players.[39]

Other distinctions among those who post harmful content may be useful for understanding their motivations, and changing their behavior. For example, some seem to be

---

[36] Brendan Maher, *Can a video game company tame toxic behavior*?, Nature, March 30, 2016. https://www.nature.com/news/can-a-video-game-company-tame-toxic-behaviour-1.19647

[37] Gregory A. Jackson, *Promoting Network Civility at MIT: Crime & Punishment, or the Golden Rule?* Oct. 12, 1994. http://www.mit.edu/activities/safe/data/mit-stopit.html

[38] *See* Maher, supra note 32.

[39] *See* Maher, supra note 32.

activated or inspired by content by or about their targets. Others – the ones I call bandwagon-jumpers – seem to be activated by news of offline events, or by posts from others about such events, which leads to surges in hateful or harmful content. Bandwagon-jumpers often cluster around news-based hashtags on Twitter, as if a hashtag were a physical microphone on a stage, surrounded by a growing crowd.

There have also been important experiments and initiatives to reduce hatred and enforce norms of tolerance online, such as Wael Ghonim's Parlio, a platform built for candid, civil discussion about matters of public importance, among people of very different views. As the first platform started with the goal of maintaining civil, tolerant discourse, Parlio[40] required its users to read the rules, one by one, and promise to obey them.

3. Reveal and publicize rules of civility

Posting rules and requiring that users read and accept them is a mild, easy intervention and is very unlikely to do any harm. Yet most internet users are unaware of the rules by which they are governed. Internet companies delete thousands of posts per minute, mainly to enforce their own internal, secret speech regulations.[41] Each company also presents a set of rules to the public, but those are much simpler than the internal rules and, in comparison, quite vague. The public rules are like constitutions, whose meaning becomes clarified when they are interpreted and applied to real cases: in other words, when decisions are taken whether to delete specific texts or images from the internet.

If the internal rules or decisions were revealed, internet users would better understand the borders between permitted and forbidden content, and many would help to enforce them. Research demonstrates that people who see the law (or in this case, rules) are more likely to obey it. This has been demonstrated, notably, among participants in online discussions on the platform Reddit – though Reddit is often notorious for incivility.[42] Internet users can also be successfully taught to encourage others to follow rules, and norms of civility.

In summary, the best responses for countering harmful speech online will be tailored, as much and as thoughtfully as possible: to types of content, to the audiences they reach, and to the

[40] Matter, *supra* note 8.
[41] Some details of internet companies' internal rules have been leaked to journalists and published. *See, e.g*. Julia Angwin and Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, ProPublica, June 28, 2017.
[42] Matias, *supra* note 7.

social, cultural, and historical circumstances in which they circulate. Some people who post harmful content online are incorrigible, just as some people commit heinous crimes like murder, no matter how severely laws against those crimes are enforced. By focusing instead on the large number of people who are *not* incorrigible, researchers can learn how to build online norms of tolerance and civility among them, and substantially decrease the rate at which harmful content appears online.