A Timeline and Catalog of Tech Company Efforts to Reduce Harm

By Lisa Schirch, University of Notre Dame/Toda Peace Institute
PREPRINT EDITION – Not for Citation, Please add comments or suggestions.
February 2023

Unlike any other industry in human history, the evolution of social media products promised users "connection", the foundation of social cohesion. As these new tech products have grown, so too has public awareness of both the positive and negative impacts of these platforms on societies around the world. Based on research with nearly 30 tech staff and dozens of other observers of tech impacts on society,[1] this paper provides a timeline and catalog of tech efforts to reduce harm and its efforts to support social cohesion.

The paper begins with a timeline of the evolution of tech efforts to reduce harm. Silicon Valley's social media products were first seen as hopeful tools for spreading democracy and peace and continue to be used to promote democracy and human rights today. But there are growing concerns that some tech products algorithmically amplify and incentivize harmful content.

Next, the paper provides a catalog of six different strategies and interventions to respond to harmful content. As tech companies began to identify harmful content on their platforms, they have taken a variety of approaches to addressing toxic or harmful content "while the plane is flying." Tech companies began to respond to harmful content online by adding *Community Guidelines* to describe what behavior is not allowed on their tech products. *User Interface* strategies determine how products present content. *Human moderation* strategies determine what content violates community guidelines. *Algorithm-based* strategies determine how tech products rank and recommend content to users and what content is available. *Policies and partnership* strategies refer to the ways companies engage with outside groups and events, such as civil society or elections. *Company infrastructure* strategies refer to how tech companies organize their internal teams to prevent or respond to harm.

Tech company staff offer a range of explanations for taking a user-centered content moderation approach. Some tech insiders interviewed for this report downplayed the responsibility of tech companies for harmful content or online polarization, asserting that technology is just a "mirror" reflecting to people who they are and what they already think. In this view, tech users generate the problem of harmful content and tech companies are building a Trust and Safety infrastructure to advance content moderation. For example, Facebook's Nick Clegg offered this argument noting, "There is no editor dictating the frontpage headline millions will read on Facebook. Instead, there are billions of front pages, each personalized to our individual tastes and preferences, and each reflecting our unique network of friends, Pages, and Groups."[2] Some interviewees noted that journalists overstate the scale of toxic content. Facebook's Clegg

[1] This report was commissioned by the Working Group on Technology and Social Cohesion. Interviews with 26 tech staff were carried out by Althea Middleton-Detzner and Lisa Schirch with support from Search for Common Ground and the KBF Canada Charitable Network in 2022. This report was written by Dr. Lisa Schirch of the University of Notre Dame and Toda Peace Institute. *All content and/or mistakes are the responsibility of the author and not the commissioning organizations or interviewees.*

[2] Nick Clegg, "You and the Algorithm: It Takes Two to Tango." *Medium.* 31 March 2021. https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2

is on record stating that the scale of harmful content online is relatively small, noting, "hate speech is viewed 7 or 8 times for every 10,000 views of content on Facebook."[3] [4]

In response to widespread reports of escalating levels of toxic digital content, Silicon Valley's largest tech companies continue to invest in building a "Trust and Safety" infrastructure[5] to reduce digital harms that contribute to polarization, primarily with approaches to content moderation. For example, after numerous media outlets published critiques of Meta's role in polarization, Vice President for Integrity Guy Rosen posted a rebuttal to charges that the company contributed to polarization and offered a listing of the various strategies Facebook is using to try to reduce polarization.[6]

The paper ends with an analysis of the incentives and disincentives felt by tech companies to respond to harmful content. Tech product teams juggle multiple priorities, including user engagement, growth, and profit on one hand, and safety issues on the other. There are tradeoffs between focusing on one area over another. Tech insiders expressed frustration with outsiders offering a myriad of ideas about how to fix tech without understanding the efforts already underway and the complexity that even small changes can result in unintended impacts. Some attempts to fix tech harms have reinforced the problem or created new ones. Reducing tech harms goes well beyond simply adding a button or tweaking product designs. There is no one "silver bullet" to reduce tech harms.

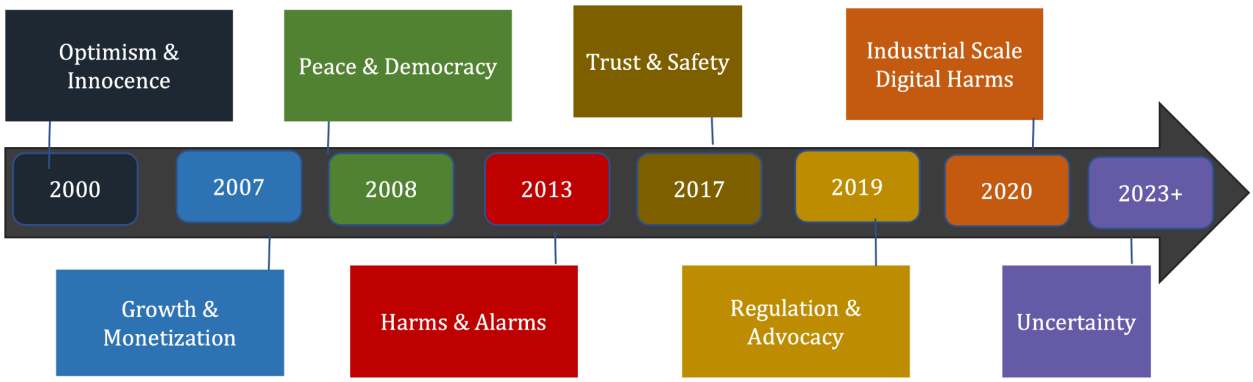
A Timeline of Tech Impacts on Polarization & Social Cohesion

A timeline of tech narratives and responses to harmful content and toxic polarization illustrates the evolution of content moderation over the last 20 years. Tech insiders interviewed for this report noted eight different "eras" of tech company responses to reports on tech roles in polarization and social cohesion.[7] Initial optimism and innocence that social media products could "connect the world" in the early 2000s paired with real world examples of peace and democracy movements such as in the Arab Spring relying on social media products for organizing and mobilizing support. But by the mid-2000s, tech companies' successful focus on growth and ad-based monetization created incentives for engagement-driven algorithms and affordances. Outsiders began pointing their fingers at this advertising model as incentivizing and amplifying harmful content. Like motorists slow down to see a traffic accident, so too do users give more attention to alarming content online.[8]

Soon there was an initial era of alarming reports of toxic polarization from Myanmar, Sri Lanka, the Philippines and elsewhere beginning in 2013. In response to media reports and public concerns about harmful content on some tech products, tech companies began to develop a "trust and safety" infrastructure and governments began floating serious proposals for tech regulation between 2017 and 2019. By 2020, there was growing evidence of the industrialization of digital harm with cyber armies using bots to wage cognitive warfare on domestic and foreign populations. Today, tech staff report an era of uncertainty with an expectation that regulation and market forces will significantly alter the landscape of social media in the next several years.

---

[3] Ibid.
[4] Without full access to internal research, it is difficult to challenge these numbers. Yet there is wide skepticism that the problem is small given the wide perception of the vast scale of false, deceptive, and hateful content on social media. A meta-analysis of research on the scale of mis/disinformation on social media related to the COVID-19 pandemic found that up to one third of Covid-related content was false or deceptive.
[5] See for example the Trust and Safety Professionals Association at https://www.tspa.org
[6] Guy Rosen. "Investments to Fight Polarization." *Meta*. 27 May 2020.
[7] See for example Katie Harbath. "Decentralization and Disruption." *Anchor Change*. 18 February 2022.
[8] This comparison comes from Tristan Harris and Aza Raskin on their podcast "Our Undivided Attention."



2000: Optimism & Innocence

The first years of tech innovation of social media products was full of optimism and innocence. Early tech leaders like Flickr's CEO Caterina Fake referred to her community-building startups like Kickstarter and Etsy as "online communities" rather than "social media." Like other tech innovators, Fake was interested in tech for social good.[9] With a mission to foster social cohesion, Stanford University's social psychologist B.J. Fogg published *Persuasive Technology: Using Computers to Change What We Think & Do*.[10] Dozens of tech executives would trace their user engagement and growth strategy back to Fogg's class and book.[11] Within tech companies, there is wide optimism that internet technology will connect the world. Outside observers still tend to view social media platforms as entertaining but of little importance.

2007: Growth & Monetization

Many Silicon Valley companies began without a business plan in mind. The goal was to create popular products and figure out how to make money with them later. As growth in users increased, companies searched for a way to monetize their platforms to generate profit. Google hired Sheryl Sandberg to create an ad structure to help them monetize their search engine. Sandberg was so successful at developing a business model for Google that Mark Zuckerberg hired her to do the same thing for Facebook. Monetization followed a similar pattern: collect user information to improve ad targeting, incorporate psychological insights and principles from persuasive technology that aim to alter user behaviors, and then tailor information to user profiles with algorithms to predict user preferences. These changes on behalf of growth and monetization later became focal points for analysis of tech harms to social cohesion. Monetization for the "attention economy" also incentivized algorithms that emotionally engaged users with polarizing content, disinformation, and hate speech. Like motorists slow down to see a traffic accident, so too do users give more attention to alarming content online.[12]All user engagement translated into more personal information to sell to advertisers for use in targeting ads, and more users viewing ads on these platforms.

2009: Peace & Democracy?

[9] Reid Hoffman. *The Right Way to Build an Online Community: 3 Rules from Investor and Flickr Cofounder Caterina Fake*. LinkedIn. 2018.
[10] B.J. Fogg, *Persuasive Technology: Using Computers to Change what we Think and Do*. Amsterdam Morgan Kaufmann Publishers, 2003.
[11] Simone Stolzoff. "The Formula for Phone Addiction Might Double as a Cure." *Wired*. 1 February 2018.
[12] This comparison comes from Tristan Harris and Aza Raskin on their podcast "Our Undivided Attention."

The late 2000s was an era of techno-optimism both inside tech companies and in wider public narratives about the impacts social media products might have on the world. In Silicon Valley, commentators noted that Airbnb and Uber, and other sharing economy platforms were creating a "peace dividend" by bringing strangers together. In the "sharing economy," people earn money by sharing their vehicles and homes. Analysts noted these technologies could build social cohesion by getting "strangers to trust each other"[13] and create an opportunity for strangers to discuss their own cultures, global challenges, and shared humanity.[14] In "Making Peace at eBay," Colin Rule laid the groundwork for how technology can work at scale to help people solve their conflicts together using core principles of mediation and conflict resolution based on his online dispute resolution work at eBay.[15] Building off of his work on persuasive technology, Stanford's BJ Fogg called Facebook a "peace technology," predicting that it would create world peace in 30 years.[16] Tech innovators in other parts of the world were innovating web platforms to support citizen journalism in countries like Sri Lanka. In Kenya, new forms of "peacetech" enabled early warning and prevention of election violence. By 2011, Arab Spring activists used Twitter and Facebook to recruit new members, organize protests, and gather reports, photos, and videos of government repression. One Tunisian activist referred to social media technology as "the GPS for this revolution" by helping to guide the leaders of democratic movements.[17]

2013: Harms & Alarms

By the early 2010s, there were growing reports of significant harms on social media. There were increased media reports of individual harms such as cyberbullying online as well as accusations of coordinated campaigns against minority populations. Civil society groups far from Silicon Valley began making the trip to Facebook headquarters to report alarming information on the weaponization of the platform. In 2013, tech observers in Myanmar shared with Facebook executives how the Myanmar government was using Facebook to mobilize violence against minority Muslim groups.[18] By 2016, Filipino journalist Maria Ressa brought Facebook executives evidence of presidential candidate Duterte's false and inflammatory information posted on its platform.[19] In 2017, US intelligence agencies and the US Senate confirmed that Russia had attempted to interfere in both the Brexit referendum and the US election by creating fake accounts and spreading memes aimed to dissuade some voters while motivating others. In 2018, a white supremacist went on Facebook Live to video-stream his murder of Muslims in two mosques in Christchurch, New Zealand. The Christchurch massacre revealed a problem of scale as Facebook users attempted to share the video 1.5 million times within 24 hours. Angry at the refusal to take responsibility for the algorithmic amplification of the video, the New Zealand privacy commissioner called Facebook's leaders "morally bankrupt pathological liars."[20] Facebook

[13] MacDonald, Chris. "Uber is Built on Trust." *IEEE Technology and Society* (10 December 2014). https://technologyandsociety.org/uber-is-built-on-trust/.
[14] Jiang, Li. "The Airbnb Peace Theory." . Accessed Dec 28, 2021. https://medium.com/@lijiang2087/the-airbnb-peace-theory-43f8640f7d38.
[15] Colin Rule. "Making Peace at Ebay: Resolving Disputes in the World's Largest Marketplace." *Quarterly Magazine of the Association for Conflict Resolution* (Fall, 2008): 8-11.
[16] B.J. Fogg, *Facebook: Peace Technology.* Scribd, 2007.
[17] Rebecca J. Rosen, "So, was Facebook Responsible for the Arab Spring After all?" *The Atlantic* (3 September 2011). https://www.theatlantic.com/technology/archive/2011/09/so-was-facebook-responsible-for-the-arab-spring-after-all/244314/.
[18] Victoire Rio. "Myanmar: The Role of Social Media in Fomenting Violence." In *Social Media Impacts on Conflict and Democracy: The Techtonic Shift*, edited by Lisa Schirch. Sydney: Routledge, 2021.
[19] BBC. "Nobel Peace Prize: Maria Ressa Attacks Social Media 'Toxic Sludge'." *BBC News,* 10 December 2021. https://www.bbc.com/news/world-59613540.
[20] Shawn Langlois. "'Morally Bankrupt Pathological Liars' at Facebook Can't be Trusted, Warns New Zealand's Privacy Commissioner." https://www.marketwatch.com/story/morally-bankrupt-pathological-liars-at-facebook-cant-be-trusted-warns-new-zealands-privacy-commissioner-2019-04-08.

biographer Steven Levy reported that world leaders called Facebook an "outrage machine," offering "an earsplitting sound system to hate groups."[21] Across the globe, from Venezuela to Zimbabwe, civil society reported examples of how social media was amplifying existing tensions and polarization.[22]

2016: Trust & Safety Efforts

In reaction to a deluge of media reports, some tech companies began investing in new trust and safety infrastructure, resources, and top-level attention to tech harms. Some tech companies began to build internal teams to conduct research and develop new policies and product features to try to minimize harms. For example, in 2016, Twitter formed a Trust and Safety Council. In the same year, Google's parent company Alphabet created Jigsaw as a think tank to explore using technology to mitigate digital threats.[23] In 2018, Mozilla Foundation created an "internet health" initiative.[24] Microsoft launched a "Digital Peace Now" campaign focused on cybersecurity.[25] Facebook's "Integrity Timeline" asserts that the company increased investment to improve safety on the platform starting in 2016.[26] Facebook's team of engineers and researchers began working on a "Common Ground Initiative" that empowered internal staff to work on social cohesion and conflict in an exploratory way, described in more detail later in this report. In 2021, Facebook media ads asserted that the company has spent $13 billion on "safety and security" since 2016 and has 40,000 employees working on preventing harm.[27] These efforts moved from being primarily reactionary to attempting to prevent abuses and "get out in front" of crises.

2019: Regulation & Advocacy

Critics of tech noted that tech companies were not moving fast enough to change their products to reduce harmful content. They chided that while big tech had not knowingly created products that could undermine democratic elections and spread disinformation and hate speech, that once they had clear information on how malevolent users were using the algorithms and monetization potential to profit from and spread harmful content, tech companies themselves were liable for the harms that occurred. This era brought increasing attention from governments and international organizations prompted by growing alarms related to the spread of harmful content on tech platforms. Civil society advocacy pushing for regulations and changes to platform algorithms began to gather momentum. The European Commission developed a Code of Practice on Disinformation signed by major social media platforms. European regulators developed the E.U.'s privacy and data protection rules in the General Data Protection Regulation (GDPR) to address digital harms. In July 2020, US civil rights organizations called upon companies to boycott advertising on Facebook in the Stop Hate for Profit campaign to protest the platform's handling of hate speech and misinformation and urge reforms.

2020: Industrial Scale of Harm

Tech insiders interviewed for this report noted that by 2020 it was clear that their interventions to reduce harmful content could not keeping pace with the industrial production of harmful content online. Military and intelligence agencies have a long history of propaganda, psychological manipulation, and information operations. Harnessing the new powers of digital technology, a

[21] Steven Levy. *Facebook: The Inside Story*. New York: Blue Rider Press, 2020. Pp. 10-11.
[22] Schirch, Lisa. Editor. *Social Media Impacts on Conflict and Democracy: The Techtonic Shift*. Sydney: Routledge, 2021.
[23] https://jigsaw.google.com
[24] https://foundation.mozilla.org/en/internet-health/
[25] https://blogs.microsoft.com/on-the-issues/2018/09/28/digital-peace-now-launches-this-weekend/
[26] https://transparency.fb.com/policies/improving/timeline/
[27] Meta. "Our Progress Addressing Challenges and Innovating Responsibly." 21 September 2021. https://about.fb.com/news/2021/09/our-progress-addressing-challenges-and-innovating-responsibly/

booming new industry of "disinformation for hire" is operating at a scale that amounts to what NATO refers to as "cognitive warfare." The disinformation industry promises shadowy political actors the ability to alter the opinions and behaviors toward authoritarian candidates or away from political candidates supportive of democracy,[28] though some amount of this might be marketing hype. By 2020, Oxford University's Programme on Democracy and Technology 2020 report found 81 countries using computational propaganda and disinformation campaigns as part of their political strategy.[29] Political actors from ISIS to Russia weaponize these affordances to operate mass influence operations. Cyber troops and a booming for-profit disinformation industry generate content conduct mass cognitive warfare on social media platforms. This includes undermining public trust in democratic institutions and elections, discrediting human rights activists, and widening preexisting divisions in society. Social media affordances enable ordinary people to amplify divisive propaganda by sharing false, deceptive, or polarizing information campaigns, also known as *ampliganda.*[30]

Toxic polarization is increasing globally, contributing to violence, and hampering efforts to solve pressing problems from Covid to the climate crisis. While not the origin of social and political division, there is wide agreement that the industrial production and incentive structures on social media are amplifying and distorting polarization. Journalists and researchers across all regions of the world report social media playing a key role in further polarizing already divided societies, undermining public trust in democratic institutions, and increasing public support for autocrats.

2023: Uncertainty

Tech company responses to harmful content are mixed. Some tech company leaders continue to tout the role of tech in social cohesion. For example, in 2020, Twitter released a "Global Impact Report" that claims it is committed to "promoting healthy conversation."[31] Between 2020 and 2022, Mark Zuckerberg's annual update affirmed that Meta could build a global social infrastructure to help people overcome tribalism and work together.[32] On the other hand, in the last few years, Facebook executives stopped apologizing for content harms and became more combative toward media critics.

Most observers and insiders expect big changes in the years ahead. Interviewees described the tech sector as running on the energy of "The Next New Thing" and that tech innovators "abhor boredom." Advances in blockchain technology, web3 applications, the metaverse, and virtual reality and augmented reality will introduce more complexity to harmful content online. This is an era of uncertainty and turbulence as new regulations and technologies are likely to impact the challenges and opportunities facing technology and social cohesion.

In 2022, inflationary pressure and rising interest rates, reduced tech company stocks, and Elon Musk's Twitter acquisition resulted in layoffs of 100,000 tech workers and downsized or eliminated human rights and content moderation teams. The content moderation challenge is far greater outside the US. In India, Twitter recently fired around 180 of its 230 employees. Mass layoffs and cuts to Trust and Safety teams leave many people wondering what will happen now to the efforts to curb harmful content.

---

[28] Max Fisher. "Disinformation for Hire, a Shadow Industry, Is Quietly Booming." *New York Times.* 25 July 2021. https://www.nytimes.com/2021/07/25/world/europe/disinformation-social-media.html

[29] Samantha Bradshaw, Hannah Bailey, and Philip Howard. "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation." Working Paper 2021.1. Oxford, UK: Project on Computational Propaganda. 2021.

[30] Renée DiResta. "It's Not Misinformation. It's Amplified Propaganda." *The Atlantic*. 9 October 2021. https://www.theatlantic.com/ideas/archive/2021/10/disinformation-propaganda-amplification-ampliganda/620334/

[31] Twitter 2020 Global Impact Report. https://about.twitter.com/content/dam/about-twitter/en/company/global-impact-2020.pdf

[32] Mark Zuckerberg. "Building Global Community." https://www.facebook.com/notes/3707971095882612/.

Political polarization over digital content moderation itself is growing. Tech companies face dilemmas to define the limits of free speech online, and the social norms for digital spaces.[33] On the left, human rights and democracy activists in countries around the world argue that major tech companies do not do enough content moderation. On the right, conservative activists argue that tech companies removing posts deemed hateful, false, or deceptive is a violation of free speech. Content moderation itself, as a strategy for addressing harmful content, is a highly contentious process.

2023 also brings an era of possibility. New tech startups seem to be trending toward inclusive, participatory, and user-controlled spaces that are less centralized. New Venture Capital funds are looking to invest in tech that supports social cohesion. One interview observed that the Wild West period of technology will end, and the peacemakers will build new institutions and policies to civilize these tools over time.

Incentives and Disincentives

The evolution of narratives about tech impacts on society links to the incentives and disincentives tech company staff experience. Staff balance competing motivations including profit incentives related to growth via engagement on one hand; and negative media attention, public outrage, shareholder pressure, and simply wanting to do the right thing to reduce tech impacts on polarization and increase tech contributions to social cohesion.

The chart below contrasts factors increasing tech companies' motivation with those factors that make reducing harmful content, changing tech product designs, or improving social cohesion challenging. Media reports and public pressure to remove harmful content are powerful incentives for tech companies. Yet significant challenges inhibit corporate action, including the complexity of the task and the scale and pace of toxic content.

Factors Incentivizing Tech Attention to Social Cohesion	**Factors Inhibiting Tech Attention to Social Cohesion**
Achieving the tech company mission to "connect" people and growing the user base of people who want a safe place to communicate	Hesitating to change affordances or algorithms that amplify polarizing content because it is profitable
Committing to social responsibility to prevent harm while also reducing charges of political bias	Lacking staff and leadership preparation to manage a global digital town square or to understand how to design products to foster social cohesion
Managing reputational risks from journalist reports and//or public boycotts that might impact the use or investment in the tech product	Managing an escalating amount of harmful content from individual users and industrial producers is creating a sense of futility that content moderation is an endless game of "whack a mole"

[33] Valerie C. Brannon. "Free Speech and the Regulation of Social Media Content." *Congressional Research Service*. 27 March 2019. https://sgp.fas.org/crs/misc/R45650.pdf

Preventing further government regulation that might sanction companies for harmful content	Classifying harmful content to be able to remove it is difficult

Interviews for this report with staff from large social media and search engine companies highlighted the commitment to address the problem of harmful content. Many interviewees insisted that harmful content does not benefit the company's profit model. Social cohesion matters because most technology companies hold a mission to serve the public good by providing information, entertainment, connecting people, etc. Interviewees noted that a tech company that brands itself as strengthening community but then is charged with enabling genocide or undermining democracy has a serious problem. Others stated that a tech company that faces widespread charges of harming society is failing its mission, which will make it more difficult to retain and attract good staff.

Interviewees reported that staff want to feel good about the company that employs them and feel that their efforts are contributing toward a positive corporate mission. Within tech companies, interviewees noted that there is a "huge appetite" for achieving company missions that align with the public good, and great concern about tech-related harms. Some also noted that reports of tech harms have reduced the number of applicants applying to big tech companies, and driving a brain drain away from big tech as some staff left after not seeing enough effort or will to implement needed changes. Other interviewees noted that recent media reports from whistleblowers leaking internal documents have created a sense of distrust which undermines trust and communication within companies, leading to more secrecy and restriction of information and data for researchers.

Interviewees for this report made the case that harmful content is bad for business. As an example of this argument, Facebook Nick Clegg stated in a recent article,

> [It's] not in Facebook's interest — financially or reputationally — to continually turn up the temperature and push users towards ever more extreme content. The company's long-term growth will be best served if people continue to use its products for years to come. If it prioritized keeping you online an extra 10 or 20 minutes, but in doing so made you less likely to return in the future, it would be self-defeating. And bear in mind, the vast majority of Facebook's revenue comes from advertising. Advertisers don't want their brands and products displayed next to extreme or hateful content — a point that many made explicitly last summer during a high-profile boycott by a number of household-name brands.[34]

One interviewee noted that over the long term, some people are going to leave tech products that generate anger, recrimination, conflict, and some will gravitate towards other tech products that create empathy, connection, belonging dignity, and a sense of inclusion. One interviewee in a tech startup noted that "If you build a system to give people justice, transparency, and a place where they feel heard, and they feel fairly treated, they will come back, and they will reward you with more money."

Several interviewees noted they were never in a room where anyone spoke about how a product or algorithm change aimed at reducing harm might reduce profits. Several insiders asserted they never directly observed tension between profits over safety or public goods like social cohesion. Other interviewees noted the ad-based profit models are an unacknowledged obstacle to the bigger changes that might reduce harm and increase benefits. They note the profit model incentivizes keeping users on their product longer to collect more information and show more ads to users.

[34] Nick Clegg. "You and the Algorithm: It Takes Two to Tango." *Medium.* 31 March 2021.

Who is in the room when important decisions are made? Other interviewees noted that while profits might not be discussed during a crisis, the overarching push for growth, user engagement, and profit is the main central framework for employees seeking to climb the ranks. Interviewees also observed that even mainstream news organizations optimize their content for user engagement.

While tech company spokespeople like Clegg challenge the claim that tech company profit models incentivize polarizing content, other observers noted that the boycott Clegg references had little visible impact on Facebook. More than a thousand of the 9 million companies that advertise on Facebook joined the Stop Hate for Profit boycott of Facebook, including large advertisers. The boycott did result in a short-term decrease in company profits.[35] While the boycott harmed Facebook's reputation, boycotts against social media companies have not yet met a threshold to cause shareholder harm to the company. To date, user boycotts and advertiser boycotts have had little impact on profits.

Media reports and public pressure to remove harmful content are powerful incentives for tech companies to act. Yet significant challenges inhibit corporate action, including the complexity of the task and the scale and pace of toxic content. Some interviewees argued that they balance competing motivations including profit incentives related to growth via user engagement on one hand; and negative media attention, public outrage, shareholder pressure, and simply wanting to do the right thing to reduce tech impacts on polarization on the other hand.

Tech companies are investing far more in efforts to reduce digital harm rather than promote prosocial content. But interviewees noted that there are studies indicating frustration and counterintuitive impacts of content moderation. Harvard Kennedy School found that improving the amount of truthful information had a more powerful effect than removing misinformation.[36] Correcting people on Twitter leads to more toxic and less accurate future retweets. Researchers found causal evidence on Twitter that the experience of being corrected increases the partisan slant and language toxicity of a user's subsequent retweets and had no significant effect on the user's primary tweets. Researchers inferred that those individuals felt defensive after being publicly corrected by another user, which shifted their attention away from accuracy concerns. The researchers note this presents an important challenge for social correction approaches.[37]

A main challenge of moderation is to find a way to analyze nuance at scale. Facebook has over 3 billion users, creating an unimaginable amount of content requiring classification systems in dozens of different languages in contexts that change rapidly over time. Metaphors for hate speech may evolve quickly as companies remove one term, and users begin creating new terms or symbols representing the same hateful content. People rapidly innovate new ways of dehumanizing and demonizing others without using explicit hateful terms, or even mentioning the group in question. In Myanmar, for example, people on some social media products were praising the qualities of the Buddhist Burmese. By default, they were excluding the Muslim groups in the country as an insult by erasing them from the narrative.

To date, there has been relatively little effort to look beyond content moderation to design technology that contributes to healthy, pro-social content or social cohesion. Some interviewees noted that it is natural that a company would start from the place where they are getting the most criticism by removing "bad stuff"

---

[35] Tiffany Hsu and Eleanor Lutz. "*More Than 1,000 Companies Boycotted Facebook. Did It Work?" New York Times. 1 August 2020. https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html*

[36] Alberto Acerbi, Sacha Altay, Hugo Mercier. "Research note: Fighting misinformation or fighting for information?" Harvard Misinformation Review.12 January 2022. https://misinforeview.hks.harvard.edu/article/research-note-fighting-misinformation-or-fighting-for-information/

[37] Mosleh, M., Martel, C., Eckles, D., & Rand, D. (2021, May). Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. https://doi.org/10.1145/3411764.3445642

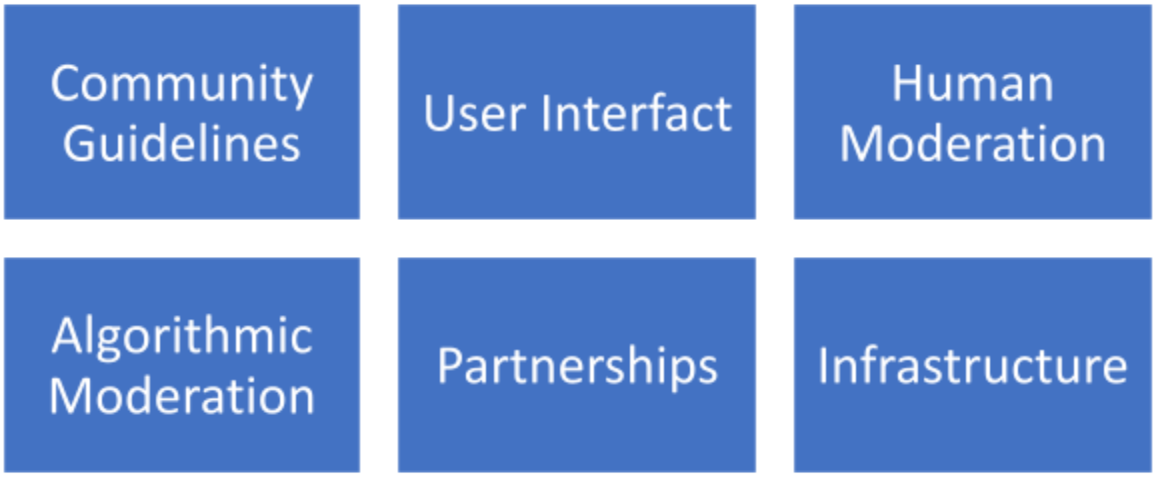
from showing up on their products. A negative experience can be more impactful than a positive one for users.

Tech company efforts to avoid partisan decisions on content moderation are proving unavoidable. Some tech staff assert they are committed to free speech, and thus minimize content moderation. Some use the term "social engineering" to the deliberate psychological manipulation of users through some forms of content moderation. Conservative critics of big tech companies like Facebook and Google note that even tech efforts to reduce harms are a form of social engineering. For example, the Redirect program sends user search queries for white supremacy content to organizations such as Life After Hate, founded and run by former white supremacists to prevent the spread of white supremacy. Some groups view this as a form of censorship rather than viewing it as an effort to reduce harm.[38]

A Catalog of Tech Efforts to Reduce Harm

Flooded with unsolicited advice from all corners of society about what to do with harmful content, tech companies ask for tactical recommendations informed by what has already been tried. This section of the paper provides a catalog of six categories of strategies to reduce harmful content on digital platforms. These include changes to guidelines and norms, user interface, moderation policies, ranking algorithms, internal company infrastructure, and external partnerships.[39] *Guidelines* refer to how people can use the platform. *User Interface* strategies determine how platforms present content. *Human moderation* strategies determine if the content violates community guidelines. *Algorithmic moderation* determines how platforms rank and recommend content to users and what content is available. *Partnership* strategies refer to the ways platforms engage with outside groups and events, such as civil society or elections. *Company infrastructure* strategies refer to how platforms organize their internal teams to prevent or respond to harm.

[38] Bronwyn Howell. "Consequences of the Christchurch Call: Social Engineering by Internet Platforms?" *American Enterprise Institute*. 23 September 2019.
[39] These categories draw from the work of Jigsaw and Jonathan Stray's research.



While in no way comprehensive, the descriptions here provide a catalog of efforts, experiments, and proposed changes to platforms. This typology focuses on efforts within tech companies and does not include a complete catalog of interventions in fact-checking and other strategies used by governments, news media, or civil society organizations to address tech-related harms.

Guidelines and Norms: How do people engage with platforms?

The first intervention tech companies used to try to reduce harms was to create community guidelines. In the early 2000s, the scope of harmful digital content related to photos of bare midriffs and nursing mothers on the early photo-sharing site Flickr. Then CEO Caterina Fake recognized the Orwellian nature of creating and enforcing community guidelines. Tech platforms created community guidelines to help users understand what was permitted or not. While nursing mothers is still an issue on many platforms, the scope of digital harms is now far beyond what early tech CEOs imagined. The evolution of community guidelines continues as users ask for greater transparency in decisions related to content moderation. For example, at Meta what started as a Holocaust denial policy expanded over time to be a genocide denial policy, and then to be a guideline for how to respond to a mass casualty incident denial. Community guidelines set the rules for the "*edges*" of what is acceptable behavior.

Norm Setting

Unlike rules which define the border of acceptable behavior, norms set the pattern for how people behave *most of the time.* Norms are set in a variety of ways. The tone the platform itself uses to communicate with users sets a norm. Group moderators who post content create norms for discussing issues.

Tech companies are exploring ways to set digital communication norms. For example, researchers on Facebook's "Compassion Team" reportedly iterated ways of helping users learn how to ask another user to take down a photo or how to communicate about difficult topics with dialogue rather than outrage.[40] Jigsaw's research brings together anthropologists and psychologists to understand how humans are

[40] Larry Magid. "The Inside Story of Facebook Reactions: Beyond 'Like'" *Huffpost.* 6 December 2017. https://www.huffpost.com/entry/the-inside-story-of-faceb_b_9307108

improvising new digital norms, and what might be able to foster better social cohesion online.[41] Ebay's dispute resolution center videos that users need to watch if they post hurtful content on the platform. Norm-setting videos on social media could model active listening, and what group dialogue looks like when conflict is expressed in healthy rather than toxic ways.

Strategies for norm-setting on social media include popup "nudges" such as TikTok's "Take a Break" videos suggesting that users put their phones down and go outside to protect their well-being.[42] During the pandemic, the U.N. supported a "Pause Before You Share" campaign to encourage people to reduce sharing misinformation.[43] Norms could be shaped by rating content with G-rated content open to everyone and x-rated content requiring a license or age verification.

User Interface: How is content presented on platforms?

Tech platforms serve a variety of purposes. Some but not all have algorithmically curated News Feeds. For Facebook, the News Feed is "king" in terms of the hierarchy of platform design. Everything else is sort of in support of the News Feed. There are a variety of affordances tech platforms use in a way to reduce harm and influence social cohesion.

Buttons

Facebook's "Like" button is a design feature that is a way users can communicate with each other through a click rather than a comment. It can show appreciation or care, but the number of likes on someone else's post may also trigger negative social comparisons. When Facebook explored adding a "Dislike" button…? These guidelines continue to evolve as new types of threats and harms occur as users improvise new ways of abusing user interface, moderation, and algorithm strategies. the platform did not want to encourage people to be divisive in disliking someone's experience that they shared. People can interpret and use symbols in different ways. While Mark Zuckerberg said he hoped to incentivize people to empathize with each other, he noted that it was "surprisingly complicated to make an interaction that you want to be that simple."[44]

Language seems particularly important in political disagreements. Intriguingly, replacing the usual "like" button with a "respect" button increased the number of clicks on counter-ideological comments, that is, people were more likely to "respect" something they disagreed with than to "like" it (Stroud et al., 2017).

User-controlled Blocking, Hiding, and Bozo Filters

Some platforms allow users to block or hide certain content. Bozo filters began as an affordance to early websites that allowed people to send messages or content. A bozo filter keeps unwanted messages or people out. While such an affordance might reduce harms from unwanted sources, it does not proactively build social cohesion.

User-controlled Content Hiding

[41] Gillian Tett. The human factor — why data is not enough to understand the world. *Financial Times.* 28 May 2021. https://www.ft.com/content/4f00469c-75da-4e29-baf3-b7bec470732c

[42] Caroline Burke. "Ever Spent Hours On TikTok Without Realizing It? The App Is Trying To Fix That." *Bustle.* 19 February 2020.

[43] https://news.un.org/en/story/2020/06/1067422

[44] Larry Magid. "Facebook So-Called 'Dislike' Button For Kindness, Not Meanness." *Forbes Magazine*. September 2015. https://www.forbes.com/sites/larrymagid/2015/09/16/facebook-so-called-dislike-button-for-kindness-not-meanness/?sh=291ee68a713e

Facebook added little Xs to the top of every post to allow users to hide content. This gives Facebook a community signal to augment and kickstart more "mature" detection and machine learning models, especially in languages and cultural contexts where it has a more rudimentary ability to work on integrity.

User-controlled Upvoting and Downvoting

Many platforms have asked users to upvote or downvote content as a method of collective ranking. Reddit still primarily operates this way today. Some tech companies have explored the idea of allowing users to upvote or downvote content as a method of individual moderation. Unlike the Pol.is platform described earlier in this report which uses upvoting and downvoting on specific policy proposals, when platforms offer users this affordance in settings where people are sharing their identities or ideas, rather than their policy proposals, the affordance enables users to downvote someone's identity or personal information in ways that cause harm.

User Training of Algorithms

"Likes" are not just social signals to other users, but are data used to train platforms. Facebook offers users the ability to "train" the algorithms so that users see more relevant content. The platform encourages users to "Like" and "Follow" relevant "Pages," "Groups," and "Favorites." Users can also manually select content via the "See First" control. Facebook also offers users the possibility of viewing the News Feed in chronological order rather than on what an algorithm anticipates that users may want to see. Users can also click on the News Feed through the "Why Am I Seeing This?" tool to understand why Facebook's algorithms are showing certain content.

User-controlled Algorithmic Choice

Another human-centered option discussed at Twitter was to allow users to choose their own algorithms. CEO Jack Dorsey noted, "We need to open up and be transparent around how our algorithms work and how they're used, and maybe even enable people to choose their own algorithms to rank the content or to create their own algorithms, to rank it."[45] Several policy researchers have explored the possibility of regulations requiring support for third-party ranking algorithms.

User-controlled Flagging

Some platforms give users an option to flag content they find offensive. The idea behind this affordance was to help moderate and keep the platform safe. However, like other affordances, users have found ways to abuse this power. A common harassment tactic now is coordinated mass flagging until someone's content is removed. On Facebook, only a small fraction of removed content is originally flagged by users, as opposed to platform content moderators or algorithms.

Digital Coaches, Warning Labels, and Accuracy Nudges

Tech companies can put warning labels on harmful content, noting that the post may contain content that is harmful, deceptive, or false. For example, some social media companies added warning labels on posts about the Covid-19 pandemic directing users to sites with verified information. Some interviewees noted that some tech companies coach users about the tone of their posts and provide prompts to help users compose more productive and less harmful digital communication. For example, eBay coaches users who are unhappy with the products they receive to communicate in a way that is more likely to result in a satisfactory outcome. Some interviewees noted tech companies could offer broader coaching or warning labels to users composing a post that includes content that might be harmful to others in order to raise their awareness of potential harm.

[45] Lauren Jackson with Desiree Ibekwe. "Jack Dorsey on Twitter's Mistakes." *New York Times*. 7 August 2020.

Accuracy nudges are digital coaches or warning labels on user content that might contain misinformation. A study of Twitter's "This claim has been disputed" tags on users in the US found that while they reduced the frequency of sharing for Democrats and Independents, they did not have an impact on the sharing of misinformation among Republicans.[46]

Inoculation Posts

Platforms could use inoculation posts to coach users to identify hateful, manipulated, or false content. Inoculation posts can build up people's resistance or "mental antibodies" by first prompting an audience to understand that forces are trying to manipulate people, then explaining your source of information and why it is credible, and then offering a "microdose" of a misleading message.[47] Research found that exposing people to apolitical inoculation messages about the techniques used in disinformation can build "transferable immunity" relevant to a wide range of types of disinformation a person might encounter.[48]

Jigsaw has teamed up with scholars at the Universities of Cambridge and Bristol to develop short videos that inoculate against five of the most common misinformation techniques that apply in a wide variety of contexts online (scapegoating, fearmongering, ad hominem attacks, incoherent logic, false dichotomies. [49] Research at American University's Polarization and Extremism Research Innovation Lab (PERIL) on the efficacy of inoculation against extremist propaganda found that an inoculation message before exposure to extremist propaganda can reduce potential support for extremist messages.[50]

Removal of the Dislike Count

Like the abuse of flagging, people used YouTube's "Dislike" button and count to coordinate targeted dislike campaigns or "review bombings." Some people used the dislike count to humiliate and attack people of color, LGBTQ+, women, or religious minorities. People treated the dislike count as a trolling scoreboard. Some paid attention to the "Like" to "Dislike" ratio. In 2021, YouTube decided to remove the scoreboard or count of dislikes a video received. Research by YouTube found that removing the scoreboard curtailed the harmful game. Content creators can still view the Dislike count, but it is no longer public where it can be embarrassing and stressful.[51]

Friction and Limits to Sharing

Tech companies generally aim for a frictionless user interface, making the products easier to use. Some companies experiment with altering the user interface to limit or add friction to make it more difficult to reshare information since viral content sharing is a signal for potentially harmful content. Slowing down viral sharing also provides tech companies time to analyze the content of viral sharing that is spiking on a platform.

WhatsApp reduced the number of shares that somebody could do on a piece of content to stop some of the spread of misinformation on that platform when you couldn't see the content.[52] At Facebook, executives rejected a "sparing sharing" proposal that would have reduced the content of extremely active users stop

---

[46] J. Lees, A. McCarter, and D.M. Sarno. Twitter's Disputed Tags May Be Ineffective at Reducing Belief in Fake News and Only Reduce Intentions to Share Fake News Among Democrats and Independents. *Journal of Online Trust and Safety*, *1*(3). 2922. https://doi.org/10.54501/jots.v1i3.39

[47] Jigsaw. "Can "Inoculation" Build Broad-Based Resistance to Misinformation?" 17 March 2021. https://medium.com/jigsaw/can-inoculation-build-broad-based-resistance-to-misinformation-6c67e517e314

[48] Stephan Lewandowsky and Sander van der Linden. "Countering Misinformation and Fake News through Inoculation and Prebunking." *Null* 32, no. 2 (2021): 348-384.

[49] Beth Goldberg. "Psychological Inoculation: New Techniques for Fighting Online Extremism." *Medium. Jigsaw.* https://medium.com/jigsaw/psychological-inoculation-new-techniques-for-fighting-online-extremism-b156e439af23

[50] Kurt Braddock. "Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda." *Null* 34, no. 2 (2022): 240-262.

[51] "Update to YouTube's Dislike Count." 11 November 2021. https://www.youtube.com/watch?v=kxOuG8jMIgI

[52] Brian Barrett. "Will WhatsApp's Misinfo Cure Work for Facebook Messenger?" *WIRED.* 4 September 2020.

at two hops. Facebook brought on Netflix recommendation director Carlos Gomez Uribe to lead the newsfeed Integrity Team in January 2017. Uribe wanted to reduce the influence of hyper-partisan users by changing the algorithm that offered more influence to those who Liked, Shared, or Commented on 1500 pieces of content. Uribe argued "super sharers" drowned out people who shared less often. Hyper-partisan users sometimes spend up to 20 hours on the site and act more like bots or shift workers. Uribe's "sparing sharing" proposal would have reduced the content of hyperactive users and combatted spam from Russia or other political actors. Zuckerberg decided to weaken the influence of super sharers by 80%.[53]

User Verification

Some platforms offer user verification tags. Some have explored a reputation system where people's abusive behaviors would accrue on a scoresheet that would follow them to other platforms. Platforms might require users to earn an "internet driver's license" in order to gain access to more advanced tools and powers such as sharing or hosting a group. Those who understand the rules of the road could earn a verification tag. Users might be able to achieve verification tags specifically for engaging in or facilitating healthy conversations. Users might achieve rewards and recognition for the positive roles they play in communities or public posts.

Karma and Reputation Accrual Tags

Platforms such as Reddit and Stack Exchange offer users an affordance that lets them improve their "karma" or community reputation. On Reddit, users accrue karma by adding the total amount of upvotes and subtracting downvotes. Users seek karma because the Reddit algorithms use it to determine the ranking of user content including both their posts and their comments on others' posts or community forums. and which posts or comments it shows to other users. Research finds that a variety of factors contribute to what content receives upvotes, including whether the user has relevance, the title of the post, and other factors.[54] Reputation accrual systems might also work similarly to the internet driver's license concept discussed above, where users gain further access to more tools once gaining a certain level of social credit or karma. Reputation accrual could also help to incentivize content that builds social cohesion. For example, users might be able to offer an upvote on content that attempts to show multiple points of view or that identifies common ground between people discussing an issue. A karma system based on how the content improves the quality of a discussion would be distinct from a karma system based on the ideas of the content.

Groups and Group Moderators

Creating options to form online groups was one strategy tech companies used to attempt to improve social cohesion by providing a user interface that enabled smaller, more private conversations instead of the public posts on Newsfeed. Group moderators can create and enforce shared guidelines. Some groups have "onboarding" to teach the group's social norms and rules to new members who may have different expectations and assumptions. Moderators could remind group members and enforce the rules. Some platforms gave moderators the tools to "boost" content by highlighting or adding positive comments to help set the tone and the norms for the group.

As platforms realized that some groups were being used to spread divisive, false, and hateful content to recruit new members, moderators received more guidance to foster more explicit pro-social norms to steer them toward more constructive communication. Facebook researchers identified a variety of suggestions for decreasing harm in Groups on their platform. For example, one strategy was to allow moderators to

---

[53] Horwitz, Jeff and Deepa Seetharaman. "Facebook Shut Efforts to Become Less Polarizing --- the Giant Studied how it Splits Users, then Largely Shelved the Research." *The Wall Street Journal,* 27 May 2020.

[54] Himabindu Lakkaraju, Julian McAuley, Jure Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 2021. 311-320. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14408

create a temporary subgroup as a space for people arguing on difficult topics so that other users would not witness and enter the conflict. Other options included changing the algorithms to ensure individuals would see a broad range of options for "Group Recommendations" to decrease the likelihood of people groups that spread disinformation or extremism.

Groups were not the engine of social cohesion tech companies had hoped. Group administrators found that content moderation and managing the group's dynamics was time-consuming and a difficult job even for those highly trained in facilitating in-person dialogues.

Human and AI Moderation: What content is available?

Any type of intervention in this area requires some form of human coding of what is or is not acceptable. Tech platforms identify the scope of harmful content in their community guidelines, often detailing prohibitions against spam, sexually explicit content, hate speech, bullying, harassment, and incitement to violence. Decisions regarding content moderation are widely debated: Who should have the power to develop classifying systems: the tech company, governments, civil society, or some combination?

Identifying harmful content (text, photo, video, and live streaming) is difficult for both human moderators and AI-driven algorithms. It requires consideration of identifying the intent behind the content; and whether it is true, false, misleading, partially false, and partially or fully threatening to contribute to offline violence. This can be difficult, especially right after an episode of violence or an election.

Human moderators interpret content in different ways, based on their own experiences. AI cannot reliably detect ambiguous content or make difficult decisions based on protecting principles of defending the freedom of speech. Users must be able to understand why content might have been deleted or demoted. Algorithms and policies that work on one platform in one context might be irrelevant or even harmful on another platform or context.[55] Most AI content moderation only works in a handful of languages that have data sets coding and classifying key terms.

It is not clear how well moderation works to reduce harmful content. Users seem to experience moderation as a punishment and it makes users angry, possibly fueling more online outrage. Users complain of censorship and a lack of transparency on what they said that got them in trouble. Users ask for better platform communication on what was considered harmful or false, and what violated the community standard. Without this, users may feel victimized by platforms.

Interviews for this report noted that any form of content moderation, demotion, promotion, or redirecting users is only a stopgap measure. It does not address the root causes of the individual posting the content.

Algorithmic Interventions: How is content identified and ranked?

Tech companies use algorithms for content filtering and selection in two main ways: to identify content that should not be available (moderation), and to select the content that each user sees (ranking).

Automated systems have become an essential part of content moderation rather than human moderators for two main reasons. First, the Algorithms work at a scale impossible to reach with only human moderators. Second, tech companies originally thought that algorithms would be more "neutral" than human moderators, each of whom might define harmful, deceptive, and divisive content in different ways. However, every algorithmic process will have some set of effects on various stakeholders, good or bad, so the design of such systems is never a value-free choice. In many cases, whether through biased training

[55] Google. "How Google Fights Disinformation." February 2019.

data or the disparate impacts of a facially neutral approach, in *Weapons of Math Destruction* and *Race After Technology* researchers Cathy O'Neil and Ruha Benjamin explain how bias and oppression are built into information retrieval algorithms.[56] Other researchers explore a taxonomy of the various dimensions of fair information access.[57] And others research how identity and beliefs vary in how one perceives toxicity. This research found that individuals who themselves held racist beliefs were more likely to rate African American English as toxic while less likely to rate anti-Black language as toxic.[58]

After unacceptable items are removed during moderation, a potentially vast number of remaining items must be filtered down to a much smaller, human-sized set. This is known as content "ranking" because most such algorithms operate by assigning a numeric "relevance" score to each item, then selecting only the top-ranked items. The same results may be shown for all users, as is typical with search engines, or individual results may be highly personalized, as is typical with recommender systems. Harmful content may be demoted instead of removed outright, but there are potential interventions for social cohesion that go beyond this. One commonly proposed intervention is to diversify the displayed results, perhaps along ideological lines. Most production recommender systems include a diversification mechanism of some sort, though not typically primarily for social cohesion purposes. Going further, it may be possible to algorithmically favor content that reduces polarization – assuming that other users or publishers are creating such content.[59]

Interviewees noted that tech companies are most interested in algorithmic solutions because solving technology's problems at scale with a computational approach is "in the DNA" of tech companies.

Building Classifiers and Taxonomies

Algorithmic interventions begin by collecting data sets, building taxonomies, and developing machine learning content classification systems. These tasks begin with people creating and sorting different types of information. Many interviewees for this report detailed the successes of building taxonomies for human moderators to use as a guide, or machine learning classification systems to use as a training corpus for machine learning. Given that algorithms reflect the people who make such taxonomies and classification systems, there is a growing demand for more tech partnerships with civil society organizations to build these systems (discussed later in this report.) External groups and partners have helped to build these data sets consisting of lists of words and phrases related to digital harms, with additional contextual information.

Platforms have been building nuanced classifiers for everything from child sexual abuse material to hate speech. A classification system for hyperpolarized content could help to identify future polarized content and its severity. Sentiment analysis and natural language processing (NLP) can help companies see the patterns around specific topics.

For example, in 2021, Apple introduced child sexual abuse material (CSAM) detection technology called Neural Hash. The program is touted as a breakthrough in building data sets and classification systems that will be able to identify users' CSAM content without violating their privacy. However, privacy and civil

---

[56] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code.* Cambridge, Medford, MA: Polity, 2019; Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* The Crown Publishing Group, 2016.

[57] Michael D. Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz. Fairness in Information Access Systems." *Foundations and Trends in Information Retrieval*. 16:1-2

[58] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, Noah A. Smith."Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection." 2022. arXiv:2111.07997

[59] Jonathan Stray. "Designing recommender systems to depolarize." *First Monday.* Volume 27, Number 5 - 2 May 2022. Aviv Ovadya, "Bridging-Based Ranking: How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy", Belfer Center, Harvard Kennedy School, 2022.

society advocate caution that governments or other powerful interests could replicate the system to detect other materials or to falsely implicate innocent people.[60]

Tech companies can also build classifiers for predictable events that may lead to the "cascade of events" that lead from events to online harm and back to physical violence. For example, an incident like the murder of George Floyd created a cascade of events including conspiracies, mis/disinformation, extremist recruitment, and calls for more violence than could be predicted. Online incidents of hate and harassment spike around real-world events.

Ranking algorithms select content from an impractically large set of candidates, by assigning a numeric score to each item and then displaying the top-scoring candidates. There are two main kinds: search engines require a query and typically return similar results for all users, while recommender systems can operate without a query and may produce highly personalized selections. Ranking is determined by a variety of factors but is typically significantly influenced by the predicted probability of user interactions such as clicks, likes, shares, comments, or time spent, collectively known as "engagement." This prediction in turn depends on past user interactions, so such systems respond strongly to user feedback. determine the value or rank of content by user testing and feedback through surveys. Using machine learning algorithms, platforms ask users to mark content that they perceive is offensive or harmful. User values determine the weight of content and whether it is promoted or demoted on a News Feed or Search engine. Many platforms also collect feedback through user surveys which may ask whether the content is valuable or harmful, or from paid annotators who follow elaborate instructions (e.g. Google's 170-page search result rater guidelines).[61] Attention to user feedback as the main guide is one way of ensuring that product engineers do not tune algorithms to promote certain ideological agendas – though of course, the users themselves may have such agendas.

Recommender systems algorithms suggest content a user might find of interest, which might be selected from accounts or groups that a user has followed (e.g. the Facebook News Feed), in accordance with user controls or topic settings (e.g. Google News), or from content available across the platform (e.g. YouTube).regardless of whether a user has asked for recommendations.[62] Most recommender systems make their choices in large part to maximize predicted engagement.[63] In some circumstances, engagement is a signal of value and relevance, while in others it may benefit the platform at the expense of the user.[64] Some social media platforms and some search engines use recommender algorithms to keep users on their platforms longer. Some social media platforms use recommender algorithms to keep users on their platform longer. Critics point out that some recommender algorithms have shown users extremist content, contributing to polarization. For example, someone who watched several credible videos on 9/11 may then be presented with a conspiracy theory video. An interdisciplinary group of experts published a report explaining recommender systems and how they go about ranking content. Companies have been exploring potential changes to the algorithms that might downgrade harmful content while amplifying pro-social content.[65]

---

[60] Zack Whittaker. "Apples' CSAM Detection Tech Under Fire – Again." *TechCrunch.* 19 August 2021. https://techcrunch.com/2021/08/18/apples-csam-detection-tech-is-under-fire-again/

[61] Danny Sullivan, "An overview of our rater guidelines for Search. 2021 https://blog.google/products/search/overview-our-rater-guidelines-search/

[62] Jonathan Stray. "Designing recommender systems to depolarize." *First Monday.* Volume 27, Number 5 - 2 May 2022

[63] Jonathan Stray, et al. "Building Human Values into Recommender Systems: An Interdisciplinary Synthesis." ArXiv abs/2207.10192 (2022).

[64]Priyanjana Bengani, Jonathan Stray, Luke Thorburn. "What's Right and What's Wrong with Optimizing for Engagement." 2022. https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851

[65] Jonathan Stray, et al. "Building Human Values into Recommender Systems: An Interdisciplinary Synthesis." *ArXiv* abs/2207.10192 (2022).

Demotion is a strategy that reduces the distribution of harmful content that is clickbait or found to be sensational, misleading, or false. Tech companies demote content deemed as "spam" or "click bait" which includes content that misrepresents or impersonates to deceive or manipulate users. Research on the metrics of demotion proves that the strategy has been successful.[66] Some tech companies demote harmful content instead of removing it as a way of giving a nod to free speech while recognizing potential harms.

Demonetizing removes the monetary reward for people to engage with harmful content, decreasing the incentive for posting this type of content.

Deplatforming removes a user's account so that they no longer have access to a tech platform. When conservatives in the US were kicked off big tech platforms for spreading disinformation about the election, many users moved to platforms such as Gab and Parlor. According to one interviewee who has conducted interviews with some of these users, some realized that there was no one there for them to argue with and no actual substantive debates. They stated that they "miss being in the arena" of the big tech platforms. Some circumvented policy removals and bought whole new routers so that they could get new IP addresses to get back on Facebook and Twitter because they missed "the arena."

Promotion
Some platforms promote what they deem to be high-quality information. Some tech companies are looking for ways to incentivize positive content with algorithms that post credible news sources adjacent to questionable content on issues such as Covid vaccines.

For example, in 2017, Facebook changed its News Feed algorithm to increase the prevalence of posts deemed "Meaningful Social Interactions" (MSI) defined as "meaningful interactions with emotional, informational, or tangible impact that people believe enhance their lives, the lives of their interaction partners, or their personal relationships." CEO Mark Zuckerberg explained, "Our focus in 2018 is making sure Facebook isn't just fun, but also good for people's well-being and for society" and touted people should feel their time on the platform is "time well spent." The goal was to encourage people to strengthen connections with family and friends and reduce time spent on passive consumption of professionally produced content, which their research suggested was harmful to users' mental health.

Time Magazine reported that the algorithm change resulted in users spending 50 million fewer hours on the platform, reducing the company's stocks by 4% in the first quarter of the year (although by the end of the year, stocks had increased).[67] The Wall Street Journal offered another perspective in its article, "Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead." Unfortunately, the algorithms weighting meaningful social interaction looked for emotional engagement and the number of comments on a post. If a user gave an "angry" emoticon to a post, its rank was five times greater than a simple "like." Instead of promoting friendly conversations, the algorithms seemed to heighten traffic in outrage. Political parties in Europe attributed a shift in the most successful promotional strategies. The new algorithm meant that a political figure needed to post something controversial or evoke a strong emotional response to be seen by the public, to compete with similarly boosted emotional messages of all kinds.[68]

---

[66] Google. "How Google Fights Disinformation." February 2019.
https://kstatic.googleusercontent.com/files/388aa7d18189665e5f5579aef18e181c2d4283fb7b0d4691689dfd1bf92f7ac2ea6816e09c02eb98d5501b8e5705ead65af653cdf94071c47361821e362da55b

[67] Kate Reilly. "Facebook Users Spent 50 Million Fewer Hours Per Day on the Site Last Quarter." *Time*. 31 January 2018.
https://time.com/5127913/facebook-daily-usage-drop-earnings/

[68] Adi Robertson. "Political parties told Facebook its News Feed pushed them into 'more extreme positions'." *The Verge*. 15 September 2021.
https://www.theverge.com/2021/9/15/22675472/facebook-wsj-leaks-news-feed-social-media-politics-polarization

Providing Context

Tech company ranking algorithms may promote news stories garnering engagement from a broader user base. Some tech companies address mis/disinformation by first identifying topics or words that are contentious and then posting other information sources alongside the user's post to prompt users to see other points of view on a topic. For example, Google provides a wider context on its search engine by setting its algorithms to post content that is checked by expert panels and/or fact-checking organizations.[69]

Promotion takes the form of "Recommendations." Critics continue to claim that despite tech company efforts, "Recommendation" algorithms on some platforms, specifically Facebook and YouTube, continue to push users to engage with or watch extremist, false, and harmful content. Facebook algorithms choose what "top" comments. Reducing hateful speech in visible comments seems to encourage others' posts to have a more positive tone.

The Redirect Method

In researching how to reduce violent extremist digital recruitment, Jigsaw harnessed Google's targeted ad strategy to develop the Redirect Method. When an individual searches for a term related to a group like ISIS using any keywords or phrases Jigsaw has found to correlate with potential recruits, the Redirect Method takes that user instead to Arabic- and English-language YouTube channels with preexisting videos that might "undo" ISIS's brainwashing. These might include video clips from former extremists or Muslim leaders who have condemned ISIS's corruption of Islam.[70] Organizations devoted to preventing or "off ramping" individuals attracted to violent extremism such as Moonshot CVE are partners with Google to redirect digital traffic away from violent extremist groups. Moonshot CVE is not a content producer. Instead, they recognize that their value is in redirecting people to existing content supportive of human dignity. Facebook has its own Redirect Initiative. Facebook users who search for terms related to white supremacy in the US, https://counterspeech.fb.com/en/initiatives/redirect/, an organization founded by former violent extremists that provide crisis intervention, education, and support groups.[71]

Partnerships: Who is involved in content analysis?

The fifth category of tech strategies to reduce harm involves a variety of partnerships.

Partnerships with News and Journalist Organizations

News organizations are working with tech platforms on a variety of initiatives to help determine effective strategies for moderation, promotion, and demotion. The First Draft Coalition explores how to best combat disinformation online, especially in the run-up to elections. The Trust Project explores how journalism can signal its trustworthiness online with eight indicators of trust that publishers can use to better convey why their content should be seen as credible. Poynter's International Fact-Checking Network (IFCN) coordinates fact-checking organizations from different countries.

Support for Public-Interest Media

The Google News Initiative (GNI) is investing $300 million over 3 years to strengthen quality journalism and evolve news media business models to drive sustainable growth and technological innovation in the digital age.

Support for Elections

[69] Google. 2019.
[70] Andy Greenberg. "Google's Clever Plan to Stop Aspiring ISIS Recruits." *WIRED*. 7 September 2016.
[71] https://counterspeech.fb.com/en/initiatives/redirect/

Some tech companies are using all three strategies in the run-up to elections. For example, Facebook adds pop-ups and flags on users' posts related to an election. Facebook may tune its algorithms to reduce viral sharing of election posts. Facebook also has provided free training to campaign professionals and political parties, so they have skills and tools to protect themselves from attacks and interference. Tech companies may also identify malicious actors during election processes to determine where they originate, disable their accounts, and then share threat information with other companies and law enforcement officials.

Stopping Political Ads
While the product teams are proactive in trying to build features that reduce harm or proactively find and remove harmful content, policy teams are reactive to identify content that violates community guidelines and build the external relationships necessary to enforce them. Tech companies have an outsized impact on elections in part because of political advertising to targeted groups, enabling a political campaign to send ads targeted to user profiles and interests without seeing the ads that other groups are seeing. Google, for example, stopped political ads in some cases or reduced the targeting options for political ads.[72]

Partnerships with International Mediators
UN officials observe that social media increasingly is another theatre of conflict and war, and can disrupt delicate diplomacy and peace processes. Facebook established a "Trusted Partner" agreement with UNSMIL to address hate speech, incitement to violence, and mis- and disinformation. At the request of UNSMIL, Facebook removed social media posts aimed at discrediting or harming activists, youth, and peace promoters. UNSMIL also worked with local stakeholders to produce a digital code of ethics to reduce harmful content on social media.[73]

Partnerships with Civil Society
Tech companies explored community partnerships first in the US, where they reached out to group administrators and tried to build support for them within the company and with other groups. These groups sometimes represented political interests. A Republican Facebook group known as the "Deplorables" included posts that were demonstrably hateful toward other groups. In an effort to be nonpartisan, Facebook went looking for content from Democratic Facebook groups like Clinton's "Pantsuit Nation" that just did not have the same kind of toxic content.

As tech companies began building networks of "Trusted Partners" these included working with international NGOs as well as in-country civil society organizations (CSOs), especially where tech companies do not have a presence on the ground. Tech companies also invest in included funding think tanks such as the Atlantic Council's DFR Lab.

Hate Speech Lexicons
Ethnographic teams can provide a list of words and phrases considered to be hate speech. These data pipelines need to be generated in the local context and sent to tech companies for use. For example, a civil society group in Ethiopia can provide a list of words and phrases that are considered hate speech in the context today. But then the terms and metaphors for hate speech can change in just a few days. Getting ahead of the firefighting model and feeding data in the other direction requires broad civil society partnerships working with tech companies to constantly build out the training data for algorithms to identify the evolving hate speech.

Good Speech Lexicons or "Positive Motifs"

[72] Google. "Political Content." https://support.google.com/adspolicy/answer/6014595?hl=en
[73] David Lanz, Ahmed Eleiba, Enrico Formica, Camino Kavanagh. "Social media in peace mediation: a practical framework." Bern: Swisspeace and UN Department of Political and Peacebuilding Affairs. June 2021.

Tech companies are just beginning to explore the metrics of positive conversations or high-quality communication on their platforms. It is challenging to determine how terms and phrases may be divisive or cohesive in different contexts. Ideally, such lexicons could be the basis of data sets used to amplify positive content.

Crisis and Safety Centers
In emergency situations such as the posting of false information immediately before an election or right after a violent incident, tech companies may use "break glass measures" to moderate a higher level of content when it is not possible to review it in a timely way to prevent further harm.

Regulation
Tech companies are also promoting regulation as a way of reducing tech harms. Facebook ran an ad campaign saying the company supports regulation of the Internet. Skeptics note that they support regulation because they do not want to be held responsible for toxic content. They want there to be rules imposed upon them so that if they abide by the rules, then any problems become the fault of those responsible for setting the rules. Some observers believe government regulations are necessary to address polarization and improve social cohesion.[74] This report did not fully explore these policy recommendations.

"Trust and Safety" Infrastructure: How do tech staff work on content?

Tech companies' trust and safety infrastructure is expanding and evolving.

Informal Coordination and Learning

Tech Trust and Safety protocols and policies spread from one company to another as staff moved between tech jobs, bringing their experiences with them. Most of the learning has been informal, between staff moving jobs or friendships between staff working for different tech companies. As tech companies began to compete for users, the non-disclosure forms they signed made sharing such strategies more difficult.

Several interviewees pointed to the successful coordination between tech companies to prevent child sexual exploitation through the national database of hashes as an example of high levels of coordination in terms of a third-party kind of entity. A second example is tech coordination to prevent the use of platforms by terrorist groups.

Interviewees noted a lack of coordination space related specifically to building social cohesion.

Research, Policy Development, and Capacity Building

Some tech companies set up independent research initiatives to understand tech harms, and capacity-building funds for universities and civil society organizations to coordinate their research and policy work. As noted in the timeline, Mozilla Foundation created an "internet health" initiative. The Omidyar Network, built from the profits from Ebay, invested in a Tech and Society Initiative. The Hewlett Foundation, built from the profits from Hewlett Packard, also invests in a Cyber Policy Initiative. Microsoft initiated the Digital Peace Now campaign on cybersecurity issues.

Google's parent company Alphabet created Jigsaw as a think tank to explore using technology to mitigate digital threats.[75] Jigsaw describes itself as "a unit within Google that explores threats to open societies and builds technology that inspires scalable solutions." Jigsaw works with academics to bridge behavioral

[74] Paul M. Barrett, Hendrix, and Sims. "Fueling the Fire: How Social Media Intensifies Polarization." New York University Stern Center for Business and Human Rights. September 2021
[75] https://foundation.mozilla.org/en/internet-health/

science with tech products and policies in research on disinformation, censorship, toxicity, and violent extremism. Jigsaw's mission aligns closely with a social cohesion agenda, noting on its website that "Toxic language online silences important voices. We're exploring how machine learning can reduce toxicity online and create more space for healthy conversation." Jigsaw sits adjacent to Google but remains somewhat independent.

Company Infrastructure

Different companies are building different types of internal teams to address the challenges related to digital harms and polarization. Tech platforms use different names for such teams, including the Trust and Safety Team, the Integrity Team, Well-being Team, the Protect and Care Team, the Responsible Innovation Team, the Compassion Team, and the Common Ground team. The term "Trust and Safety" is emerging as the most common way of identifying these efforts across the tech industry. This research project was not able to collect enough data to provide a comparative analysis of the size, titles, or functions of these teams.

The architecture and hierarchy of these teams seem to be frequently shifting and reorganizing. TikTok has a variety of structures related to social cohesion alongside its mission to "inspire creativity and bring joy." The TikTok Trust and Safety Team includes an "Integrity and Authenticity Policy Team, a "Responsible Innovation Team," and an "Outreach and Partnerships Team." According to a job advertisement looking for staff, "The Trust & Safety team at TikTok helps ensure that their global community is safe and empowered to create and enjoy content across all of our applications. The Responsible Innovation team was formed in response to society's growing concern about the role of big tech in society. As the technology sector increasingly takes steps to address both the intended and unintended impact of innovation (e.g algorithmic bias), TikTok has created a dedicated team focused on ethical technology and innovation."[76]

Instagram's well-being team evolved in response to a widely-publicized survey by Britain's Royal Society for Public Health (RSPH), a health education charity, which ranked Instagram as the #1 worst social media network for mental health and wellbeing.[77] The Wellbeing team's job is to "make people feel better while using Instagram.[78]

Facebook Case Study

Research for this report was able to gather more information from Facebook's evolution of its Trust and Safety teams both through desk research and interviews. As such, a more thorough analysis of these internal teams at the largest social media platform provides insight into the development of this architecture.

At Facebook, there is a "Central Integrity Team" as well as smaller integrity teams embedded in different units. These teams research to understand digital harms, as well as experiments designed to reduce harms. For example, Arturo Bejar, the director of engineering for the Facebook Protect and Care team, revealed to Radiolab that he and his team try to prevent suicides by making subtle adjustments.[79] Kelly Winters, a product manager on Facebook's designated "Compassion Team," is a group of

[76] https://www.themuse.com/jobs/tiktok/data-scientist-analytics-responsible-innovation
[77] Royal Society for Public Health. "Instagram Ranked Worst for Young People's Mental Health." 19 May 2017. https://www.rsph.org.uk/about-us/news/instagram-ranked-worst-for-young-people-s-mental-health.html
[78] Instagram's New Wellbeing Team. https://wiidoomedia.com/instagrams-new-wellbeing-team/
[79] Laura Entis. Facebook Updates Its Suicide Prevention Tools. 26 February 2015 https://www.entrepreneur.com/article/243393

designers, engineers, researchers, social scientists, and psychologists who put together advice on handling close relationships and family breakups on the platform. [80]

The Central Integrity Team deals with "gnarly issues" around misinformation, hate speech, bullying, harassment, and the policy constructs around them. Interviewees noted that the team often "bumps up against" social cohesion. Facebook's "Product and Process Team" sits within the Trust and Safety Team and is the link between the policy team and the product team.

One interviewee described these efforts like this,

> A team's research agenda is set by product needs. To develop partnerships for research, a Trust and Safety team first has to convince a product team to care about a particular issue. Likewise, any product feature developed by another team must go through layers of approvals from different teams, including the Trust and Safety team. There is a review process whereby the Trust and Safety team vets the product decision before it is rolled out. Facebook is trying to embed a responsible innovation process in the overall product review process. These teams look at products from an inclusion perspective, from a data and AI ethics perspective, and from a human rights perspective.

Based on interviews with social cohesion and peacebuilding experts, perhaps the most significant team related to social cohesion was Facebook's "Common Ground Initiative" which sat within what is now the Integrity Team. Led by Lisa Conn who had previously worked on depolarization with Twitter and MIT, this team conducted research and experiments to reduce markers of outrage and brought in experts on polarization and social cohesion to consult with Facebook staff.[81] Conn, engaged with bridge building and peacebuilding organizations such as Braver Angels[82] and Search for Common Ground[83] to explore how the platform might contribute to intergroup dialogue. Facebook had sent in a production team to film one of their dialogue workshops and create a film to show Facebook employees that they were interested in these issues.

The Common Ground Initiative emphasized Facebook's neutrality by arguing that the company should not attempt to change people's beliefs, prevent conflict, limit opinions, or stop people from forming groups. Data scientists with the Common Ground Initiative found that hobby-based groups that did not include political ideologies could successfully bring people from different backgrounds together. A 2018 document states, "We're focused on products that increase empathy, understanding, and humanization of the 'other side.'" The Common Ground team recommended that the company form partnerships with academics and nonprofits to increase its credibility for changes affecting public conversation. Researchers in the Common Ground Initiative also found that most fake news, spam, and clickbait came from a small group of "hyper-partisan" users. There was a more extensive infrastructure for spreading such polarizing content on the right in the U.S. than on the left. The team also warned that combating polarization might reduce user engagement and described some of its proposals as "antigrowth" and "requiring Facebook to take a moral stance".[84]

Trust and Safety Professional Association (TSPA)

[80] https://www.facebook.com/compassion
[81] "Healing Societal Division Through Community and Technology with Lisa Conn" Interview with Marsha Druker on *Create Community Podcast*. Episode 15. https://www.createcommunitypod.com/episodes/lisa-conn
[82] https://braverangels.org
[83] https://www.sfcg.org
[84] Jeff Horwitz and Deepa Seetharaman, "Facebook Shut Efforts to Become Less Polarizing --- the Giant Studied how it Splits Users, then Largely Shelved the Research." *The Wall Street Journal,* 27 May 2020.

Facing government regulations, in 2018 tech insiders formed the Trust and Safety Professionals Association.[85] The TSPA is a global community of professionals who develop and enforce principles and policies that define acceptable behavior and content online. New platforms like TikTok joined the TSPA in May 2021.[86] The TSPA creates a space for looking at what other companies have done related to responsible innovation.

Embedding International Frameworks

Big tech teams began expanding their Community Guidelines by looking for international frameworks that would bring clarity and legitimacy to some of the difficult content decisions they were making. For example, in 2019 Facebook hired Miranda Sissons to be the Director of Human Rights and create a human rights policy for Facebook based on the international consensus in the UN Guiding Principles for Business and Human Rights and the Universal Declaration of Human Rights. Sissons notes the importance of using internationally recognized frameworks rather than trying to come up with their own ethical standards.[87] Several of the interviewees noted that the lack of a standardized social cohesion framework is a significant obstacle to building in product review, research, and team support for social cohesion.

Staff Initiatives

In some tech companies, internal staff are urging higher-level executives to do more related to digital harms. Interviewees noted that there are different factions at tech companies. Some understand the "techlash," but believe they are fighting the good fight too.

The Integrity Institute

The Integrity Institute is made up of a group of former tech staff including engineers, product managers, researchers, analysts, data scientists, operations specialists, policy experts, and more. These former staff are working toward a "social internet" that research problems and have experience in both failed and successful attempts to improve platforms. These staff are openly critical of how social media platforms can "use bad design practices or fail to build responsibly, systematically rewarding bad behavior in ways that affect individual well-being, social trust, and the stability of governments and institutions."

Trusted Partner Network

Interviewees described how tech companies are creating "trusted partner networks" (TPN) to help them navigate in contexts where there is an authoritarian government, a lack of rule of law, and repression. These TPNS help companies improve platform options and defenses against abuses so that as the user base grows, platforms are equipped to give them a positive user experience. TPNs also help with social concepts like user rights on the platforms, community guidelines, how to achieve redress on the platform, and how to identify and anticipate harmful activity and trends.

TPNs evolved from companies' hiring anthropologists and market researchers to explore "emerging markets" where company staff did not speak the language or know what was being said on the platform. Some interviewees noted that TPNs are a way to divert responsibility on local partners rather than hiring Trust and Safety professionals who are familiar with local languages, dialects, and shifting political symbols and events that shape harmful content before a platform begins to operate and recruit users in new regions.

International Organizations

[85] https://www.tspa.info/
[86] https://newsroom.tiktok.com/en-us/tiktok-partners-with-the-trust-and-safety-professional-association
[87]

International organizations like the UN and the World Bank have been slow to engage with tech companies or create their own tech tools. In general, governments and international organizations have not fully harnessed the power of social media for good. They have lagged behind advertisers who have more resources for getting out ads on their products. Some tech companies are establishing liaison offices with the UN, OECD, and other Bretton Woods organizations.

Meta, for example, holds a "comprehensive dialogue" with the UN. Interviewees noted that the Meta-UN dialogue includes four approaches. First, there is a dialogue on digital policies on AI, content moderation on hate speech and mis/disinformation, and how algorithms impact content in different countries. Second, this office also provides support for UN diplomatic staff to be more efficient users of the Facebook platform and to learn how to use the power of Facebook for high-level messaging and how to protect the privacy and safety of staff. Third, the office helps international organizations get out its message on climate change, the pandemic, and other issues. This can build social cohesion as the public needs to hear from the leaders of international organizations to understand and trust these organizations. Fourth, social media companies need to be in conversation with international organizations about how to respond to internet shutdowns and political leaders or actors who are using social media to incite violence. When a government is using Facebook to provoke violence, as with the Myanmar government's use of Facebook to promote violence against Rohingya Muslims, Facebook needs to work closely with the UN to respond to ensure they are working with the right actors and are in line with a broader international approach. For example, the Facebook office for international organizations works with the UN on Resolution 1325 on Women, Peace, and Security to make sure that gender equity is central to Facebook's engagement with the Global South.

Other interviewees noted that much of this tech company outreach to international organizations might be more for public relations value than actual interest in collaboration.

Conclusion

This report offers a timeline of tech narratives and eras related to responding to how social media platforms affect toxic polarization and social cohesion. It then offered insights into how interviewees reported the incentives and disincentives for addressing harmful content. The third section of the report then provided a catalog of six different strategies tech companies are using to try to reduce harmful impacts.

Interviewees for this report noted several insights related to these tech efforts to reduce harmful content related to polarization.

Proportional Growth and Effort?

First, interviewees asked questions about whether tech responses are proportional to the threats. Interviewees described a significant increase in resources and attention toward reducing tech harms in the last five years. Tech companies tout their investments in hiring thousands of content moderators and specialists in "integrity" and "trust and safety" on their platforms.

What is unclear is whether the scale of these strategies are reducing harm at the scale and speed of tech expansion in new countries and the industrialization of harmful content. Have these strategies reduced harm at the scale and speed of the growth of tech platforms in countries around the world? Relative to the *seriousness* of the alarms sounding around the world over the last decade, are tech companies investing enough resources and making enough effort to respond to the challenge?

Interviewees noted that relative to the extent of digital harm, the size of these efforts seems to be too small and too slow. Relative to the size of the companies and their growth over the last twenty years, the size of the internal teams and partnerships also seems to be too small.

Focus on Reducing Harm or Building Tech to Support Social Cohesion?

Second, interviewees asked whether tech companies focus too much on reducing harms in comparison to designing technology to increase social cohesion. Trust and safety efforts focus far more on content moderation than in designing tech to support social cohesion. Many interviewees described moderation as a *"whack a mole"* strategy that cannot keep up with the scale of digital harms, especially with the growth of the disinformation industry. Moderation addresses the symptom rather than the causes of worsening polarization. If tech companies want to "connect the world" and be a force for building community, democracy, and peace, how might they incentivize this social cohesion through the same six categories identified above?

No Silver Bullet

Tech insiders expressed frustration with outsiders offering a myriad of ideas about how to fix tech without understanding the efforts already underway and the complexity that even small changes can result in unintended impacts. Some attempts to fix tech harms have reinforced the problem or created new ones. Reducing tech harms goes well beyond adding a button or platform design. There is no one "silver bullet" to reduce tech harms.

A Note on Social Engineering

Like governments, technology companies have a tremendous amount of power to steer human behavior. Governments contribute to social engineering by providing public schools, enforcing a criminal justice system, and building roads and bridges. These activities encourage people to behave in "prosocial" ways that encourage humanizing and expressing concern for others. Societies encourage social cohesion when they use benevolent manipulation to incentivize and structure prosocial behavior.

When tech companies focus on removing harmful content, they can fend off some accusations of political bias by focusing on their platforms primarily for entertainment purposes. But this era may be coming to an end, as commentators increasingly view platforms as playing a role in social engineering conversations about elections, abortion access, health policies, and a myriad of other issues. When tech companies take steps to ensure that a minority group can have a voice on social media platforms or to be represented in a Google or Airbnb search, this proactive step can foster social cohesion and prevent harm. But it also can be viewed as social engineering.