The Case for Designing Tech for Social Cohesion: The Limits of Content Moderation and Tech Regulation

By Lisa Schirch, University of Notre Dame
lschirch@nd.edu
February 16, 2023

Apple CEO Steve Jobs described computers as "bicycles for the mind" that amplify human energy. But the metaphor of a bicycle suggests the internet is a road we can travel in any direction. In reality, digital platforms restrict what we can and cannot do.

Recognizing the power of computer infrastructure on human behavior, Stanford Psychology professor BJ Fogg taught a generation of Silicon Valley innovators how to design tech products to harness psychological insights in his course based on his book *Persuasive Technology*.[1] Digital products are persuasive technologies; they engineer how humans communicate. The design of tech products may amplify some human behaviors, thoughts, and relationships and distorts, obscures, or downgrades others. Small changes to algorithms and user interfaces on social media products can influence what people buy, whether they vote, who they vote for, etc.

Harvard law professor Larry Lessig's book *Code and other Laws of Cyberspace* described the internet as a socio-technical institution; code is law. Technology products also reflect the biases and perspectives of those designing affordances and algorithms.[2] Computer engineers embed their values into the affordances and algorithms that govern human interaction online. The architecture of technology products enables what people can and cannot do. Lessig warned that the digital architecture of the web could enable freedom and privacy, or the contrary; it could enable business and government to surveil and control.[3] Lessig's point applies to polarization and social cohesion. Digital infrastructure can amplify polarization through the code. *And* digital infrastructure can persuade people to build social cohesion.

Since the beginning of Silicon Valley's tech industry, there has been a thriving "tech for good" movement. Yet some of the tech products, particularly social media platforms, have become superhighways for disinformation, hate speech, and other forms of harmful content. The design of digital products shapes the direction we can pedal. Digital technology affordances and algorithms can amplify hate and disinformation online that spill over into real-world violence. Digital tools can also help us to build bridges online to improve social cohesion.

Toxic polarization is increasing globally.[4] While not the origin of social and political division, there is wide agreement that harmful content on social media amplifies and distorts polarization. Toxic polarization refers to harmful levels of distrust and dysfunction in divided societies.

Divisive digital content influences the way political actors, traditional media, and the public frame public issues even for people who do not use social media. The challenge of harmful content online is increasing. Political actors, cyber armies, and a growing for-profit disinformation industry amplify and

[1] B.J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. (Amsterdam: Morgan Kaufmann Publishers, 2003).
[2] See, for example, 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.* (New York: The Crown Publishing Group, 2016).
[3] Lawrence Lessig. *Code and Other Laws of Cyberspace*. (New York: Basic Books. 1999).
[4] Thomas Carothers and Andrew O'Donohue. *Democracies Divided: The Global Challenge of Political Polarization.* (Washington DC: The Brookings Institution. 2019).

incentivize individual producers of divisive digital propaganda aimed at polarizing societies with a "divide and conquer" strategy.

Since around 2017, some tech companies have built a Trust and Safety infrastructure with thousands of staff overseeing a global content moderation effort to remove, demote or disincentivize harmful content. But in 2022, the tech sector's relatively new "Trust and Safety" infrastructure laid off 120,000 tech workers and downsized or eliminated human rights and content moderation teams due to reduced tech company stock prices, Elon Musk's Twitter acquisition, and other global factors.[5] The politicization of content moderation is increasing. In the US, conservatives tend to criticize content moderation as censorship while liberals tend to view content moderation as a matter of life or death for threatened minority groups and democratic institutions targeted by online harmful content. In other countries, there is an even more dangerous polarization with repressive governments using content moderation on human rights and democracy activists or even just shutting off internet access altogether.

This paper draws on nearly 60 interviews with staff at tech companies, critics of big tech, civil society groups impacted by tech-amplified social media, and new tech startups designing platforms to reduce polarization and support social cohesion.[6] The interviews took place between 2021 and 2022, primarily in the US and Europe. Interviews revealed three distinct but complementary narratives or approaches to thinking about polarization and social cohesion in digital spaces.

The ***"User-Centered" Narrative*** describes harmful content online as *generated by users*, with social media products and search engines acting as a *mirror* of society. Several interviewees described the defeating feeling of playing "whack a mole" against the growing tide of individual and state-sponsored industrialized harmful digital content. This narrative points to the need for content moderation on user-generated content and digital media literacy to help the public navigate information and communication on the internet.

The ***"Tech Design Regulation"*** **Narrative** describes harmful content *as amplified by tech product designs* including the *affordances and algorithms* that are optimized for user engagement, advertising, and shareholder profit. Many social media companies optimize their product designs for user engagement to maximize their ad-based profits. Machine learning algorithms promote emotionally alarming and divisive content which tends to garner more attention, just as cars slow down driving past a car accident and as news outlets use the "if it bleeds, it leads" principle to prioritize alarming news. From this point of view, some tech products incentivize harmful content that drives toxic polarization. This narrative presses for government regulation to extend beyond privacy to regulating tech profit models, algorithms, affordances, and designs that amplify toxic content.

The ***"Social Cohesion by Design"*** **Narrative** describes tech products that *amplify and scale social cohesion* by designing affordances and algorithms optimized for these purposes. "Peacetech" engineers with training and expertise in social cohesion can design products that contribute to social cohesion.

---

[5] See Laurel Wamsley. "It's the end of the boom times in tech, as layoffs keep mounting." *NPR.* 16 November 2022.

[6] This research was commissioned by a working group formed to launch a Council on Tech and Social Cohesion, including the Center for Humane Technology, Search for Common Ground, the Toda Peace Institute, Braver Angels, More in Common, the University of Notre Dame, and the Alliance for Peacebuilding. David Jay from the Center for Humane Technology and Althea Middleton-Detzner provided a list of and introduction to tech company staff that could be interviewed for this report. Funding to support the research came from two main sources. Search for Common Ground secured funding from KBF Canada to hire Althea Middleton-Detzner and the Toda Peace Institute supported research by Lisa Schirch, based at the University of Notre Dame. The two conducted most of the interviews together. Schirch wrote this report receiving important feedback from colleagues and interviewees.

These digital products can support human agency to participate in civic action, bridge divided communities, and build trust between the public and institutions.

The first half of this article provides explores the complex relationship between toxic polarization and digital spaces and analyzes these three frames or paradigms for understanding the role of digital spaces in toxic polarization. The second half of the paper focuses on examples and case studies of "social cohesion by design." The paper concludes with a call for governments and tech companies to move beyond content moderation to invest in technologies that will improve societies' ability to solve problems and prevent violence. The paper argues that governments can incentivize social cohesion by design. As a complement to content moderation and government regulation, designing tech to support social cohesion should be a primary strategy for addressing the crisis of toxic polarization.

I. Understanding Polarization and Social Cohesion

Polarization occurs when diverse identity groups in a society are divided along an axis into two sides.[7] In general, polarization or differences of belief are not necessarily harmful and can be opportunities for positive social change. Polarization over the ethics of slavery, colonialism, and women's rights, for example, led to civil rights movements, policy proposals to improve equality, and eventually to social change.

In technical terms, there are different types of polarization. *Issue* polarization describes a normal situation where different groups hold different views but can listen to each other and solve problems that arise through democratic processes because of a shared sense of human dignity and trust. Issue polarization can be managed when there is social cohesion. A society with social cohesion addresses conflict or issue polarization as an opportunity for improving society. Conflict is a normal and important aspect of human relations signaling that there are issues needing attention. Groups of people with different experiences and interests often experience conflict.

Toxic polarization, also known as *affective polarization*, occurs when groups distrust and/or dehumanize others with us-vs-them narratives, viewing others as existential threats, and refusing to solve problems without violence.[8] *Political polarization* refers to a society where political party affiliation becomes a defining element of identity; overshadowing how an individual may feel about an issue.[9] In the US, polarization is not just spilling over from elite political polarization; a growing number of people at the community level hold contempt for people of other political parties.[10] A growing body of evidence suggests that political polarization exaggerates the actual policy differences between groups.[11] In other words, there is a perception gap. People think they disagree more than they actually do.[12]

Affective and political polarization can be toxic to society. Toxic polarization can reduce a society's ability to interact with each other and respond to complex problems like the climate crisis or the pandemic. As public mistrust of other social groups and public institutions decreases, so does an

[7] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. "The Origins and Consequences of Affect Polarization in the United States." *Annual Review of Political Science* 22, (2019), 129-146.
[8] Ibid.
[9] Peter T. Coleman. *The Way Out: How to Overcome Toxic Polarization*. (New York: Columbia University Press, 2021).
[10] Daniel DellaPosta. "Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization." *American Sociological Review*, 85(3), (2020), 507–536.
[11] See Chris Bail. *Break the Social Media Prism: How to Make Our Platforms Less Polarizing.* (Princeton, NJ: Princeton University Press, 2021).
[12] Daniel Yudkin, Stephen Hawkins, and Tim Dixon. "The Perception Gap: How False Impressions are Pulling Americans Apart." *PsyArXiv*, (14 September 2019).

individual's belief that change is possible and that civic engagement is an effective route to change.[13] Societies with low levels of social cohesion have a weaker ability to solve problems together and have an increased likelihood of intergroup violence.[14]

Social cohesion is the opposite of *toxic* polarization, as illustrated in Figure 1. Social cohesion enables a society to function in a way that addresses the needs of all members and to be resilient to shocks, stressors, and crises such as a pandemic or natural disaster. Social cohesion is the glue that keeps a society together. The United Nations defines social cohesion as "the extent of trust in government and within society and the willingness to participate collectively toward a shared vision of sustainable peace and common development goals."[15]

Toxic Polarization		**Social Cohesion**
Individuals feel isolation, humiliation, and frustration and behave as though they are not able to participate in decisions that affect them	Individual agency	Individuals feel a sense of safety and dignity, and behave with the skills and capacity to *influence* and *participate* in decisions that affect them
Individuals feel a sense of *exclusion* and behave with contempt and distrust toward other people	Horizontal cohesion within and between groups	Individuals feel a sense of *belonging and inclusion* and behave with *empathy and trust* toward other people
People in society feel a sense of *exclusion, contempt, and distrust* toward leaders and institutions which are seen as *corrupt and captured* by elite interests	Vertical cohesion between institutions and the public	People in society feel a sense of *inclusion, investment, and trust* in leaders and institutions which are seen as *transparent and accountable* to the public.

Figure 1: The Polarization-Cohesion Spectrum

Social cohesion is both a *goal* and an *approach.* The UN uses the term social cohesion to describe *the goal* of its efforts in peacebuilding, dialogue, participatory governance, prevention of violent extremism, and bridge-building interventions. UN Peacebuilding initiatives have grown out of local peacebuilding and bridge-building efforts to coordinate diverse stakeholders and activities in support of social cohesion.

For more than four decades, the field of peacebuilding has been researching, experimenting, and practicing the science and art of facilitating dialogue, negotiation, and mediation to depolarize divided societies and address the root causes of conflict.[16] International organizations like the UN and World Bank invest large sums in peacebuilding, as they recognize its value in preventing violence, which negatively affects people, business interests, and the planet. The field of peacebuilding is an umbrella term that includes the concepts of conflict resolution, conflict management, and conflict transformation. Within the US, there are a wide range of movements and organizations whose work can be categorized as

[13] Ethan Zuckerman. *Mistrust: Why Losing Faith in Institutions Provides the Tools to Transform Them*. New York: W.W. Norton and Co. (2021), 20.
[14] I. Olawole, et al. *Strengthening Social Cohesion for Violence Prevention: 10 Lessons for Policymakers and Practitioners*. Washington, D.C.: Mercy Corps, 2022; A. Lichtenheld, et al. Understanding the Links Between Social Cohesion and Violence: Evidence from Niger. Washington, D.C.: Mercy Corps. 2021.
[15] United Nations Development Program (UNDP), *Strengthening Social Cohesion: Conceptual Framing and Programming Implications. (*New York: UNDP, 2020).
[16] Fletcher D. Cox and Timothy Sisk. *Peacebuilding in Divided Societies: Toward Social Cohesion*. (Cham, Switzerland: Palgrave MacMillan, 2017).

peacebuilding, including for example groups that aim to protect democracy, address social justice, or build bridges between groups.

The OECD uses the term social cohesion to *describe* a society that "works towards the well-being of all its members, fights exclusion and marginalization, creates a sense of belonging, promotes trust, and offers its members the opportunity of upward social mobility."[17] The OECD defined social cohesion as characteristic of a society that values "the well-being of all its members, fights exclusion and marginalization, creates a sense of belonging, promotes trust, and offers its members the opportunity of upward social mobility."[18]

Based on the work of Search for Common Ground, there are three elements related to social cohesion.[19]

1. **Individual Agency** exists when individuals feel a sense of safety, dignity, and capacity (skill) to *influence* and *participate* in decisions that affect their lives within society and with governing institutions. Individual agency requires an ability to communicate about difficult issues in a "healthy" way with communication skills that focus on problem-solving while recognizing the dignity of oneself and others.

2. **Horizontal Cohesion** exists when individuals feel a sense of *positive relationships, belonging, and trust* within and between identity groups based on politics, religion, ethnicity, class, education, region, or other shared identities. The term "polarization" refers to an absence of horizontal social cohesion. Horizontal cohesion requires skills for healthy expression of conflict and solving problems through inclusive, collaborative, non-violent processes in both bonding and bridging networks. It also includes efforts to improve horizontal cohesion through dialogue and research, building trust through working together in areas where there is common ground, and reality checking, as often people misperceive the intentions and beliefs of others. Horizontal cohesion is also called "horizontal social capital."

   ***Intra*communal cohesion**, also known as "bonding social capital," refers to the quality of relationships within an identity group (e.g., relationships among black Americans).

   ***Inter*communal cohesion**, also known as "bridging social capital" refers to the quality of relationships between identity groups (e.g., between black and white Americans).

3. **Vertical Cohesion** exists when individuals and groups in society feel a sense of *trust, transparency, accountability, and collaboration* with public institutions including government, as well as news media, academic institutions, and corporations. This is also called "vertical social capital." In an active democracy, citizens engage with governments. Civic engagement is an expression of vertical cohesion paired with individual agency. Vertical cohesion exists when public institutions recognize basic human rights and serve community members equitably affording public goods such as equal treatment under the law, safety, healthcare, and education to all.

Countries with higher levels of social cohesion had fewer deaths.[20] Social cohesion enables a society to function in a way that addresses the needs of all members and to be resilient to shocks, stressors, and

[17] OECD, *Perspectives on Global Development 2012: Social Cohesion in a Shifting World*, (Paris: OECD Publishing, 2011).
[18] Mike Colledge and Chris Martyn. "Social Cohesion in the Pandemic Age." IPSOS, (October 2020).
[19] This schema synthesizes similar frameworks for social cohesion, and draws specifically from this report: Institutional Learning Team. "Building Social Cohesion in the Midst of Conflict: Identifying Challenges, Measuring Progress, and Maximizing Results." Search for Common Ground. (November 2020).
[20] Adam Taylor, "Researchers are asking why some countries were better prepared for covid. One surprising answer: Trust." *Washington Post*. (1 February 2022).

crises such as a pandemic or natural disaster. Social cohesion enables societies to work together to solve problems.[21] In the panoply of catastrophes facing humanity today, social cohesion enables societies to work together to solve problems including climate change, poverty, inequality, racism, and violence.[22]

Cohesive societies are more likely to reduce disparities in income and unemployment, are more likely to address problems collectively, and people are more likely to have a sense of belonging in the places they live. Societies with low levels of social cohesion have a weaker ability to solve problems together and have an increased likelihood of intergroup violence.[23] During the Covid pandemic, countries with low levels of social cohesion are suffering more deaths from Covid.[24] A lack of social cohesion can mean that people did not feel a sense of agency to work for change, did not trust their neighbors to wear a mask or get a vaccine, and/or did not trust their government to give them accurate information about the pandemic and vaccine. Countries with higher levels of social cohesion had fewer deaths.[25] Similarly, countries with high levels of social cohesion can make climate policies more acceptable to citizens.[26]

A society with social cohesion addresses conflict as an opportunity for improving society. Conflict is a normal and important aspect of human relations signaling that there are issues needing attention. Groups of people with different experiences and interests often experience conflict. The goal of social cohesion is not to suppress conflict or to reduce differences between groups. Authoritarian governments tend to view conflict itself, such as citizens voicing a critique of government policy, as dangerous. The goal of social cohesion is to provide democratic processes and spaces for public deliberation and creative problem-solving to address conflicts between groups.

[21] World Policy Forum. "Social Cohesion and the State: What can the G20 do to improve social cohesion and trigger responsibility in business and politics?" Summit 2022. Found at: https://www.global-solutions-initiative.org/global-table/social-cohesion-through-business-and-politics/ Accessed 1 January 2023.
[22] Ibid.
[23] I. Olawole, et al. *Strengthening Social Cohesion for Violence Prevention: 10 Lessons for Policymakers and Practitioners*. Washington, D.C.: Mercy Corps, 2022; A. Lichtenheld, et al. *Understanding the Links Between Social Cohesion and Violence: Evidence from Niger*. (Washington, D.C.: Mercy Corps. 2021).
[24] Loring J. Thomas, et al. "Geographical patterns of social cohesion drive disparities in early COVID infection hazard." *Proceedings of the National Academy of Sciences in the United States of America*. (14 March 2022).
[25] Adam Taylor, "Researchers are asking why some countries were better prepared for covid. One surprising answer: Trust." *Washington Post*. 1 February 2022.
[26] Daniele Malerba. "The Effects of Social Protection and Social Cohesion on the Acceptability of Climate Change Mitigation Policies: What Do We (Not) Know in the Context of Low- and Middle-Income Countries?" *The European journal of development research*, 1-25. (6 May 2022).

Definitions

Toxic polarization occurs when people perceive other people as existential threats, distrust and dehumanize others with us-vs-them narratives and justify the use of violence against others. *Toxic polarization* includes three dimensions:

- Individual isolation and a loss of human agency to participate in civic life
- Divisions between groups into narratives of "Us vs them" with emotional contempt for the "other"
- Lack of trust between the public and institutions in government and public-interest media

Social cohesion refers to the glue that keeps society together; it is the opposite of toxic polarization. Three dimensions of social cohesion include:

- Individual agency to participate in civic life
- Horizontal relationships within and between social groups
- Vertical relationships between public institutions and society

Bridge building and peacebuilding are types of ***prosocial interventions*** that support the goal of social cohesion in three ways. 1) Increasing individual agency to participate in civic life; 2) Bridging relationships between groups; and 3) Building public trust between society and governing institutions.

Technology or tech refers in this report to digital tools, with a particular focus on social media. ***Affordances*** are the features of a tech product that shape behaviors. The Like, Share, and Comment features of most social media products are examples of affordances. ***Algorithms*** are the computational settings of a tech product that determine what content users can see.

PeaceTech refers to technology that both supports the analysis of polarization and bridge building or peacebuilding interventions to support social cohesion.

II. Social Media, Harmful Content & Polarization

Many of the tech insiders interviewed for this report questioned the link between technology and social cohesion. Polarization between groups is as old as human history and was increasing globally before the advent of digital technology. There is a robust literature on the impact of social media products on polarization.[27] Research both supports and questions the link between technology and polarization.[28] Some studies have found that polarization is growing more among groups with less internet usage.[29] But research surveys consistently find that social media platforms impact social cohesion by altering social

[27] See for example: 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; Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, Ralph Hertwig. "Digital Media and Democracy: A Systematic Review of Causal and Correlational Evidence Worldwide." *SocArXiv*, 22 (Nov. 2021); Jay J. Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, Anni Sternisko, "How social media shapes polarization." *Trends in Cognitive Sciences*, Volume 25, Issue 11, (2021): 913-916; Almog Simchona, William J. Bradyc, Jay J. Van Bavel. "Troll and divide: the language of online polarization." *PNAS Nexus*, (2022), Vol. 1, No. 1.

[28] See for example, Jonathan Stray. "Designing recommender systems to depolarize." *First Monday.* Volume 27, Number 5 - 2 May 2022; Gideon Lewis-Kraus. "How Harmful Is Social Media?" *New Yorker.* 3 June 2022; Gideon Lewis-Kraus. "How Harmful Is Social Media?" *New Yorker.* 3 June 2022; Lydia Laurenson, Polarisation and Peacebuilding Strategy on Digital Media Platforms. Part 1: "The Current Research." And Part 2: "Current Strategies and their Discontents." Tokyo: Toda Peace Institute, 2019. https://toda.org/assets/files/resources/policy-briefs/t-pb-44_laurenson-lydia_part-1_polarisation-and-peacebuilding-strategy.pdf;

[29] Levi Boxell, Matthew Gentzkow & Jesse M. Shapiro, "Greater Internet use is not associated with faster growth in political polarization among US demographic groups." *Proceedings of the National Academy of Sciences in the United States of America.* 19 September 2017.

networks and fragmenting public conversations on issues, rapidly spreading false information and the dysfunction of digital governance and norms.[30]

A survey study by New York University's Stern Center for Business and Human Rights asserts that while big tech companies like Meta, Twitter, and Google were not the source or largest factor in rising U.S. political polarization, these products amplified "divisiveness" and its "corrosive consequences."[31] According to the Pew Research Center, 64% of Americans believe social media is negatively affecting the US, and express concern about the misinformation and the hate and harassment they see on social media.[32]

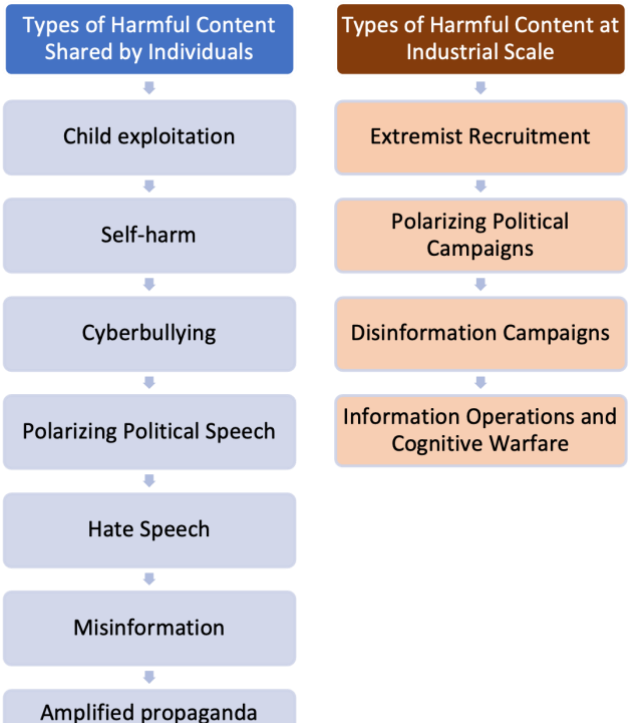

Figure 2: Typology of Harmful Content

Political actors are exploiting social media and search engines to spread false propaganda to divide citizens, aggravate existing social divisions, foment violence, and sway elections. False and deceptive information both online and offline synergize with hateful content, violent extremism, and repressive states to pit "us vs them." Harmful content online contributes toward "toxic polarization."

The problem of harmful content on these tech products started small. Early tech products like eBay and Flickr wrestled with "individual rule breakers" posting spam, fraud, and nudity. But an avalanche of other problems soon followed. The internet became a superhighway for child sexual abuse and exploitation. Some social media products became boxing arenas for verbal jousts and hateful commentary by average people. Individuals spreading harmful content and inadvertent rule breakers soon were joined by industrial-scale producers of harmful content.

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 to undermine public trust in democratic institutions and elections, discredit human rights activists, and widen 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*.[33]

By 2020, the University of Oxford Programme on Democracy and Technology warned of "industrialized disinformation" by over 80 countries with cyber armies spreading computational propaganda.[34]

[30] Sandra González-Bailón and Yphtach Lelkes, "Do social media undermine social cohesion? A critical review." *Social Issues and Policy Review*, 00, (2022), 1– 26.
[31] Barrett, Hendrix, and Sims, (2021).
[32] Brooke Auxier. "64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today. *Pew Research Center*. (15 October 2020).
[33] Renée DiResta. "It's Not Misinformation. It's Amplified Propaganda." *The Atlantic*. (9 October 2021).
[34] Samantha Bradshaw, Hannah Bailey & Philip N. Howard. "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation." Oxford, UK: Programme on Democracy & Technology, (2021). demtech.oii.ox.ac.uk.

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.[35] The impact of industrialized disinformation campaigns is what some call "the liar's dividend" or "epistemic insecurity" where the public senses chaos, feels confused and views everything as questionable resulting in the collapse of truth. It is not uncommon to hear people refer to the weaponization of social media or refer to some tech products as weapons of mass distraction[36] and mass destruction.[37]

Just like a small amount of toxins can pollute a river or lake, even a small amount of harmful content online can create toxic *information ecosystems* that enable autocratic political actors to undermine social cohesion and democracy. The Center for Humane Technology describes "polarization spills" on social media as unique. Unlike toxic oil spills, a polarization spill not only causes harm in dividing society. It also makes it difficult to govern. While an oil spill does not in itself make it more difficult to find regulatory solutions to prevent more oil spills, toxic polarization spills do make it more difficult for political actors to find regulatory solutions to digital amplification of polarization.[38]

III. 3 Approaches to Reducing Polarizing Content on Digital Spaces

There are three distinct but complementary narratives or approaches to thinking about polarization and social cohesion in digital spaces. The first approach blames users for generating harmful content. In this view tech products are neutral mirrors of society. Content moderation focuses on removing user-generated harmful content. The second approach blames tech companies for designing harmful affordances and algorithms that amplify toxic content. This approach advocates government regulation of tech algorithms. The third approach views the challenge as a need to incentivize new digital spaces with affordances and algorithms designed to support social cohesion. Table 1 below synthesizes these three approaches.

Table 1: 3 Narratives on Harmful Digital Content

	Perception of the Challenge	**Interventions**
User-Centered Narrative	Tech insiders often frame the problem as **user-generated content**. In this view, technology is just a **"mirror"** of society.	Tech companies have built a "Trust and Safety" infrastructure to address how people use the internet to cause harm. The bulk of Trust and Safety initiatives focus on moderating content by developing data classifiers and using human moderators paired with machine learning and AI to remove harmful content. Tech insiders often refer to this as a **"whack-a-mole"** effort that cannot keep up with the scale of user-generated harmful content.
Tech Regulation Narrative	Tech critics often frame the problem as **harmful tech products** with profit models	Tech critics identify the need for **government regulation** of tech products not only in terms of data privacy and cybersecurity but also of tech

[35] Lisa Schirch, Editor. Social Media Impacts on Conflict and Democracy: The Techtonic Shift. (Sydney: Routledge, 2021).
[36] Christina Nemr and William Gangware." Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age." (Washington DC: Park Advisors, 2021).
[37] Sarah Jacobs Gamberini. Social Media Weaponization: The Biohazard of Russian Disinformation Campaigns. Center for the Study of Weapons of Mass Destruction. *Joint Forces Quarterly*, (2020).
[38] Center for Humane Technology. "Addressing the TikTok Threat." *Your Undivided Attention Podcast.* 8 September 2022.

	that incentivize affordances and algorithms that distort and amplify the worst aspects of human behavior.	profit models, product affordances, and algorithms.
PeaceTech Narrative	Social cohesion experts frame the problem as **a lack of tech products** that can scale social cohesion efforts to build individual agency, bridge intergroup relationships, and build public trust.	Social cohesion experts see the need for content moderation, tech regulation, and **incentivizing prosocial tech product designs** that amplify the best aspects of human behavior and improve rather than detract from social cohesion.

IV. The *User-Centered* Narrative

Research for this paper found that most tech company staff 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. The logic of externalizing the problem of hateful content is part of a communication strategy for tech companies like Meta. 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."[39] Some interviewees noted that journalists overstate the scale of toxic content. Facebook's Clegg 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."[40] [41]

Tech companies draw on a catalog of tech strategies to improve trust and safety. 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[42] 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 toxic 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.[43]

Flooded with unsolicited advice from all corners of society, tech companies are open to ideas but ask for tactical recommendations informed by what has already been tried. 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.

[39] Nick Clegg, "You and the Algorithm: It Takes Two to Tango." *Medium.* (31 March 2021).
[40] Ibid.
[41] 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.
[42] See for example the Trust and Safety Professionals Association at https://www.tspa.org
[43] Guy Rosen. "Investments to Fight Polarization." *Meta*. (27 May 2020).

As of 2022, tech companies are taking a variety of steps to reduce digital harm. *Guidelines* refer to how people can use the tech product. *User Interface* strategies determine how products present content. *Moderation* strategies determine what content is available. *Algorithm-based* strategies determine how tech products rank and recommend content to users. *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.

A. Incentives for Addressing Toxic Polarization on Tech Products

Tech companies have incentives and disincentives for responding to online polarization. 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.

Incentives include staff desire to achieve their tech company mission to "connect" people and grow the user base of people who want a safe place to communicate. Some identify a broader commitment to social responsibility to prevent harms. Several 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. 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 noted that people 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 drove 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. Tech companies may also respond to digital harms as a way of managing reputational risks from media attention to digital harms or public boycotts that might motivate investors to withdraw support from the company. Tech companies are also trying to prevent further government regulation or sanctions for harmful content.

Yet significant challenges inhibit corporate action. Many companies simply lack the staff necessary to manage a global digital town square and the scale of industrial-scale disinformation and hate speech. 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."

B. Analyzing Nuance at Scale

The task of content moderation is itself difficult, as harmful content first needs to be classified in order to be identified by machine learning algorithms. But classifying disinformation, hate speech, and other forms of harmful content requires ongoing analysis and public debate on the many pieces of content that may be protected as free speech.

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.

C. The Politicization of Content Moderation

Tech companies face dilemmas to define the limits of free speech online, and the social norms for digital spaces.[44] 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.

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.[45]

D. Profit Model Considerations

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. Many interviewees insisted that harmful content does not benefit the company's profit model and 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.[46]

Yet, other interviewees insisted the profit model of user engagement underlies all company decisions about designs and algorithms. Other interviewees noted that while profits might not be discussed during a crisis, the overarching push for growth, user engagement, and profits remain as a central framework for employees seeking to climb the ranks. Other interviewees noted the ad-based profit models are an unacknowledged obstacle to the bigger changes that might reduce harm and increase benefits. One interviewee noted that over the long term, some people are going to leave tech products that generate anger, recrimination, and 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."

---

[44] Valerie C. Brannon. "Free Speech and the Regulation of Social Media Content." *Congressional Research Service*. (27 March 2019).
[45] Bronwyn Howell. "Consequences of the Christchurch Call: Social Engineering by Internet Platforms?" *American Enterprise Institute*. 23 September 2019.
[46] Nick Clegg. "You and the Algorithm: It Takes Two to Tango." *Medium.* 31 March 2021.

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.[47] 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.

E. The Limits of Content Moderation

Tech companies are investing far more in efforts to reduce digital harm rather than promote prosocial content. By the end of 2022, an increasing number of tech insiders and analysts expressed dismay at the limits of content moderation.[48] Moderating user-generated content is expensive, slow, and requires a vast global infrastructure because of the inability of AI automation to identify content to remove.

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.[49] 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.[50]

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

V. The *Tech Regulation* Narrative

Tech critics hold a different view. While user behavior and demand generate harmful content, and there are also offline factors fueling polarization, most large social media products are not a simple or accurate mirror reflection of society.[51] Rather than blaming users for harmful content online, tech critics point to how the affordances and algorithms of tech products incentivize and reward harmful content. Instead of focusing on the "symptom" of harmful content, tech critics argue the focus should be on regulating tech. The tech regulation narrative suggests that regulation should go beyond privacy and antitrust issues to address tech profit models and the affordances and algorithms baked into the design of some social media products.

---

[47] Tiffany Hsu and Eleanor Lutz. "More Than 1,000 Companies Boycotted Facebook. Did It Work*?" New York Times. (*1 August 2020).

[48] See for example, Ravi Iyer. "Content Moderation is Dead." *The Psychology of Technology Newsletter*. (October 7, 2022); CRS. Online Content Moderation and Government Coercion. *Congressional Research Service.* Legal Sidebar. (13 May 2022).

[49] Alberto Acerbi, Sacha Altay, Hugo Mercier. "Research note: Fighting misinformation or fighting for information?" *Harvard Misinformation Review.* (12 January 2022).

[50] Mosleh, M., Martel, C., Eckles, D., & Rand, D. 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*, (May 2021), 1–13.

[51] Dean Eckles. "Algorithmic transparency and assessing effects of algorithmic ranking." Testimony before the Senate Subcommittee on Communications, Media, and Broadband. (9 December 2021).

A. Optimized for User-Engagement and Profit

The Center for Human Technology insists that technology is not neutral.[52] Tech product design features foster "a race to the bottom of the brainstem" and an "attention economy" that rewards divisive content, resulting in "polarization spills." Tristan Harris of the Center for Humane Technology describes Twitter as a "gladiator stadium, like a Roman Coliseum, where people are being told that they need to debate free speech and ideas in a marketplace of ideas with balls and chains and arrows and swords and goring each other."[53] According to a Pew Survey, a minority of highly active users post most tweets, and nearly half of Twitter users in the US are silent observers of the brutal clash of the most extreme and violent users.[54]

In 2021, Facebook whistleblower Francis Haugen revealed damning internal reports documenting staff concerns that the company was driving polarization in countries around the world. Countless researchers and journalists from outlets such as the *World Street Journal* and the *New York Times* were documenting the evidence.[55] In 2020, the *Wall Street Journal* published an article claiming Facebook was ignoring polarization or undermining efforts to work on it. The article suggested Facebook's internal research in 2019 and 2020 found that its algorithms were increasing polarization by exploiting "the human brain's attraction to divisiveness." The article cited a 2018 slide from an internal presentation that noted that "if left unchecked [Facebook algorithms optimized for profit would offer] more and more divisive content to gain user attention and increase time on the platform." The article stated that Facebook researcher and sociologist Monica Lee gave a presentation in 2016 that detailed how Facebook was fueling extremism. The 2016 slides state that extremist content that is "racist, conspiracy-minded and pro-Russian" is found in a third of all large German Facebook groups, and "64% of all extremist group joins are due to our recommendation tools."[56]

Tech critics point to engagement-based profit models that incentivize and optimize for polarizing and extremist content that keeps users engaged with emotional content. Harvard Business School Professor Shoshana Zuboff refers to the social media profit model as *surveillance capitalism*. Tech companies capture more private user information and attention to ads the longer a user stays on the product and the more they engage. User-engagement metrics translate to profit as ad companies pay more to access more users.[57]

The Center for Humane Technology calls this the "race to the bottom of the brainstem." The user-engagement profit model driving social media tech companies translates into "design choices that will create a more addicted, distracted, outraged, polarized, validation seeking, and narcissistic society" while also leaving people vulnerable to domestic and foreign political actors waging divisive psychological influence campaigns.[58] Harris states, "Twitter's business model of engagement is about making sure that every post, every moment of anger, every moment of controversy is as maximally visible and interactive with as many other people as possible."[59] Just as CNN found that its profits increased when it offers round

[52] The Center for Humane Technology. "The Myth of Neutrality." (31 March 2022).
[53] Tristan Harris. "Humane Technology on 60 Minutes." *Your Undivided Attention Podcast*. (10 November 2022); Tristan Harris and Aza Raskin, Center for Humane Technology. "Elon, Twitter and the Gladiator Arena." *Your Undivided Attention Podcast*. (27 October 2022).
[54] Meltem Odabas, "5 facts about Twitter 'lurkers'" *Pew Research Center*. (16 March 2022).
[55] John D. McKinnon and Ryan Tracy. "Facebook Whistleblower's Testimony Builds Momentum for Tougher Tech Laws." *The Wall Street Journal*, 5 October 2021.
[56] 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.
[57] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. (New York: Public Affairs. 2019).
[58] Tristan Harris. "Humane Technology on 60 Minutes." *Your Undivided Attention Podcast*. (10 November 2022).
[59] Ibid.

the clock crisis coverage, social media companies also profit more when their "trauma inflating" algorithms that amplify anger and injustice.[60]

The film *The Social Dilemma* portrays three challenges social media poses for society. First, there is a *mental health dilemma* that relates to internet addiction, depression, anxiety, and a loss of an ability to have agency, or the ability to make decisions. Second, there is a *discrimination dilemma* that relates to the subjective biases and prejudices in algorithms that amplify oppressive dynamics. Third, there is a *democracy dilemma* that relates to the role of some tech products in undermining public trust in democratic institutions, public interest journalism, and elections.

Tech design affordances may reduce individual agency, a marker of social cohesion. *Engagement-driven affordances* such as "likes" and "shares" fuel social comparisons and foster greater use or even addiction-like obsessions, deteriorating mental health and well-being, and keeping users scrolling rather than taking actions to improve the quality of life for themselves and their communities. *Engagement-driven algorithms* rank content to promote and recommend divisive content designed to keep people on the tech product longer. *Engagement-driven data collection* on user location, ideas, behaviors, beliefs, networks, and identities creates a databank of information that governments and political actors can use to surveil the public. This surveillance may fuel distrust between the public and institutions with access to their data. Institutions may use this surveillance to repress certain identity groups or civil society groups advocating for human rights or democracy.

B. Disrupting Public Interest Journalism

In addition to tech affordances and algorithms that seem to amplify polarizing content, there are also a wider set of digital impacts on social cohesion.[61] Digital advertising is diverting money away from public interest media where it helped to fund local news and investigative journalism.[62] The decline in the availability and quality of legacy media (newspapers, radio, TV) may enable disinformation online and offline to spread more freely and rapidly.

Media fragmentation both online and offline interacts to reinforce users encountering similar conspiracies, partisan, or false information in a hybrid online/offline media ecosystem that reinforces political divides. Public surveys document a decline in public trust in journalism.[63] At the same time, the growth of partisan media, a growing disinformation industry, and state-based cyber troops waging *cognitive warfare*[64] on both domestic and foreign publics all seek to maximize algorithmic rewards for outrage and division. Together, these forces seem to contribute to *epistemic insecurity* where the public is unsure who to believe or what is true.[65]

Cumulatively, the design of digital spaces and their optimization for user engagement and profit comes at the expense of social cohesion.

C. The Limits of Tech Regulation

[60] Tristan Harris and Aza Raskin. "Can Psychedelic Therapy Reset Our Social Media Brains? with Rick Doblin. *Your Undivided Attention Podcast*. (15 December 2022).
[61] Lisa Schirch, Digital Space and Peace Processes. (Geneva, Switzerland: Interpeace, Fondation Hirondelle, ICT4Peace. May 2022).
[62] Derek Wilding, Peter Fray, Sacha Molitorisz, Elaine McKewon, The Impact of Digital Platforms on News and Journalistic Content, (Sydney: University of Technology NSW, 2019).
[63] Katherine Fink. The biggest challenge facing journalism: A lack of trust. *Journalism*. Vol 20, Issue 1, (2019).
[64] Samuel C. Woolley and Philip N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media.* New York: Oxford University Press, 2018. See also François du Cluzel, *Cognitive Warfare*. Brussels: NATO Innovation Hub, (2020).
[65] Emanuel Adler and Alena Drieschova. The Epistemological Challenge of Truth Subversion to the Liberal International Order. International Organization, 75(2), (2021), 359-386.

Perhaps unwittingly, social media products are a defacto digital public sphere; a space for discussion of issues that affect people's lives. But the engineers and tech innovators who created most social media products had no training or preparation to design a digital public sphere. Few companies consulted social scientists. Most tech companies lack staff with appropriate backgrounds to anticipate and respond to governing new digital town squares or addressing toxic polarization. Tech companies also lack the political legitimacy to do the policing of these new town squares, particularly for moderating "political entrepreneurs" who use polarizing political messages to instill fear and division within potential voters, and for industrial-level harmful digital content created by cyber armies and disinformation industries.[66]

A Congressional Research Service report offers a summary of the Townsquare doctrine; a legal theory that says that certain types of spaces even though they might be technically private, still have certain types of protections for freedom of expression. In other words, when a technology product grows and becomes widely used as a town square, it has public responsibilities.[67] Most governments are not yet prepared to keep up with tech innovations that create new public spaces that demand new types of guidelines and regulations. Tech companies face dilemmas to define the limits of free speech on their products and the social norms for these spaces.[68] While Lessig implored humanity to understand that "code is law," he also writes about the inability for government laws to adequately regulate code. Digital spaces have become resistant to regulation.[69]

To users and government regulators, a company can tout its product as a neutral communication platform where anyone can communicate. To advertisers and investors, a company can tout its product as an "advertising" or "marketing" platform where "users" and their private information and attention are the product being sold.[70] It might take researchers 20 years to determine exactly how much technology companies are responsible for harming human agency, polarizing communities, and undermining trust in democratic institutions. But there is a precedent for not waiting for the absolute scientific consensus on tech impacts on polarization when the stakes are so high.[71]

Government regulation of tech companies to date has focused on privacy and cybersecurity concerns, not the affordances and algorithms that amplify toxic polarization. Tech products optimized for user engagement, advertising, and profit incentivize the spread of false and hateful posts. Regulating algorithms can curb tech platform's prioritizing profit over people. Yet the speed of the movement for government regulation of technology platforms harmful impacts on society is nowhere close to catching up to the impacts of digital amplification of disinformation and hate speech on intergroup relations, the escalation of threats to electoral integrity, or the decline in public trust in institutions.

VI. The *Social Cohesion by Design* Narrative

A third way of approaching technology companies' roles in responding to or preventing harmful content goes a step further. In *addition* to content moderation and tech regulation, tech companies can design tech products with affordances and algorithms that support social cohesion.

---

[66] Jennifer McCoy and Murat Somer. "Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies." The Annals of the American Academy of Political and Social Science. Volume: 681 (1), (20 December 2018), 234-271.
[67] Valerie C. Brannon. "Free Speech and the Regulation of Social Media Content." *Congressional Research Service*. (27 March 2019).
[68] Valerie C. Brannon. "Free Speech and the Regulation of Social Media Content." *Congressional Research Service*. (27 March 2019).
[69] Lessig (1999).
[70] Tarleton Gillespie. "The Politics of 'Platforms'" *New Media and Society*. Vol 12(3). (2010), 47-364
[71] Jonathan Haidt. "Yes, Social Media Really is Undermining Democracy: Despite What Meta has to say." *The Atlantic*. (28 July 2022).

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.

At the 2022 Trust and Safety Research Conference at Stanford University, former Twitter vice president of trust and safety Del Harvey urged tech companies to look beyond content moderation toward designing for health. Harvey oft described as Silicon Valley's "chief sanitation officer" for her role in removing harmful digital content,[72] explained that it is not enough for tech companies to remove the sickness of harmful content. Tech companies could learn from public health lessons to move beyond reaction to prevention. When asked if tech companies had an appetite to "design for health" most panelists indicated they did not see evidence of such interest.

Several interviewees for this report noted that there had been some internal experiments to incentivize positive content to build social cohesion. Citing concerns from conservatives in the U.S., Facebook hired Republican leader Joel Kaplan in 2011 to be policy chief and to vet proposed changes. Kaplan expressed concern that changes to encourage better communication skills were "paternalistic," calling the vetting process "Eat Your Veggies." According to the *Wall Street Journal* and tech staff interviewed for this report, Kaplan approved some changes but blocked other proposals because they lacked "rigor and responsibility" related to effectiveness and possible unintended consequences. Facebook ended the Common Ground initiative citing political bias, social engineering, and cognitive manipulation. The Common Ground team disbanded around the end of 2018. The central Integrity Team also disbanded, though other dispersed integrity teams continued.[73]

Yet despite robust attention to a Trust and Safety approach that would design for civic health from within Silicon Valley, elsewhere in the world, a range of new pro-social technologies or "peacetech" aim to decrease polarization, improve social cohesion, and advance computational democracy offer a compelling alternative paradigm for thinking about Trust and Safety or the dilemmas of content moderation.

eBay's dispute resolution product designer Colin Rule believes that computer code can itself act as a mediator and offer people the agency and capacity for being good to each other. A tech product can provide the structure to coach users on what they can say to have the best chance of a positive encounter with another person. In this case, Rule asserts that tech product designs are a form of "benevolent manipulation." As system designers, tech companies can provide the "walls" to structure positive behavior and individual agency to enhance social cohesion online. Other tech products could learn from eBay's example of coaching users to communicate more effectively.

Rule estimates that 90% of individual bad behavior is from someone with a first offense. A warning and "the first mistake is free" approach offers education about the community guidelines and a video or some other educational tip so users do not re-offend. Rule notes that tech products could send a message describing why a piece of content was harmful. For example, a prompt could tell users "you said a hurtful thing on the forums, and people were upset about it, so your content got flagged. Please watch this short video to learn more about healthy conflict and effective communication." If someone makes another offense, these users could have reduced access to the product or forum. They might, for example, only be allowed to post 30 times a month. And the next time they offend, they could only post 15 times a month.

[72] Kashmir Hill. "Meet Del Harvey, Twitter's Troll Patrol." *Forbes Magazine*. (2 July 2014).
[73] 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.

And then on the fourth offense, they might be "de-platformed" or lose all posting ability for 6 months. Building on this example, other tech companies could use harmful content as an opportunity to coach users in effective communication. This might not be welcome but might fuel less anger and rebellion than perceptions of censorship.

Peacetech is an umbrella term referring to forms of technology that improve social cohesion. Peacetech enables pro-social content. It is part of a broader field of *public interest technology* that uses technology to advance the public interest, generate public benefits, and/or promote the public good. Peacetech may include other public interest technologies such as *civictech* which informs citizens on public interest issues and services, connect people with others, and facilitates communication with their government; and *govtech* which helps governments to facilitate communication with citizens to improve public services and public engagement.

VII. The History and Scope of PeaceTech

The idea to design technology to support social cohesion has a long history dating to the 1980s.[74] The term "peacetech" emerged from multiple places in the early 2000s.[75] The Swiss think tank ICT4Peace began research on peacetech in 2003. In 2004, the US Institute of Peace in Washington, DC initiated what is now known as the PeaceTech Lab. In the mid-2000s, local tech innovators in Sri Lanka and Kenya designed tech products to support early warning of violence and citizen journalism. In 2007, the tech company Ushahidi began using tech for the prevention of election violence. In the same year, Stanford psychology professor BJ. Fogg began teaching courses and researching ways technology could be used to support peace, which he called "peace technology."[76] Building on this research, Stanford Peace Innovation Lab continues to create models for peacetech.

In 2020, the UN Secretary-General released a Roadmap for Digital Cooperation, detailing a robust "digital transformation" agenda supporting the innovation of new tech products that support the UN's Department of Peacebuilding and Political Affairs. The United Nations is also investing in a suite of technology tools to support the UN Department of Political and Peacebuilding Affairs (DPPA). Countless NGOs are also exploring peacetech and digital peacebuilding. The NGO Build Up works with partners around the world to support civil society in learning how to use peacetech and authored Search for Common Ground a Digital Peacebuilding Guide which provides insight into how to choose what type of technology to use to support social cohesion. The NGO swisspeace conducts research and an online course on digital peacebuilding. The Alliance for Peacebuilding hosts a community of practice on "Digital Peacebuilding" that offers monthly meetings to learn about new types of peacetech.

There are now centers around the world devoted to peacetech, including the University of Waterloo's Grebel Peace Incubator, the University of Bristol's Interdisciplinary PeaceTech Group, and the University of Notre Dame's PeaceTech and Polarization Lab. In Florence, Italy, the European University Institute held the first Global Peacetech Conference in November 2022.[77]

---

[74] For a longer history, see Lisa Schirch. "25 Spheres of Digital Peacebuilding and PeaceTech." Tokyo: Toda Peace Institute, (2020).

[75] See for example, Yiannis Laouris *Information Technology in the Service of Peacebuilding: The Case of Cyprus. World Futures* 60(1). (December 2003), 67-79; Helena Puig Larrauri and Anne Kahl. "Technology for Peacebuilding." Stability: International Journal of Security and Development, 2(3), (2013); Ioannis Tellidis & Stephanie Kappler. "Information and Communication Technologies in Peacebuilding: Implications, Opportunities and Challenges." Cooperation and Conflict. Vol. 51(1) (March 2016); Pamina Firchow, Charles Martin-Shields, Atalia Omer, and Roger Mac Ginty. "PeaceTech: The Liminal Spaces of Digital Technology in Peacebuilding." *International Studies Perspectives*, Volume 18, Issue 1, February 2017; "Social Media Impacts on Social & Political Goods: A Peacebuilding Perspective." L. Schirch. Toda Peace Institute. Policy Brief #38. (April 2019);

[76] B.J.Fogg, *Peace Technology*: *Why a class about Facebook Apps?* Scribd. (2007). Found at https://documents.pub/document/dr-bj-fogg-facebook-peace-technology.html Accessed on 1 January 2023.

[77] See Kalypso Nicolaidis and Michele Giovanardi. Global PeaceTech : unlocking the better angels of our techne, EUI RSC, 2022/66, Global Governance Programme-481, Europe in the World.

VIII. Functions of Pro-Social PeaceTech

Prosocial technology contributes to social cohesion in four broad ways. First new tech platforms help to analyze digital harms and polarization. Second, technology products can improve human agency to participate in civic issues affecting their lives. Third, technology products can support intra-group and inter-group communication and joint problem-solving. Fourth, tech platforms can improve public trust and inclusion in governance.

A. Tech for Analyzing Digital Harms and Polarization

Understanding the dynamics of polarization is an essential element of planning effective social cohesion programs. Each context has a unique information ecosystem and a unique set of conflict dynamics. For many decades, conflict analysis and context assessment tools have been essential to developing effective peacebuilding and development programs.[78] Growing polarization and state-sponsored disinformation campaigns highlight the need to add an analysis of information ecosystems and how digital spaces and their interaction with offline spaces may be driving conflict.[79]

Strategic planning on the use of digital tools to support social cohesion begins with first analyzing information ecosystems.[80] The United Nations is investing in a suite of technology tools to support social media analysis.[81] Sparrow is a social media analysis tool created by and for the UN Department of Political and Peacebuilding Affairs (DPPA) for analyzing Twitter to identify trending topics, hashtags, and key influencers.

Another example is called Phoenix. The peacebuilding NGO Build Up and the technology company DataValuePeople partnered to create Phoenix, an open-source, non-commercial, customizable process and tool to support peacebuilders and mediators who want to work ethically with social media data to inform programming. Local communities first develop contextually grounded problem statements that address peacebuilding objectives. The groups then use Phoenix to create a data pipeline to add social media sources, along with labeling and visualization tools. Phoenix offers new ways to understand the drivers of conflict and the opportunities for peace.

B. Tech to Support Individual Agency

Some tech products support individual agency so that people have the capacities and belief that they can participate in civic action to work on issues that affect their lives. These platforms can help people feel that they have a voice by providing tools for them to share their identity, experiences, beliefs, and passions. Some platforms offer affordances such as hashtags to enable isolated individuals to find each other to form larger movements, such as with the hashtags #MeToo and #BlackLivesMatter.[82]

Other tech products help individuals to reality check their perceptions to help individuals recognize they there is more common ground between people than commonly assumed. For example, digital quizzes

---

[78] Lisa Schirch. *Conflict Assessment and Peacebuilding Planning: Toward a Participatory Approach to Human Security*. (Boulder, Colorado: Lynne Rienner Press, 2013).
[79] Fondation Hirondelle, Demos, Harvard Humanitarian Initiative and ICREDES. "Influencers and Influencing for Better Accountability in the DRC." (July 2019).
[80] Branka Panic, *Data for Peacebuilding and Prevention Ecosystem Mapping: The State of Play and the Path to Creating a Community of Practice* (New York: NYU Center on International Cooperation, 2020).
[81] See for example, United Nations, *Digital Technologies and Mediation in Armed Conflict.* Helsinki: Department of Political and Peacebuilding Affairs; Centre for Humanitarian Dialogue, 2019; Global Pulse, *E-Analytics Guide: Using Data and New Technology for Peacemaking, Preventive Diplomacy and Peacebuilding* (New York: United Nations, 2019).
[82] Sarah J. Jackson, Moya Bailey, Brooke Foucault Welles. #HashtagActivism: Networks of Race and Gender Justice. Cambridge, MA: MIT Press, (2020).

such as The Perception Gap, developed by the bridge-building organization More in Common, offer individuals an online opportunity to reflect and test whether their perceptions of other groups match reality.[83] People in different countries could take the quiz and find out how realistic their view was of people on the other end of the political spectrum. This serves an important role in "reality testing" and challenging people's presumptions about other groups. Helping individuals realize they do not accurately understand their political opponents might prompt them to be curious to learn more so that they may correct their perceptions and understanding.

There are a variety of tech products to help people learn effective communication skills and to model how to have a healthy conversation on difficult issues or conflicts. For example, Games for Peace uses Minecraft games between Israeli and Palestinian youth.[84] In addition, individual influencers are offering conflict resolution advice on products like TikTok using hashtags such as #resolveconflict. Another example comes from Karin Tamerius and her group called Smart Politics created an "Angry Uncle" Chatbot to help coach people in effective communication skills for having political conversations at holiday dinners.[85] The Canadian-based Suzuki Foundation created a climate conversation coach bot called CliMate. Other organizations are organizing cooperative video games between groups in conflict.[86]

Based on his experience building eBay's Online Dispute Resolution (ODR) system, Colin Rule helped to set up the mechanism for eBay users (sellers and buyers) who had disputes. Rule found in his dispute resolution work with eBay is that if you have a dispute between a buyer and a seller, and you give an open text box to the buyer, they are the complainant, this is not helpful. Tech product designers do not need to let "everyone" talk to "everyone." This gives people too much ability to generate more anger and havoc for themselves and others. They may engage in threats and insults because that is the way they think they can get a sense of fairness and they are angry and frustrated and want the other side to understand that in hopes of influencing their response. Instead of giving the complainant an open textbox where they vent that anger, instead tech products can structure more constructive communication by giving them a forum where they can go through and make selections. What kind of problem do you have? What kind of solution do you want?[87] Users leave with a positive sense of resolution and empowerment, a key element of social cohesion. eBay has resolved millions of disputes through this system. eBay bots coach complainants to rephrase and reframe their messaging to take out insults. The seller has an incentive for that buyer to be happy because the buyer is unhappy, and they leave them negative feedback that's going to impact their ability to sell on the site.[88]

Similarly, some have suggested that popups, a box, symbol, or window that appears when you begin writing on a computer, might offer users feedback on their tone. On Twitter, such a concept includes informing users with a popup stating, "I see you might be headed for an uncivil conversation"?[89]

C. Tech to Support Horizontal Cohesion

As described above, there are two forms of horizontal cohesion. Intra-group cohesion is known as "bonding" social capital. Inter-group cohesion is known as "bridging" social capital.[90]

---

[83] See The Perception Gap at https://perceptiongap.us/
[84] See Games for Peace at https://www.gamesforpeace.org/
[85] Karin Tamerius. **"**How to Have a Conversation With Your Angry Uncle Over the Holidays." *New York Times.* 18 November 2018.
[86] David Suzuki Foundation. "How and why to have climate change conversations." Accessed at https://davidsuzuki.org/what-you-can-do/how-and-why-to-have-climate-change-conversations/
[87] Amy J. Schmitz and Colin Rule. "Lessons Learned on eBay" in The New Handshake: Online Dispute Resolution and the Future of Consumer Protection. *American Bar Association Section on Dispute Resolution*. (2017), 33 – 46.
[88] Interview with Colin Rule, February 18, 2022.
[89] Molly Wood. "Twitter hires social scientists to help figure out our conversation problem." *Marketplace.* (25 September 2019).
[90] Robert D. Putnam, *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster, (2020).

There are several examples of new tech startup companies that focus on intra-group bonding, particularly for individuals meeting online who are meeting for social or work purposes. Gatheround is a video conferencing tech product that describes itself as "a team bonding and community engagement platform for people-focused organizations seeking to build relationships and strengthen teams in an era of disconnection and distraction." Co-founded by Lisa Conn, formerly director of the Common Ground Initiative at Facebook, described Gatheround as "Unlike Zoom" in that it is "designed for how humans connect. We make it simple and effortless to bring people together online in a way that's both wildly fun and deeply meaningful."[91]

As in real life, individuals on Gatheround do not see themselves, just the other participants to whom they are talking. Participants are encouraged to be kind to each other by the video focusing on people at a "nose-biting" distance. Gatheround offers conversation prompts for people to share their experiences, so they feel more heard and seen. There are no affordances to mute or turn off the camera, making it impossible for people to check out of the conversation. There are no backgrounds so people on the product see the context of where you are sitting. Gatheround has a share/facilitation feature where, like a talking stick in a dialogue, a question is asked, and people form a line with equal time to speak. This ensures that existing power dynamics and structures might be altered on the product, providing more equity to all participants.[92]

A second example of tech-supported intra-group cohesion is Marco Polo, a social media product focused on well-being and happiness in a closed social network. Marco Polo offers a video chat or video voicemail with a front-facing camera. Marco Polo emerged from a sense that people had turned to lower-quality text-based communication and had stopped calling each other and having conversations. Text-based products may increase the *quantity* of relationships at the cost of the *quality* of relationships. Co-founder Vlada Bortnik describes the "thoughtful, human-centered design" as focusing on the quality of the connection. As a person records a video chat, they look into the camera, like looking in the mirror. There are no filters or glamor, but rather an encouragement by design to be authentic and intimate. This may make it more likely to present positive body language and less likely to spew hate at someone. The home screen in Marco Polo is chronological. There are no counts of friends, likes, or emojis, as the product does not want to have "vanity metrics." The experience in Marco aims to be enriching and nourishing, to increase happiness, and attempt to curb the epidemic of loneliness. Marco Polo does not want to increase anxiety by urging users to have competitive streaks. Instead, the design aims to be a tool for humans intrinsically motivated to use rather than manipulative of a person's time. Marco Polo does not sell user data and does not advertise on Google or other monetized products that collect information because that would allow Google to gather user information and Marco Polo wants to protect the privacy of users. The staff at Marco Polo assert that because people on Marco Polo connect more intentionally with their closest friends, the threat of encountering harmful content is lower. Marco Polo does allow people to block someone in their network. But because the video chat is asynchronous, a user cannot talk over somebody. Marco Polo aims to be a product that encourages people to listen to each other.[93]

Other tech-support platforms aim to improve inter-group cohesion to build "bridging" social capital between people that may belong to different social groups. Intergroup relations can improve in a variety of ways. These tech products offer affordances for people to explore their differences as well as their common ground. Some products seek to create safe or "brave" spaces for dialogue across the lines of conflict to build trust and empathy. The examples here illustrate that technology can scale empathy and understanding between groups, as well as increasing the group's capacities for solving problems together.

[91] Interview with Lisa Conn, December 18, 2021.
[92] Ibid.
[93] Interview with Vlada Bortnik, March 2, 2022.

For example, the tech designers at Soliya set out to pair technology with the power of dialogue in 2003, before the start of most other social media products. Soliya is a Virtual Exchange product to foster "high impact inter- and cross-cultural education facilitated through digital technology." Soliya hosts dialogues between 15,000 young people per year in small, diverse groups to share their perspectives on identity and current events. Soliya has held a special focus on intercultural dialogue between young adults in the West and the Arab and Muslim World. Soliya is unique in part because participants' videos show up in a circle, surrounding a prompt for the dialogue that asks a question or keeps the participants focused on a topic.[94]

Another approach to building social cohesion is to invite people to engage in a conversation not to prove who is right or wins the argument but rather to see who can change another person's views on a topic. This exercise requires users to listen to other points of view and then try to build a bridge between worldviews. Scottish teenager Kal Turnbull created Reddit's "ChangeMyView" community which invites people to "post an opinion you accept may be flawed, to understand other perspectives on the issue and to encourage users to enter with a mindset for conversation, not debate."[95] Users rewarded compelling arguments with a delta symbol (Δ) to indicate when someone changed their mind.

Turnbull extended the subreddit community by creating a new website. "Change a View helps internet commenters see eye-to-eye, where the forum breaks us out of our online filter bubbles, and where we relearn how to talk to each other online." Users note the digital space feels like an "oasis" while journalists call it "our best hope for civil discourse." Change A View uses Jigsaw's comment-ranking engine, called Perspective API, which scores comments, demotes harmful content, and eases moderator loads. Change a View provides a template for how to improve difficult conversations online. It could be adopted by news agencies for discussing news articles, or in major platforms when people in a conversation begin moving toward harmful content.[96]

Researchers examine the affordances of "ChangeMyView" that enable effective communication on the platform, namely the "game" elements and the social norms of the community. Gamification is a method of turning an activity into a game to increase motivation. Gamification provides enjoyment and social approval through competition, with the award of a delta sign, which accumulates into "delta scores." Participants told researchers that the incentive of earning a delta encourages them "to be civil to one another." Users observed that the people who were able to change the views of others were "polite in their posts." Users also noted the role of moderation of trolls and people there who were rude or not open-minded.[97]

The peacebuilding organization Build Up launched "The Commons" as an intervention to depolarize political conversations on Twitter and Facebook in the USA. Paid facilitators with the project initiated thousands of conversations across some of the most polarized individuals and polarizing topics. The goal was to help people engaged in polarized conversations to have an experience of a more positive conversation, to increase interest in taking further action to promote civility, and to shift how they engage with people on social media. Facilitators found polarized conversations by curating a list of top hashtags and content creators or influencers at the center of US political conversations. Bots would then identify who was open to a conversation, and then humans would engage with them to attempt to depolarize.[98]

---

[94] Marta Guarda. "Giving voice and face to other cultures: the Soliya Connect Program and the development of intercultural communicative competence". In Carte d'Occasione vol. V, Padova: Unipress, (2013), 111-131.
[95] https://www.reddit.com/r/changemyview/
[96] Arielle Pardes. "'Change My View' Reddit Community Launches Its Own Website." *Wired.* (6 April 2019).
[97] Shagun Jhaver, P. Vora, and A. Bruckman. *Designing for Civil Conversations: Lessons Learned from ChangeMyView, GVU Center Technical Reports*. (2017).
[98] Anooj Bandari. "The Commons: Where are we at in 2021? *Medium.* (27 September 2021).

There are a wide variety of other new tech startups aiming to improve intergroup dialogue. Some are exploring how to use virtual reality to foster intergroup dialogue and empathy. The group HackthePlanet offers a variety of VR programs to build intergroup understanding.[99] A platform called Kazm bills itself as a "conversation engine." Kazm describes itself as a social platform built for community, as opposed to other platforms built for an audience. Kazm's affordances offer support and tools for dialogue facilitators and community administrators, saying, "At Kazm, we believe real communities are built on conversations - but managing conversations is hard work. We make it easy."[100] Kazm offers bridge-building and dialogue groups like Living Room Conversations affordances including a way to have members connect via video dialogue, join events, access content, and comment on discussion boards. Kazm also offers video coaching to guide the conversation, prompts to focus a dialogue, videos, polls, and word clouds. Noting its role in social cohesion, Kazm states that there are no ads or trolls on Kazm.[101]

Few of the tech start-ups designing new products to support social cohesion are reaching the scale necessary to address toxic polarization. Big tech companies with the scale to shift societies away from polarization and toward social cohesion will need to learn from and adapt the design affordances and algorithms from smaller startup tech companies.

For example, Twitter drew inspiration from Pol.is to create incentives for individual agency and participation in negotiating the validity or truthfulness of digital posts. Pol.is engineers optimized the platform to contribute to social cohesion and to put guardrails on the platform to limit harmful content. Learning from Pol.is' affordances and algorithms, Twitter staff developed a program called Community Notes (formerly Birdwatch) to empower Twitter users to add helpful notes to Tweets that might be misleading. Wired Magazine calls Twitter's experiment "one of the most exciting content moderation innovations ever to come out of not just Twitter, but any major platform."[102]

Aviv Ovadya and Jonathan Stray have been writing about the potential of big tech companies to adopt the types of bridging ranking systems found in platforms like Pol.is and Remesh.[103] There are more opportunities for big tech companies to test the use of affordances and algorithms in divided communities.

D. Tech to Support Vertical Social Cohesion and Public Trust

Other new tech startups aim to improve vertical cohesion by enabling citizens to participate in governance through" civtech" which enables citizens to engage in collective problem-solving on policy topics, and "govtech" by enabling governments to design more inclusive processes for consulting with citizens on public issues.

These tech products recognize that social cohesion does not require preventing the expression of tension or conflict or making people be superficially "nice" to each other. Social cohesion also does not require people to make personal relationships with each other, or for that matter to have any direct contact with each other.

---

[99] See https://www.hack-the-planet.io/
[100] See https://about.kazm.com/
[101] See "Scaling Facilitated Dialogue with Conversation Engines." Alliance for Peacebuilding. (16 November 2021). *YouTube.* https://www.youtube.com/watch?v=qwkhc_8jZaI
[102] Carl Miller. "Elon Musk Embraces Twitter's Radical Fact-Checking Experiment." *WIRED Magazine*. 28 November 2022.
[103] 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.

For example, the Ushahidi platform enables citizens to report potential violence to governments, enabling violence prevention efforts. Ushahidi is a crowdsourcing and mapping tool that functions on simple mobile phone applications allowing citizens to report potential outbreaks of violence. In 2007, Ushahidi helped to prevent election violence in Kenya. Local people reported where tensions were rising in the streets. This information was shared with local civil society mediation and peace teams as well as police. Since then, Ushahidi has grown significantly and now the platform is used to enable citizen reporting and coordination about civil society and governments to respond to public issues. It provided real-time information to defuse electoral-related violence in the streets.[104] Ushahidi has been used in Haiti and Nepal to coordinate relief efforts, monitor and report on corruption in Indonesia, help address sexual violence in Egypt, and map police violence in Portland, Oregon.

Other platforms enable inclusion and participation in decision-making by making it easier for people to participate and by creating incentives to identify common ground or consensus. The 2014 tech start-up Remesh began with the mission to create a technology that would "represent the will of the people and amplify their collective voice." Conflict mediators, civil society groups, or governments can use Remesh to dialogue with and poll the public. Remesh software can extract key themes and draw insights from a dynamic and open-ended "conversation" with up to 1,000 people.[105] The UN used Remesh in Libya to gather stakeholder opinions on a proposed interim government. In Yemen, the UN used Remesh to listen to public perceptions of a cease-fire and opinions on the prospects for a peace process. The UN is now considering using Remesh for peace support in Sudan, Mali, Afghanistan, and Iraq.[106]

Inspired by insights from social cohesion efforts in nonviolent communication and attempts at collective decision-making in the Occupy Movement, Colin Megill designed the tech platform Pol.is to improve computational democracy. Experiments in Taiwan and the UK showed that Pol.is could help a divided public find areas of common ground and develop policy solutions on polarized public issues. In Taiwan, the government has used Pol.is dozens of times on different issues and resulting in government action 80% of the time.[107]

The Polis platform is designed to be optimized for consensus building, finding common ground, and fostering citizen engagement. Polis provides "a real-time system for gathering, analyzing, and understanding what large groups of people think in their own words, enabled by advanced statistics and machine learning." Polis enables "collective intelligence" and fosters mutual "listening at scale" through digital citizen assemblies that use tools to support "computational democracy." Polis can "listen at scale" by enabling thousands or even millions of people to engage in conversation with each other. The platform gathers both qualitative data and quantitative data. Unlike other platforms, on Polis users do not reply to each other's posts. Rather users submit an idea (one at a time) that others can up-vote or down-vote. This affordance enables users to reward ideas that address the interests of most people and generate new and better solutions. The lack of a "reply" affordance prevents trolling and abuse, and thus removes the pain and heat from discussions. Polis operates on open-source code allowing anyone to use the platform to host public dialogues seeking to find consensus.[108] Pol.is seems to incentivize participants on the platform to develop creative options that meet the interests of diverse stakeholders and enables "thinking outside the box" to envision positive future coexistence.

---

[104] Juliana Rotich. "Ushahidi: Empowering Citizens through Crowdsourcing and Digital Data Collection." Field Action Science Reports. no. 16. (2017), 36-38.
[105] Interview with Andrew Konya, March 20, 2022.
[106] Jordan Bilich, Michael Varga, Daanish Masood, and Andrew Konya. "Faster Peace via Inclusivity: An Efficient Paradigm to Understand Populations in Conflict Zones." AI for Social Good workshop at NeurIPS. Vancouver, Canada. (2019).
[107] Josh Smith, Toby O'Brien, Harry Carr. "Polis and the Political Process." *Demos*. (3 August 2020).
[108] Interview with Colin Megill, February 20, 2022.

IX. Conclusion

Toxic polarization online requires a multi-stakeholder, multi-pronged response.

Big tech companies will continue to need to develop new methods of disincentivizing and removing harmful content. Content moderation is difficult and insufficient for addressing the scope of harmful content online. Interviewees for this research with technology staff captured a shared sense of "great concern" about tech-related harms such as polarization and want to "feel good about the company that employs them." Tech staff reported feeling a "huge appetite" for achieving company missions to "connect" people. Even staff at companies who have hired tens of thousands of content moderators expressed dismay at the task of managing a "tsunami of harmful content" without adequate resources, particularly in the Global South where they lack staff who speak local languages.

Tech regulation is also difficult and insufficient. While governments focus on issues like privacy and cybersecurity, the challenge of regulating algorithms and design affordances on tech platforms will likely be slower and more challenging. Governments will need to create incentives for tech companies to reduce harmful content amplified by their algorithms and design features, either by changing their profit model and/or paying taxes on their polarization spills to help fund social cohesion efforts. New tech startups along with private and public funders can invest in building new tech platforms aimed to improve social cohesion.

This leaves several other options for addressing harmful content online. While this paper did not address digital media literacy,[109] there is a movement afoot to strengthen public immunity to harmful disinformation by inoculating people to "prebunk" conspiracy theories and other false and deceptive content online.[110] A mass public education effort will also take time and has only begun in a few countries such as Finland.[111]

Designing technology to support social cohesion is an alternative and a complement to these other approaches. This article makes the case that technology platforms are never neutral. Computer engineers with training in social cohesion designed some of the technology platforms described in this article. Others started as initiatives of the UN or NGOs in partnership with tech startups to create products that would support bridge-building and peacebuilding work.

Addressing the tsunami of false and hateful information online requires this type of innovation – designing and scaling tech products that support social cohesion. Content moderation, tech regulation, digital media literacy, and designing tech for social cohesion can be complementary. There is no one silver bullet to heal toxic polarization online. Together, they offer a way forward to address the *system* and not just the *symptom* of harmful content online.

[109] Media Literacy: A Definition and More, Center for Media Literacy at http://www.medialit.org/media-literacy-definition-and-more

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

[111] Jenny Gross. "How Finland Is Teaching a Generation to Spot Misinformation," *New York Times*, 10 January 2023.