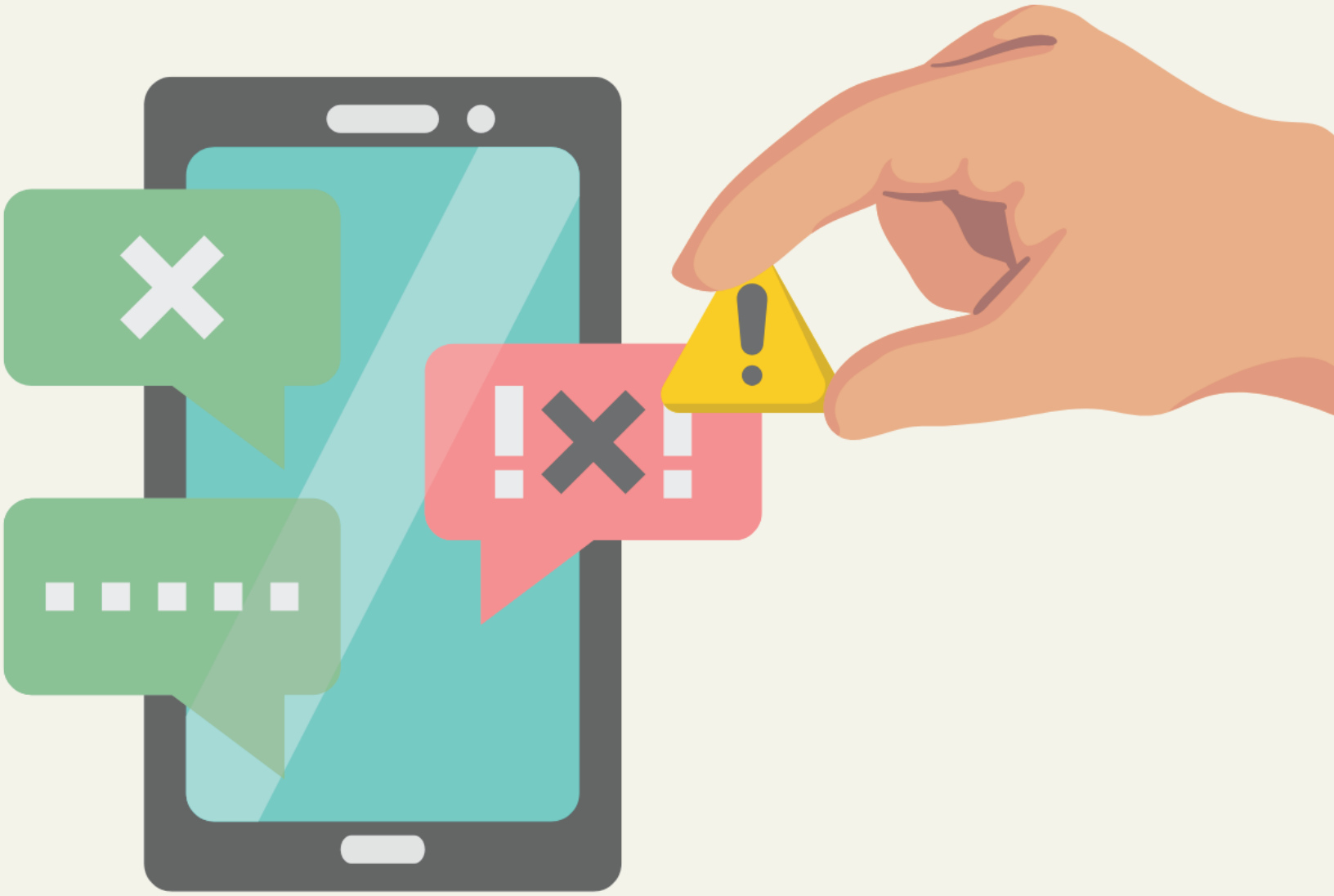
PUBLIC KNOWLEDGE

A POLICY PRIMER FOR FREE EXPRESSION AND CONTENT MODERATION



2025 | JAN

LISA MACPHERSON
MORGAN WILSMANN

A Policy Primer for Free Expression and Content Moderation

Lisa Macpherson & Morgan Wilsmann

Acknowledgements

This paper is the result of a year-long project to revisit seven years of analysis and advocacy for free expression and content moderation and refresh it for a contemporary context. We appreciate the research, recommended readings, active conversation sessions, dialogue and debate, and thoughtful reviews of the ideas expressed here, particularly from Shiva Stella, Will McBride, John Bergmayer, Sara Collins, Harold Feld, Nick Garcia, Elise Phillips, and L'Allegro Smith of Public Knowledge. We also have the deepest appreciation for the many researchers, academics, journalists, whistleblowers, civil society partners, and activists who work so hard to shine a light on the impact of platform content moderation policies and processes on free expression in the United States and around the world.

Table of Contents

Introduction: A New Vision for Free Expression and Content Moderation

A lot has changed since Public Knowledge began a dialogue almost seven years ago about the role of dominant digital platforms in public discourse. Our earliest analysis (like that of many other civil society groups) focused on the concern that platforms would moderate user content *too much*, or without due process for users. At that time, almost daily news reports recounted how content moderation decisions – such as disabling user accounts or removing or "de-monetizing" posted content – left users at a disadvantage. Users found themselves with no recourse and no alternatives because many platform markets were not (and are still not) competitive. While we shared civil society concerns about the hate speech and harmful rhetoric already swirling on platforms, we focused our own analysis on gatekeeper power. The "Santa Clara Principles," unveiled in 2018, constituted one of the first comprehensive frameworks proposed by collective civil society organizations to create accountability for internet platforms' content moderation. They, too, focused on ensuring user rights and called for a highly restricted role for the government in shaping platforms' content moderation approaches.

Public Knowledge unveils a vision for free expression and content moderation in the contemporary media landscape. Our goal is to review important changes in the media, technological, political, and legal landscape – including, most recently, a significant political backlash to the study of disinformation, and the first few Supreme Court cases entailing the role of government in platform content moderation – and describe how policymakers should think about content moderation today. Most importantly, **we will frame the appropriate policy interventions to ensure the right balance between free expression and content moderation while guaranteeing citizens the information they need to enable civic participation.** We will focus on social media and entertainment platforms that distribute user-generated content, search, and non-encrypted messaging channels, all relying on algorithmic curation.

In Part I: Centering Public Interest Values, we provide a brief historical perspective on platform content moderation, review the values that Public Knowledge brings to this topic, and discuss the importance of rooting content moderation approaches and policies in user rights. We also consider our first theory related to content moderation: that user rights should include the right to hold platforms liable if they don't enforce the community standards and/or product features they contract for in their terms of service.

In Part II: Empowering User Choice, we discuss the structure of digital platform markets and the necessity of policy choices that create healthy competition and user choice. We also center digital platforms in the broader ecosystem of news and information, and discuss how policy interventions may offset the impact of poor platform content moderation on the information environment by promoting other, diverse sources of credible news.

In Part III: Safeguarding Users, we discuss an additional array of policy interventions designed to bring about content moderation that respects the imperative for free expression and is in the public interest. These include a product liability theory, limiting data collection and exploitation, and requirements for algorithmic transparency and choice.

In Part IV: Tackling AI and Executing the Vision, we discuss the implications of the new "elephant in the content moderation room," generative artificial intelligence, for free expression and content moderation. We also discuss how our recommended policy interventions can be made durable and sustainable, while fostering entrepreneurship and innovation, through a dedicated digital regulator.

Readers will note that in the interest of brevity and clarity, we have chosen not to describe or link any of the hundreds of thousands of incidents and articles available to us to highlight the failures of digital platforms to effectively moderate violative content on their sites, including hate speech, harassment, extremism, disinformation, non-consensual intimate imagery, child sexual abuse material, and other toxic content. We assume readers will already be familiar with this context and bought into the imperative to forge policy solutions that protect the benefits of digital information distribution platforms while mitigating their harms. Readers who are looking for more information about content moderation can visit our issue page, read more about the harms associated with algorithmic curation of content, and explore why multiple policy solutions will be required to ensure free expression and effective content moderation.

Part I: Centering Public Interest Values

A Very Brief Historical Perspective on Content Moderation

In the early, utopian days of the Open Internet, informal norms and social codes were sufficient to maintain civility in online communities. The earliest computer networks connecting Department of Defense researchers and universities served a highly homogeneous population with similar values. These users could view and experience the online landscape as open, decentralized, democratic, and egalitarian. The preference was for moderation by the *community* itself, using *collaboratively developed norms* instead of centralized rules. This self-moderation approach was relatively easy to accomplish when users comprising the same or similar groups of people largely brought the same lived experiences and worldviews to small and highly specific online forums. But the harmonious homogeneity was short-lived: The introduction of the Mosaic and then Netscape Navigator "browsers" (among other technical developments) brought new audiences, non-institutional computers, and new perspectives onto the internet in droves. By 1995, Netscape Navigator had about 10 million global users.

After several conflicting judicial outcomes about companies' liability for third-party content on their services, Congress passed a new law: Section 230. (It was originally part of a far broader piece of legislation focused on the distribution of pornography, the Communications Decency Act of 1996, most of which was struck down). Section 230 is one of the most consequential – and misunderstood – provisions governing the internet. It shields online services from liability when managing third-party content on their platforms. By doing so, Section 230 allows users to express themselves freely without the threat of over-moderation by online services seeking to reduce their own legal liability.

As internet access widened, it meant that a much larger cross-section of people could be in the same dialogue – and they brought with them different lived experiences, values, and views. Online forums were seen by young activists as the antidote to corporate consolidation in media, increasing suppression of the social justice and anti-war movements, and other political forces. There was an explosion of creativity – especially among communities of color marginalized by established media channels – and the Open Internet's potential for aiding affinity groups and creators to connect, mobilize, and innovate became real.

But the democratic promise of the Open Internet also soon came to be compromised by vitriol, harassment, hate speech, and other forms of online abuse, requiring new forms of content management in order to maintain civility in online communities. Despite their extraordinary contributions to the creation of the internet, its supporting technology, and the connected networks that brought it to life, Black and women users were subject to some of the most violent abuse. In their concomitant quest for scale, a new form of online service provider – platforms – centralized content moderation and sought to make it more efficient. The platforms rising to dominance, most notably Google and Facebook, adopted an advertising-based business model, which encouraged distribution of content based largely on its profit potential. Bad actors rushed in to exploit the economics of provocative and extreme content.

Content moderation took on new urgency in the 2010s with the growth of social media and the speed and ubiquity of the mobile web. "Gamergate" (in 2014) demonstrated how focused online communities could orchestrate devastating harassment campaigns and "Pizzagate" (in 2016) demonstrated the destructive power of online conspiracy theories. Thanks to the work of researchers and journalists, we learned more about the people, rules, and processes that made up the systems of governance of the dominant platforms – "the new governors" of online speech – and "Trust and Safety" became a legitimate career path. Infinite scrolling, notifications, an explosion in video content enabled by 4G networks, and other aspects of the mobile web compounded the ease, scale, and velocity associated with the sharing of content. We learned through the Cambridge Analytica scandal how our personal data could be "harvested" without our informed consent and used for highly targeted distribution of political (and other) ads. Ultimately, thanks to the COVID-19 pandemic and the 2020 election, we also learned about the horrifying real-world harms that could come from platforms' failure to manage strains of misinformation and disinformation effectively. During this same time period, both political and social polarization in the United States increased dramatically.

Through it all, different stakeholders criticized the new speech governors' efforts as being too much. Or too little. Naive. Or corrupt. Politicized. Or indifferent. In an attempt to avoid scrutiny, platforms evolved and re-evolved their content moderation policies, experimented with community-centered approaches, and funded initiatives to make content moderation more independent. A dangerous new counter-narrative put forward by those who use disinformation as a potent political tool led to hearings and court cases claiming any government efforts to collaborate with platforms in the interest of national security and public health were "censorship." Some of these challenges have reached the level of the Supreme Court (where the claims were rejected). Academic institutions and civil society organizations focused on

understanding and mitigating disinformation narratives faced expensive lawsuits and lost a lot of their funding and their talent. All the while, Americans were losing news organizations that use ethical professional techniques to source, verify, and correct their content. Now, citizens are losing faith in not only the free press but also many democratic institutions.

And despite hearing after hearing and wave after wave of legislative proposals in Congress – to ostensibly reform the industry's Section 230 liability shield, regulate algorithms, protect privacy, ensure election integrity, "rein in Big Tech," and save the children from harm – with one exception, *Congress has not passed a single material law regarding platform liability*. (That exception, SESTA-FOSTA, the combined package of the Stop Enabling Sex Traffickers Act and the Allow States and Victims to Fight Online Sex Trafficking Act that passed Congress in early 2018, is a case study in unintended consequences and demonstrates the need for a nuanced approach to platform regulation.)

Bringing Public Interest Values to Free Expression and Content Moderation

One thing that hasn't changed since Public Knowledge's first analysis in this space is the set of core values that we bring to the discussion. Since our founding over 20 years ago, whether the topic is intellectual property or telecommunications or the internet, our bedrock has been the **value of free expression**, including **individual control** and **dignity**. But we also value **safety**, both for individual communities online and for the safety of the conversation itself, including by ensuring privacy through technologies like encryption. We also bring a core value of **equity** – that is, in a pluralistic society with diverse voices, how do we ensure equitable access to the benefits of technology, including the chance to speak? We advocate for **marketplace competition**, which helps promote consumer choice of avenues for expression. And we explicitly seek to support, not undermine, **democratic institutions and systems**.

Such are the values that must be balanced to create content moderation in the public interest. If anything, these ideals have become more important at a time when democratic backsliding is happening in the United States and around the world. From the beginning of the American experiment, civic information and the ability to both express and hear differing views have been among the pillars of democracy, and both free speech and a free press have been protected rights. But both expressing and hearing diverse or differing viewpoints require civility. *We believe the government has an affirmative responsibility to promote an environment that allows this civility, and to further a competitive marketplace that encourages a diversity of views.* Unmoderated harassment and hate speech deter the speech rights of some, and the greatest impact invariably falls on already marginalized communities. These online scourges are also incompatible with the principles of a multi-racial democracy, civil rights, and social justice. Simply put, ***we've learned that free expression for all requires content moderation***.

But better content moderation is also about capitalism and free markets. Unmoderated platforms may serve a specific demand among a subset of internet users, but they can also lack commercial value, as we have recently witnessed in the reduced ad dollars funding X, formerly Twitter. Content moderation standards have the potential to be the platforms' principal means of

competitive differentiation, especially if pro-competition policies like interoperability – which we favor – diminish the importance of network size. And even an "unmoderated" platform is not politically neutral in its impact (we're looking at you, X).

Rooting Content Moderation Policy in User Rights

As we've noted, Public Knowledge's earliest analysis of platform content moderation focused on ensuring user rights, specifically the concern that platforms would moderate content without due process for users. Our perspective was – and remains – rooted in the most basic of constitutional rights, including those found in the First, Fifth, *and* Fourteenth Amendments. Today, users' rights on platforms are bounded by the platforms' terms of service, which represent contracts between users and the platforms. However, these terms of service are mostly designed to give the *platforms* expansive rights, including the right to use all posted or shared content without being liable to the user, and to collect, use, and potentially share extensive user data. They also generally require users to use an arbitration process to resolve disputes. *At a bare minimum,* users should be able to understand the terms of these agreements, understand what they imply in terms of online experience, and expect platforms to enforce them consistently, including providing due process rights for action on content.

A Consumer Protection Theory of Content Moderation

One theory rooted in user rights goes further, holding that platforms should be held liable for defrauding users if they don't consistently enforce the community standards and/or product features they contract for in their terms of service. Infringements would include failing to moderate content that violates the platform's stated community standards, or not enforcing product features such as parental controls. This consumer protection theory calls upon the Federal Trade Commission and other consumer protection regulators to enforce the contracts the platforms already have with their users. Under its Section 5 authority, the FTC could sue companies that defraud users by violating their own terms of service contracts. (The FTC has already sued Facebook for violating the privacy promises it makes in its terms of service.) The FTC could also use its rulemaking authority to define how platforms must spell out and enforce their terms of service.

Users are beginning to pursue these rights in the courts. Recently the Ninth U.S. Circuit Court of Appeals accepted an argument that YOLO, a Snapchat-integrated app (since banned on the platform) that let users send anonymous messages, misrepresented its terms of service. The panel "held that the claims [of the plaintiff, the family of a teen boy that committed suicide after threats and harassment on Snapchat] survived because plaintiffs seek to hold YOLO accountable for its promise to unmask or ban users who violated the terms of service, and not for a failure to take certain moderation actions" (which would have been protected by Section 230). (Conversely, the panel rejected the plaintiffs' argument that YOLO's anonymous messaging capability was inherently dangerous, under a product liability theory we discuss in Part III: Safeguarding Users).

Although Public Knowledge is generally supportive of the consumer protection theory, we recognize it has some pitfalls. For example, users or government officials could try to hold a platform accountable because they disagree with how the platform has interpreted or applied its terms of service. However elaborate or detailed the platform's rules may be, a term like "hate speech" is subject to interpretation. Content moderation is inherently subjective, and the consumer protection theory could be misapplied. But at the same time, in our vision, user rights in regard to free expression and content moderation should extend *beyond* simply understanding what the platforms can do with users' content (and data), and expecting the platforms to explain and comply with their own rules. For example, we would favor rights for users to file individual appeals to the platforms to challenge their content moderation decisions.

Despite our emphasis on user rights, we do not believe that users have a *right* to publish on any particular private platform, nor do they have a *right* to be amplified algorithmically. (As Aza Raskin of the Center for Humane Technology first noted, "freedom of speech is not freedom of reach.") In fact, platforms have their own expressive rights that are reflected in the communities they create through content moderation. They have the legal capacity to determine what is and is not allowed on their feeds and establish guidelines for acceptable posts via their terms of service. Users who abuse platforms in defiance of their community standards should, of course, face consequences, including being cut off from the platform when appropriate. But they should also know *why* they are being cut off, and the right of due process is still required.

Part II: Empowering User Choice

Securing User Choice Through Antitrust Enforcement and Competition Policy

The best mechanism to ensure platforms respect users' rights and honor their contracts is through healthy marketplace competition. *In a world with healthy competition, platforms would only be able to optimize profits if their content moderation reflected the expressive and associational preferences of their users.* With an estimated audience of 5 billion people, the social media marketplace should boast lots and *lots* of providers competing for our attention. Yet, the structure of today's digital markets means Big Tech platforms face little to no competition. ***We believe the concentration of private power over public discourse allows platforms to optimize content moderation for profit instead of user preference, and is itself a threat to free speech.***

For these reasons, we need to anchor our next discussion of content moderation policy in an understanding of digital market structure.

Digital Markets Are Highly Conducive to Monopolization

The markets in which digital platforms operate (e.g., search, social media, e-commerce, and user-generated entertainment) are all highly concentrated, meaning a select few companies have a huge influence on how consumers create, connect, and communicate. Following years of serial acquisitions of both market competitors and disruptors, extreme digital platform consolidation has profoundly influenced the direction and dynamics of content moderation. For

example, Meta (formerly Facebook) purchased Instagram in 2012 and WhatsApp in 2014, ensuring billions of users stay entrenched in Meta's product ecosystem. Now nearly 4 billion people – half of the world's population – use at least one of the company's core products, and many content moderation policies have been harmonized across them. Meta's content moderation policies therefore have a far-reaching impact on how consumers receive and disseminate information online. The same dynamics – compounded by what has been determined to be illegal, specifically monopolistic business practices by Google and its parent company, Alphabet – are true in search. The characteristics of digital markets themselves – most notably high capital investment, the accretive effects of data, network effects, and a tendency toward tipping (an economic dynamic in which one player owns or dominates a market) – favor consolidation and monopolization. Social media networks are particularly "sticky," making it difficult for consumers to abandon their established networks or switch services because their social circles are entrenched in a specific platform.

We are seeing this play out in real time as we write this paper. After Elon Musk's acquisition of Twitter, he fired most of the platform's trust and safety team and reversed many content moderation policies, allowing hate speech and misinformation to flourish. But as anyone who has tried to switch from Twitter (now known as "X") to Bluesky or Mastodon knows, rebuilding your feeds and follower base can feel like an impossible feat. Meanwhile, Meta's Threads challenged X's dominance by allowing users to import their followers from Instagram, rapidly gaining 100 million users in just five days. It demonstrated that the ability to transfer one's social network is crucial for a platform's adoption and success. (It took the extraordinary disruption of the 2024 national elections to provide an inflection point and allow Bluesky to overtake Threads in terms of user count.)

A Briefer on Antitrust Law and Why It Matters for Free Expression and Content Moderation

Throughout U.S. history, antitrust laws have played a crucial role in ensuring healthy competition in key sectors such as railroads, oil production, and automotive manufacturing. In more recent decades, the telecommunications and technology sectors have become focal points of antitrust scrutiny. This includes the breakup of AT&T into seven regional companies in the 1980s. Later, in the late 1990s and early 2000s, the antitrust lawsuits against Microsoft centered on the company's alleged abuse of its monopoly in the PC operating system market. The main U.S. case, filed in 1998, focused on Microsoft bundling the Internet Explorer browser with its Windows operating system, which the government claimed stifled competition in the web browser market. After a tumultuous legal process, Microsoft settled in 2001, agreeing to share its application programming interfaces or APIs and allow computer manufacturers more freedom in pre-installing non-Microsoft software. Both the AT&T and Microsoft cases resulted in new waves of innovation and competition and allowed the entry of new players, some of which grew to be today's tech giants.

Fast forward two decades, and there is a renewed focus on antitrust enforcement against Big Tech. It has been a challenge, however, for antitrust enforcers to mitigate digital platform monopolization. Thanks to years of narrowing jurisprudence, antitrust regulators use the consumer welfare standard, which determines whether business conduct harms consumers in

the relevant market – and that has been interpreted to mean unreasonably high prices charged for services. Since many digital platform services are free to users, proving consumers are harmed by self-preferencing and other monopolistic behavior is much more complex. That is why the Chair of the Federal Trade Commission, Lina Khan, is advocating for a neo-Brandeisian framework, where the number of firms and size of the largest player are the most important factors in assessing anticompetitive behavior. In other words, a single dominant company is always a threat to competition, and the only true remedy is more competitors.

The decision against Google in the Department of Justice's search antitrust case, which found that Google violated antitrust laws by illegally maintaining its monopoly over search and search text advertising, could significantly enhance free expression by encouraging fairer competition in the search engine market. It is undeniable that Google made the vast and infinitely expanding web more navigable, presenting search engine results pages with the exact information you were seeking – for free! The issue here is *not* that Google has the most-used search engine in the world, facilitating nearly 95 percent of searches on smartphones. The issue is, as the court found, that ***Google's practice of using contracts and payments to make its search engine the default option for Android and Apple phones violates antitrust laws***. What's more, because of its dominant position, Google had less incentive to maintain high-quality search results, instead prioritizing advertisers and paid placements. It also used a variety of practices to keep users on its search engine results page instead of clicking through to online publishers for information. If Google faced real competition, it would be incentivized to present a high-quality search service or lose out to alternatives – like DuckDuckGo, which does not use targeted advertising. Increased competition in search text advertising would also give advertisers and publishers more choices, leading to a richer diversity of information and enhancing opportunities for free expression.

Public Knowledge strongly supports both the Federal Trade Commission and DOJ's antitrust enforcement efforts. This includes DOJ's final proposed remedy calling for both structural and behavioral mandates in its ad tech case against Google, and the 2020 FTC lawsuit against Meta alleging abuse of monopoly power and the illegal acquisition of Instagram and WhatsApp (which may also call for divestiture of these platforms). We also support the efforts of the FTC and the DOJ to curb Big Tech's consolidation by blocking new mergers and acquisitions and potentially breaking up existing industry giants.

Legislative Solutions To Tackle the Anticompetitive Digital Market

Adapting how regulators and courts tackle antitrust enforcement in Big Tech will take time, as litigation can drag on for years while technology platforms continue to rapidly innovate and evolve. Therefore, we also advocate for legislative solutions to proactively create competition and choice in digital markets as a means of furthering free expression.

One way to open up competition among digital platforms is through mandatory data portability and interoperability – that is, requiring platforms to facilitate the transfer and utilization of data and to allow communication across different systems or applications. Interoperability would reduce the impact of network effects and remove one of the highest barriers to entry for new

platforms. Data portability and interoperability will not be standard among digital platforms unless they are mandated by law and then enforced by the FTC. (Yes, in great irony to the free market absolutists, regulation is needed to have a competitive social media market.) Just as consumers can send and receive emails from Gmail to Outlook and call from one phone service provider to another, social network users should be able to send private messages and see public images from any platform and on whatever platform they like best – including smaller entrants and community-run services that can connect to the dominant platforms.

Some bipartisan bills intended to make current anticompetitive practices perpetuated by Big Tech gatekeepers illegal include the American Innovation and Choice Online Act (AICOA) and the Augmenting Compatibility and Competition by Enabling Service Switching Act (the ACCESS Act). We support AICOA because it prevents covered platforms from "self-preferencing" at the expense of competitors and prohibits those platforms from using non-public data to unfairly advantage their products, thus indirectly benefiting free expression. And we support the ACCESS Act because it promotes interoperability among large platforms without dictating their specific functionalities, thus enhancing competition while avoiding user isolation. Supporters of the ACCESS Act and interoperability broadly might also consider promoting the development of newer decentralized protocols, like NOSTR and Holochain, on which digital platforms might be built. Like HTTP and the World Wide Web in the 1980s and 1990s, new decentralized protocols may provide the technical capability that policies like the ACCESS Act envision, without specific technical mandates in statute.

In the meantime, we believe Congress should pass sector-specific legislation, like The Ending Platform Monopolies Act, which would give both the FTC and DOJ the ability to impose structural separations and line-of-business restrictions on a covered platform to restore competition to digital markets. Likewise, we believe that Congress should create a digital regulatory agency that would be able to craft rules of enforcement around these legislative mandates and develop a regulatory framework for the industry. We discuss this more in Part IV: Tackling AI and Executing the Vision.

Addressing the Broader Information Ecosystem

As they relate to information distribution and free expression, digital platform markets live within a broader ecosystem of news and information. That means we can also use policy to *offset the impact* of poor platform content moderation on the information environment by promoting other, diverse sources of credible news. The problem is, local news – at least in its traditional forms – is dying.

In the last 20 years, the number of local news outlets in the United States has shrunk by a third, down from 9,000 to around 6,000 today – and is still decreasing. Despite legacy news advocates' attempts to pin this decline completely on Google and Facebook, the reality is more complex and the problem started well before those two platforms rose to scale.

Newspapers have decried the entry of new technology players for decades. For example, the

Newspaper Preservation Act of 1970 authorized the formation of joint operating agreements among competing newspaper operations within the same media market area, in part to address publishers' claims that they needed to be able to compete more effectively for advertising revenue with radio and television. The news crisis intensified in the 1990s as consumers started migrating to the internet, which spawned new information channels that catered (and advertised) to them, ending newspapers' own monopoly on local retail ads and highly profitable classified ads. Craigslist, a classified advertisements website that went live on the internet in 1996, typified the internet's threat to the news industry. By 2013, the Wall Street Journal noted that "Craigslist obliterated the longtime business model of local journalism that relied on classified-ad revenues, which have fallen by 80 percent." Newspapers themselves also sought to capitalize on the revenue growth opportunities in the online classified advertising space; for example, Classified Ventures, LLC was a joint venture among six major newspaper publishers specifically created to start Cars.com and Apartments.com. But such efforts weren't sufficient to offset the declines in their core ad business.

Meanwhile, in their race to reach online viewers, many news organizations offered their core offering – news – for free, perhaps forever changing Americans' attitudes about the value of news. Now, most Americans get at least some of their news from online sources.

In another misplaced bid to "save" the industry, many newspapers ramped up consolidation, hoping to better compete against the internet, offset declining revenues, and reduce costs. This trend accelerated as private equity and hedge funds gobbled up newspapers, pursuing profit maximization under the guise of "synergy." The result: smaller newspapers were pushed out, newsrooms were decimated, and many reporting jobs were reinvented as "freelance" positions offering a fraction of their previous salaries and benefits. Now, with over half of daily newspapers owned by hedge-fund-run news conglomerates, the quality and quantity of local news reporting is tanking. When independent local journalism declines, so does trust in institutions and democratic participation.

The downward spiral in local news – especially newspapers, which still create the majority of original reporting – has spurred a number of policy solutions. Many of these focus on extracting revenue back from Google and Facebook, which do dominate the online ad market today. In the last decade, we have seen countries like Spain, Australia, and Canada use different legal theories to try to implement "link taxes," which require large platforms to pay a fee to news publishers for posting links to articles on news sites. In every case, the platforms have responded by removing links to news – or threatening to do so – to the enormous detriment of news organizations. The proposed U.S. version of a link tax, the Journalism Competition and Preservation Act, creates an antitrust exemption for publishers and would require that digital platforms pay for and carry content from any qualifying journalism provider.

In our view, the JCPA would threaten free and open access to information; undermine content moderation and increase harmful material online; infringe on platforms' own First Amendment rights; and risk expanding copyright law beyond traditional bounds. The bill also rewards the giant consolidating media corporations that helped create the problem in the first place; sidelines smaller news outlets; and does nothing to ensure funds raised support actual

journalists directly. Additionally, by matching tech might with media might, the bill will harm competition and entrench existing power structures rather than foster a healthier journalism ecosystem critical to supporting local news.

Whatever their legal theory or structure, link tax proposals are rooted in an inappropriate model: the idea that the current platforms unjustly enrich their own bottom lines by capturing advertising that belongs to news organizations. *Instead, we favor a public interest obligations approach to news.* As Public Knowledge has explained, "the public interest obligations approach does not turn on whether or not dominant platforms engage in bad behavior or receive an unfair benefit. Rather, the public interest obligations approach recognizes that we have a classic market failure." This approach still allows policymakers to place obligations on those most capable within the current market (dominant platforms) to correct the market failure by supporting a valuable public good (local news production). For example, our Superfund for the Internet policy proposal establishes a federal "trust fund" administered by an independent body, with contributions from applicable large online platforms in the form of a federal user fee calculated based on the platform's number of active monthly users. The funds would then be allocated to qualified news organizations that would apply strong fact-checking principles to platform content, ensuring social media platforms that distribute such news content are contributing to a healthy information ecosystem. A federal tax on platforms for the purpose of funding news would work in a similar way.

We also support solutions that spur public funding mechanisms and favor new journalism business models. This includes legislative initiatives like the Community News and Small Business Support Act, which aims to provide small business tax credits to help pay for advertising in local news outlets. It also gives payroll tax credits to local newspapers for employing local journalists.

Antitrust enforcement and competition policy may also have a role in the news industry itself. Twenty years ago, as we've noted, a flurry of mergers and acquisitions of small newspapers made it so the 25 largest news conglomerates owned 70 percent of America's daily newspapers. This consolidation in the news industry, especially by financially motivated owners (some of whom have come under criticism of late for disallowing presidential endorsements by their editorial boards), has caused documented harm to citizens and workers as well as to media localism and diversity. At a time when there is already an effort to expand the application of antitrust law beyond pricing harm and the consumer welfare standard, we can apply or adapt it to this critical democratic need.

There is ample precedent for the government to use policy to further diversity and locality of views. At the start of 2024, the Federal Communications Commission shared its efforts to bolster local news by giving preference to broadcasters that commit to local content in their license applications and also enhance the availability of region-specific news. This is on top of rules to protect local news and diversity of voices, like limiting the number of broadcast licenses (radio or television) an entity could control in the local market; prohibiting newspaper/broadcast cross-ownership in the same market; and prohibiting cross-ownership of television licenses and cable systems in the same market. Similar pro-competitive rules could extend to localism in the

news industry, such as amending the plant-closing laws to require that any chain planning to close a newspaper must give the community 90 days' notice so they might organize a bid to buy the paper. Similarly, any acquisition by hedge funds should be paused by the government to allow for alternative bid offers from local businesses or nonprofit organizations. Some of these ideas are already being pursued at the state level.

Beyond preventing further news industry consolidation, Public Knowledge supports efforts to de-consolidate, or "replant" newspapers back into communities, with the government providing financial incentives or tax benefits to local nonprofit organizations or mission-oriented businesses that buy newspapers, or hedge funds that sell them. Incentives may include loan guarantees for local organizations that acquire a newspaper, or providing payroll tax credits. Going further, we can incentivize owners of news conglomerates to sell off a newspaper to a local nonprofit by eliminating capital gains taxes in the transaction.

Part III: Safeguarding Users

Designing Policy Interventions to Safeguard Users from Harm

Before diving in, It's important to note that *an enormous amount of the focus on content moderation – and platform accountability more broadly – is in the interest of preventing harm.* At Public Knowledge, we categorize the potential harms of algorithmic curation of content into the broad categories of (1) harms to safety and well-being (including privacy, dignity, and autonomy); (2) harms to economic justice (including access and opportunity); and (3) harms to democratic participation (including through misinformation). These harms can arise from obvious issues like cyberbullying and non-consensual intimate imagery (NCII), from targeting of content that reflects and amplifies bias, and from purposeful narratives of disinformation.

While academics, health professionals, social science researchers, advocates, policymakers, and industry leaders continue to debate the actual causality between platform content moderation and user harm, there is clear and growing momentum pushing platforms to be safer for all users. With this understanding, we propose policy approaches that aim to balance user safety with the preservation of free expression, focusing on product liability, comprehensive privacy protection, and requirements for algorithmic transparency.

The Product Liability Theory: Revisiting "Big Tech's Tobacco Moment"

A theory gaining momentum among some policymakers and civil society groups like Public Knowledge is that platforms' design features – *separate and distinct from the nature of the content they serve to users, or how they serve it* – can create harms, and that platforms should be liable for the harms their design features cause. Generally, product liability can take the form of claims regarding manufacturing defects, defective design, or failure to provide instructions or warnings about proper use of a product. In the case of platforms, most of the discussion about product liability refers to *product design* that increases time spent on the service, triggers compulsive use, motivates unhealthy or chronic behaviors, overrides self-control, creates social

isolation (all of which can negatively affect self-image and mental health), and introduces unsafe connections to users. This theory holds that, as in other industries, platforms should be accountable – that is, legally liable – for the harms that are caused by the design of their products. (The other industry most often referenced under this theory, by far, is tobacco. Public Knowledge covered "Big Tech's Tobacco Moment" in 2021.) This goes farther than the consumer protection theory we described in Part I: Under the product liability theory, not only must a platform *not* be deceptive or unfair, but it must also take affirmative action for its products to be safe, and to warn users if they are not.

It's also important to note that policy proposals and lawsuits advanced under this theory, by definition, are *not* explicitly about free expression and content moderation (even though they *are* designed to address some of the same harms). Specifically, *we do not include algorithmic serving or amplification of user content in our definition of product design under this theory.* In fact, we have previously noted "...our continued belief that models that call for direct regulation of algorithmic amplification – whether it's the speech itself or the algorithms that distribute it – simply wouldn't work or would lead to bad results." One way to "test" claims or proposals advanced under this theory is to ask whether the claim or proposal requires knowledge of, or reference to, specific pieces or types of harmful content. If it does, then the claim or proposal really refers to content liability and is likely barred by both Section 230 and the First Amendment. Plaintiffs may seek to work around these prohibitions by characterizing their theory of liability in different terms (like claiming that "recommendations" are manifestations of the platform's own conduct). However, *any theory of liability that depends on the harmful contents of third-party material constitutes treating a provider as a publisher and is barred by Section 230.*

All that said, proposals advanced under the product liability theory as we define it are showing promise as a way to create platform accountability for harms without the constitutional or legal barriers associated with direct regulation of content. In this section we talk about the origins of this theory, its current focus in federal policy, and alternative paths to apply it.

Origins of the Product Liability Theory

Over the past few years, thanks to researchers, journalists, and whistleblowers, we've all become more aware of the externalities of the platforms' advertising-based business model. That model – which drives the vast majority of the revenue of virtually all the major search, social media, and user-generated entertainment platforms – incents the platforms to design product features that maximize the time and energy people put into searching, scrolling, liking, commenting, and viewing. It's simple: users' *attention*, focused via algorithmic targeting on content that is most likely to be relevant, is those platforms' only inventory. It's what they sell (to advertisers). So they design features into their products to create more of it.

The platforms call this time and energy – this attention – "engagement." It's a deliberately upbeat and positive-sounding word that platforms adopted from the traditional ad industry to describe the time users spend scrolling, viewing, liking, sharing, or commenting on other people's posts. It's catnip to advertisers, who assume their ads will benefit from it, too. But that same "engagement" – and platforms' efforts to increase it – has been associated with

compulsive use, unhealthy behaviors, social isolation, and unsafe connections. As awareness of the ad-based business model and its potential for harm grew, policymakers brought forward proposals to understand, and then regulate, the role of product design in the harms of social media. One early proposal Public Knowledge favored, the Nudging Users to Drive Good Experiences on Social Media Act (the "Social Media NUDGE Act") called for government agencies to study the health impacts of social media; identify research-based, content-agnostic interventions to combat those impacts; and determine how to regulate their adoption by social media platforms. If it had passed, we might be having more evidence-based policy discussions today.

However, an enormous amount of the subsequent focus in legal and policy circles shifted to how platforms secure more time and attention to sell to advertisers by algorithmically targeting and amplifying provocative content. As renowned tech journalist Kara Swisher puts it, "enragement equals engagement," meaning the content that elicits the most attention tends to be the most inflammatory. But court case after court case claiming harms from algorithmic distribution of content has been dismissed or lost on one of two grounds. The first is the Section 230 liability shield, which insulates platforms from liability for user content or how it is moderated. The second is the First Amendment, which gives platforms their own expressive rights in content moderation. Similarly, bill after bill in Congress focused specifically on influencing platform policies and practices regarding content moderation have been rejected – as they should be – as contradictory to Section 230 or to the platforms' own expressive rights under the First Amendment.

Obviously, some of the ill effects of social media are due to actual third-party content as well as the amplification of this content to users who didn't ask for it. This includes harassment, hate speech, extremism, disinformation, and calls for real-world violence. But *our product liability theory holds that some harms can be caused by product design features that are rooted in the platforms' ad-based business model and the need to sustain user attention*. So rather than pointing to content or content curation, the product liability theory implicates product liability law, holding platforms liable for the harm they cause as the designer, manufacturer, marketer, or seller of a harmful *product,* not as the publisher or speaker of information.

Current Focus of the Product Liability Theory

In Congress and the courts, the current focus of the product liability theory is the well-being of children and adolescents, whose attention generates an estimated $11 billion a year in digital ad revenue in the U.S. Some of this focus is rooted in research and whistleblower revelations about the impact of social media usage on kids – and what the platforms know about it. For example, early in 2023, the U.S. Surgeon General issued an advisory noting that "social media can… pose a risk of harm to the mental health and well-being of children and adolescents," since adolescence is a particularly vulnerable period of brain development, social pressure, and peer comparison. More recently, the Surgeon General has called for warnings akin to those on cigarette packages, designed to increase awareness of the risks of social media use for teens. A best-selling book puts forward a case that a "phone-based childhood," combined with a decrease in independent play, has contributed to an epidemic of teen mental illness. The

question of social media's impact on kids gained more steam recently from another round of revelations about "what Facebook knew and when they knew it… and what they didn't do about it" in regard to child safety. Policy proposals rooted in the product liability theory enjoy support from, and have sometimes been shaped by, youth advocacy organizations, including DesignItForUs and GoodForMEdia.

One major challenge to all of this: While the crisis in youth mental health is very real, research on the causality of social media is mixed at best. That said, products and services may cause harm even if they don't create a health crisis. For example, a literature review from the National Academies of Science, Engineering, and Medicine recently concluded that social media may *not* cause changes in adolescent health at the population level, but may still "encourage harmful comparisons; take the place of sleep, exercise, studying, or social activities; disturb adolescents' ability to sustain attention and suppress distraction during a particularly vulnerable biological stage; and can lead, in some cases, to dysfunctional behavior."

Policymakers also focus the product liability theory on kids because of the higher likelihood of bipartisan agreement, after years of failures to regulate Big Tech. Hauling Big Tech CEOs into hearings and demanding apologies for the genuinely heartbreaking losses families have experienced as their children confronted mental health crises or physical harm makes for viral moments. The risk is legislation propelled by "moral panic and for-the-children rhetoric" rather than sound evidence of efficacy relative to youth mental health (which most experts agree requires a more multidimensional approach).

In our view, the product liability theory may in fact be an effective way to mitigate some of the harms associated with social media while circumventing both constitutional challenges and the intermediary liability protections provided by Section 230. *But it should apply to all users of social media* – not just kids and teens.

There are three ways to apply the product liability theory to mitigate harms from product design: through litigation, through reform of Section 230, and through new legislation.

Litigation

In the past, courts have not generally distinguished between the role of product design features and the role of content distributed by algorithms in creating harm. The platform defendants didn't help: They generally argued they were exempt from liability due to their own expressive rights or the broad protection provided by the liability shield of Section 230. As a result, algorithms and an expanding array of other platform product features have been determined by judges to be the platforms' own protected speech and/or shielded by Section 230. Most cases have been dismissed quickly on the premise that Section 230 bars claims based on alleged design defects if the plaintiffs seek to impose a duty to monitor, alter, or prevent the publication of third-party content.

However, a set of more current cases makes finer distinctions between product design features and the algorithmic curation of user content – though they don't always agree where the line is.

In 2021, in *Lemmon v. Snap*, judges determined that it was one of Snapchat's own product features – a speed filter – and not user content that had created harm. Since then, almost 200 cases have been filed alleging product defects or similar claims, and some are making it past moves for dismissal on the basis of Section 230 and/or the platforms' own expressive rights. For example, in what is now a multidistrict product liability litigation against Facebook, Instagram, Snap, TikTok, and YouTube, a judge determined that some product design choices of the platform (like ineffective parental controls, ineffective parental notifications, barriers that make it more difficult for users to delete and/or deactivate their accounts than to create them, and filters that allow users to manipulate their appearance) *neither* represent protected expressive speech by the platforms (so the First Amendment does not protect them), *nor* are they "equivalent to speaking or publishing" (so they are not shielded by Section 230). A district court judge in Utah found that Section 230 does not preempt a state law's prohibitions on the use of autoplay, seamless pagination, and notifications on minors' accounts. More recently, courts have split on whether TikTok's recommendation algorithm is the platform's own "expressive activity" or whether it is liable in the tragic death of a 10-year-old girl who participated in the "blackout challenge" found on the platform. (Public Knowledge joined an amicus brief in this case, arguing that platforms may have both Section 230 immunity and First Amendment protections for their editorial decisions, including algorithmic recommendations.) The Superior Court of the District of Columbia, in its civil division, denied Meta's motion to dismiss a case claiming that personalization algorithms that leverage a variable reward schedule, alerts, infinite scroll, ephemeral content, and reels in an infinite stream foster compulsive and obsessive use because "the claims in the case are not based on any particular third-party content." For this reason, the court "respectfully decline[d] to follow the decision of the judge in the multidistrict litigation." In October of 2024, the Attorney General of New Mexico released new details of the state's lawsuit against Snapchat, which claims the company fails to implement verifiable age-verification, designs features that connect minors with adults, and fails to warn users of the risks of its platform.

The outcome of all these ongoing cases is obviously dependent on many variables, but they may signal it is possible to distinguish design features from content or algorithmic curation and whether a litigation path for the product liability theory is viable.

Section 230 Reform

An alternative way to apply the product liability theory would be targeted reform of Section 230 designed to clarify which aspects of a platform's own conduct or product design lie outside the protections of the intermediary liability shield. In our view, conduct apart from hosting and moderating content is already outside the scope of Section 230. But while past court cases (like Homeaway and Roommates) have demonstrated this for specific activities and fact patterns, the sheer number of current court cases (and the sometimes-conflicting decisions arising from them) show it may be difficult to draw the line. Such reform would not create liability for any elements of product design, but it would allow the case to be made in court.

Public Knowledge has proposed Section 230 principles to protect free expression online. Targeted reform of 230 to advance the product liability theory could be accomplished while

adhering to those principles. One principle states that Section 230 *already* does not shield business activities from sensible business regulation. Another principle is that Section 230 was designed to protect *user* speech, not advertising-based business models (which most of these product design features are meant to advance). A third principle states that Section 230 reform should focus on the platform's own conduct, not user content. (As a result of these principles, Public Knowledge also proposes that users should be able to hold platforms accountable for taking money to run deceptive or harmful ads because paid ads represent the business relationship between the platform and an advertiser, not users' free expression.) As noted, great care would have to be taken to distinguish between the platforms' business conduct and its own expressive speech (as well as speech of users), but if done well, this would be a content-neutral approach to Section 230 reform that would withstand First Amendment scrutiny.

New Legislation

As noted, most of the focus for the product liability theory by policymakers has been specifically about the safety of kids and teens. The federal proposal that has gained the most bicameral, bipartisan traction under this theory is the Kids Online Safety Act, or KOSA, and it personifies both the mechanisms and risks that accompany child-focused legislation. The bill requires that platforms exercise a "duty of care" when creating or implementing any product design feature that might exacerbate harms like anxiety, depression, eating disorders, substance use disorders, and suicidal behaviors. Platforms must limit design features that increase the amount of time that minors spend on a platform. And the bill requires platforms to make the highest privacy and safety settings the default setting for minors, while allowing them to limit or opt out of features like personalized recommendations.

Criticism of KOSA – and bills similar to it at both the federal and state level – largely centers on the belief that any requirement that platforms treat minors differently from adult users will inevitably lead to age-gating. Age-gating refers to the use of electronic protection measures to either verify or estimate users' ages in order to restrict access to applications, content, or features to those of a legal (or deemed-appropriate) age. Both age verification and age estimation have hazards, including technical limitations and the risk of bias and privacy risks. Age verification requirements have also been deemed unconstitutional at the state level as they impinge on all users' rights to access information and remain anonymous.

There are also concerns that any duty of care applied to content platforms – no matter how specifically or narrowly defined – will inevitably lead to content restrictions, particularly for marginalized groups, as interested parties deem content they find objectionable to be "unsafe." Court precedent so far would disallow such demands by plaintiffs under the First Amendment and the protections of Section 230, though that would not prevent platforms from removing certain categories of content themselves in order to avoid the legal risk. These concerns arise in part because the overall concept of a duty of care can be amorphous. Common law duties including a duty of care in other contexts have evolved over centuries, may require expert testimony to verify, and are subject to different interpretations by juries drawn from communities with differing values.

The product liability theory, which we support, has some similarities to frameworks calling for "safety by design." Combined with a national privacy standard, which we discuss below, such legislation would help users avoid the harms associated with certain product features without impacting users' or platforms' expressive rights. *But this is another case where Public Knowledge would prefer to see protections for all users, not just kids and teens.* Rather than run headlong into the buzzsaw of opposition to age-gating, policymakers could articulate content-neutral regulations that govern product features related to the platforms' advertising-driven business model. This would also remove the ambiguity about what material is "suitable" or "safe" for minors based on its subject matter or point of view. Such regulations may prohibit certain product design features (for example, dark patterns meant to manipulate user choices). They may include requirements for enhanced user control over their experience (for example, requiring that safety and privacy settings are at their highest possible setting by default). And/or, they may require a more focused "duty of care," or duty to exercise reasonable care in the creation and implementation of any product feature designed to encourage or increase the frequency, time spent, or activity of users. The regulations may also require platforms to study the impact of their product design, make data available to researchers for such studies, and make any findings available for audits or transparency reports.

Limiting Data Collection and Exploitation Through Privacy Law

Remember the 2018 Cambridge Analytica scandal? As a reminder, the British political consulting firm acquired personal data from millions of users for targeted political advertisements in 2016. A personality quiz application on Facebook, created by a psychology professor and funded by Cambridge Analytica, was used to collect user data and data from users' friends without their consent in order to run targeted political digital advertising campaigns. Although only 270,000 users consented to have their data harvested, Cambridge Analytica obtained data from around *30 million users* connected to those initial participants. While its actual impact on the Brexit vote has been shown to be minimal, the scandal created wide awareness of platforms' data practices.

Technically, the personality quiz app's transfer of user data to Cambridge Analytica violated Facebook's terms of service. However, from a legal standpoint, the acquisition, sale, and sharing of personal data by platforms or data brokers without individuals' knowledge is often permissible. After all, *there is no single, comprehensive federal privacy law that governs the handling of personal data in the United States*, so data collected online or through digital products has little regulatory oversight. The concentration of dominant platforms means a handful of giants control vast amounts of data, which may be used for privacy-invasive activity, like behaviorally targeted advertising and profiling of users for targeting of content. This can compound harms, especially to marginalized communities that are often the target of hate speech and harassment.

At Public Knowledge, we advocate for protecting consumer privacy through requirements for data minimization, informed consent, and effective user controls. We advocate for a thorough federal privacy law that provides a foundation for states to build upon and includes a private

right of action, enabling consumers to take legal action when necessary.

State of Play in Privacy Law

Companies have, time and time again, been exposed for sharing sensitive personal data without user consent. The Federal Trade Commission plays consumer protection Whac-A-Mole by slapping fines on privacy-violating companies – like the $7.9 million fine on Betterhelp, the online therapy company, which sold customers' health data to Facebook and Snapchat. Regulatory enforcement can punish bad actors, but does nothing to mitigate the privacy-invasive behavior – or the harms it may cause – in the first place. That's where comprehensive national policymaking comes in.

The U.S. has tried – in vain – to pass a federal privacy bill. Since 2021, Public Knowledge has supported – with some caveats – the Online Privacy Act and the American Data Privacy and Protection Act (ADPPA). The latter bill aimed to prevent discriminatory use of personal data, to require algorithmic bias testing, and to carefully restrict the preemption of state privacy laws, among other benefits. Most recently, in 2024, the American Privacy Rights Act (APRA) succeeded ADPPA, but Public Knowledge – and some Democratic lawmakers – came to oppose it due to the removal of key civil rights protections.

While we have expressed support for a variety of privacy-related bills, we believe that truly effective privacy protections require addressing the entire online data ecosystem, not just targeted measures. One-off actions can have minimal real-world impact, especially if aimed at a specific company or practice (looking at you, Tiktok ban). Worse, they may reduce avenues for free expression online while allowing Congress to neglect the need for comprehensive privacy protections across all communities.

The Misguided Focus On Kids' Privacy

While any attempt at comprehensive privacy legislation withers in Congress, more focused battles on child privacy pervade, for the same reasons we noted in regard to the product liability theory. The great impasse in the child privacy debate is that – you guessed it – bills tend to mandate data minimization while also proposing age verification mechanisms that would require additional collection of personal data. These proposals have taken various forms, including Section 230 carve-outs, which would make platforms liable for child privacy-invasive behavior.

The Children's Online Privacy Protection Act (COPPA), enacted nearly 25 years ago, is the original law safeguarding child users (in this case, those under 13) from websites collecting personal information without consent. COPPA re-emerged in the last couple of years as policymakers sought to update the framework to better reflect the evolution of social media. Known as COPPA 2.0, the revised bill increases the covered age to 17 and requires platforms to comply using an implied knowledge standard that a particular user is a minor. Public Knowledge supported the new COPPA framework, but not without critiques. The biggest was that – no surprise – we believe that any privacy law should be applicable to *all* users, not just kids.

There is also a slew of bills that specifically target the awful proliferation of child sexual abuse material (CSAM) online. Unfortunately, most of these proposed laws also miss the mark. Notably, the Eliminating Abusive and Rampant Neglect of Interactive Technologies (EARN IT) Act, floating around Congress since 2020, repeals Section 230 for platforms that do not act sufficiently on CSAM, exposing platforms to criminal and civil liability for its distribution and presentation. We've steadfastly opposed EARN IT, not only because repealing Section 230 would have such detrimental effects on free expression, but also because the bill will eliminate or discourage encryption services and force platforms to expand broad content moderation, which disproportionately impacts marginalized communities. We think that users, such as journalists messaging sensitive sources, have the right to communicate free from surveillance from third parties by leveraging end-to-end encrypted messaging.

Similarly, the Strengthening Transparency and Obligation to Protect Children Suffering from Abuse and Mistreatment (STOP CSAM) Act falls short, compromising privacy by discouraging end-to-end encryption. One very important note relevant to both EARN IT and STOP CSAM: Section 230 *already* has an exception for federal criminal activity, which includes the distribution of CSAM. If we want to curb child exploitation, increased surveillance of *everyone* is not the solution. Enforcing existing laws and putting resources towards victim identification and assistance would be a more productive, rights-preserving approach.

Requiring Algorithmic Transparency

The lifeblood of a digital platform is not the user, the interface, or the posts – it is the complex math equations used to organize content on your feed or in your search results, called algorithms. Digital platforms utilize machine-learning algorithms to tailor content feeds, aiming to maximize user engagement – and as we've noted, platform profits. These algorithms analyze personal data such as viewing habits, geographic location, platform history, and social connections to prioritize content users will likely engage with. Algorithms are tools created by humans to perform specific functions. They not only arrange and rank content in feeds but also enforce platform content guidelines by identifying and removing inappropriate content in an automated way (ideally in conjunction with human moderators who can understand and apply cultural and context cues). Yet, while algorithms can enhance personalized user experiences, they can also *amplify* harmful or discriminatory content.

Early social media platforms displayed content reverse-chronologically, but this approach quickly became inadequate as information volume and investor demands for monetization grew. For example, recognizing users' struggle to navigate the flood of content, Facebook introduced EdgeRank in 2007. It was one of the first sophisticated social media algorithms, and it drove both user engagement and profit optimization for Facebook's also-new ad-based business model. The EdgeRank algorithm prioritized content based on three key factors: the frequency of user interactions with friends; the types of content a user typically engaged with; and the recency of posts. This system aimed to present users with a more personalized and engaging feed, effectively filtering out less relevant content and highlighting posts deemed more likely to interest each individual user. Facebook has since fine-tuned its algorithm, now integrating tens

of thousands of variables that better predict what users would like to see on their feeds and what will keep their eyes on the platform for as long as possible. Today, each and every social media platform utilizes its own proprietary algorithm to attract and keep users glued to their feeds.

Algorithmic ranking of content, particularly in combination with design features such as endless scroll and recommendations, can exacerbate harm in several ways. It can expose users to increasingly extreme content, send them down subject matter rabbit holes, and narrow the range of views and voices they see. Effective content moderation requires an understanding of context and cultural nuances, whereas algorithms typically rely on specific terms or hashes, which may not capture the full meaning or intent behind the content. As we've described, algorithmically mediated enhancement of exposure and engagement with divisive, extreme, or disturbing content can have real-world impacts in terms of public civility, health, and safety. Algorithmic ranking can give rocket fuel to toxic online user behaviors like targeted cyberbullying, verbal abuse, stalking, humiliation, threats, harassment, doxing, and nonconsensual distribution of intimate images.

Adding to this complexity, companies frequently adjust moderation policies and practices in response to current events or political pressures. Moderation algorithms can also reflect the cultural biases of those who coded them: predominantly male, libertarian, Caucasian or Asian coders in Silicon Valley. While we can observe the effects of these algorithms, their inner workings remain largely obscure, often referred to as "black boxes." Nevertheless, malicious actors can exploit these algorithmic vulnerabilities to optimize harmful content without needing to understand the underlying code by playing on current events, political pressures, or predictable biases. The opacity, potential for biased data and feedback loops, lack of oversight, and scale of algorithms compound their impact compared to human moderation. These harms disproportionately affect historically marginalized groups, as platforms sometimes disregard or suppress research indicating discriminatory content moderation practices, leaving allegations of racism, sexism, and homophobia from users largely unaddressed.

Since the revelations of whistleblowers like Frances Haugen of Facebook in 2021, which showed that platforms knowingly amplify harmful content, policymakers have been interested in holding platforms accountable for algorithm-related harms. Various algorithmic transparency bills have been proposed in Congress, aiming to shed light on the mechanisms driving social media algorithms, potentially enabling researchers and regulatory bodies to monitor and, when necessary, intervene in their operation.

As part of Public Knowledge's broader advocacy for free expression and content moderation, we recognize that algorithms are crucial to platform operations but are currently too opaque. Rather than advocating for restrictions on algorithm use, Public Knowledge supports legislation that mandates transparency in algorithms and ensures users have a clear understanding of how platform content moderation decisions are made. Some notable examples are the decisions by X to downrank posts with links (to keep people on the platform) and by Meta to downrank news by default (ostensibly to respond to users wanting less "political" content in their feeds). Such

decisions warrant more transparency and choice for users given their impact on the availability of news and information.

Legal Challenges to Algorithm Use and Impact

In 2023, two court cases raised the question of whether social media companies are liable for contributing to harm to users under the Anti-Terrorism Act (ATA) by hosting and/or algorithmically promoting terrorist content.

In *Twitter v. Taamneh*, the Supreme Court considered whether a platform that hosted ISIS-related content could be liable under the ATA, which prohibits "knowingly providing substantial assistance" to designated terrorist groups, and whether Section 230 should shield it from liability. But the court found that a social media company that merely hosted such content (because it was open to anyone to create accounts and post material) did not meet the ATA's knowledge threshold. Because under the facts of the case, Twitter could not have been found liable, the Court did not need to decide whether Section 230 would have shielded Twitter from liability under the Act.

In *Gonzalez v. Google*, plaintiffs similarly argued that Google should be liable for algorithmically promoting terrorist-related content on its YouTube platform. The Biden administration filed a brief in this case, arguing that Section 230 did not shield platforms from liability for algorithmic content recommendations. (Public Knowledge filed a brief disagreeing with this claim.) However, given the result in the Taamneh case, Google could not have been found liable, whether or not Section 230 applied. The Court therefore did not issue a decision clarifying the scope of Section 230.

The Court may not be able to avoid ruling on Section 230 in future cases, but both *Taamneh* and *Gonzalez* demonstrate that, even without Section 230, holding a platform liable for harms stemming from content they host or recommend is difficult. Specific legal claims such as those under the ATA generally have high thresholds of culpability, such as requiring a platform to *deliberately* promote harmful material, instead of such material being swept up by a general-purpose algorithm. Further, the First Amendment largely protects platforms (and their users) from liability even for promoting false, or even dangerous material, without a showing of knowledge and culpable conduct, or the presence of a specific duty of care (such as that of doctors to their patients). While Section 230 does shield platforms from liability in some cases, it largely cuts short litigation that has little chance of success to begin with.

The Right Policy Framework Can Make Algorithms Both Helpful and Healthy

Proposed policy frameworks to regulate algorithmic decision-making range from banning the use of algorithms *entirely*, to holding platforms liable for algorithmically amplified content, to requiring transparency, choice, and due process in algorithmic content moderation. Banning algorithms outright is a narrow and impractical solution, given their dual role in promoting content and enforcing platform guidelines. As we've noted, Section 230 and the First Amendment preclude blanket liability for algorithmic curation and broad liability would result in

platform over-moderation fuelled by risk aversion. Instead, Public Knowledge believes the best solution is to require transparency into platforms' algorithmic design and outcomes as a component of better and more evidence-based regulation and informed consumer choice. It also provides the means to create accountability for platforms' enforcement of their own policies, as we recommended earlier. It is the role of Congress to pass legislation that empowers users and to address aspects of social media platforms' business models that can perpetuate harm.

For example, Public Knowledge supports the bipartisan Internet Platform Accountability and Consumer Transparency Act (Internet PACT Act), which would require social media companies to publish their content rules, provide transparency reports, and implement a user complaints mechanism. This approach ensures platforms adhere to their own rules while providing users with clear guidelines and appeal processes. It aims to enhance transparency, predictability, and accountability, with enforcement by the FTC.

Another bill we support is the Platform Accountability and Transparency Act (PATA), reintroduced in 2023. It aims to improve platform transparency and oversight by creating a National Science Foundation program for researcher access to data, setting FTC privacy and security protocols, and mandating public disclosure of advertising and viral content. The updated version of the bill no longer seeks to revoke Section 230 protections from non-compliant platforms, a change welcomed by Public Knowledge.

Despite bipartisan support, neither of these bills has advanced to a vote.

Part IV: Tackling AI and Executing the Vision

Tackling AI-generated Content

Journalists, researchers, and policymakers are hand-wringing about the potential of generative artificial intelligence (GAI) to further erode trust in information institutions. The worry proved especially salient in 2024, a big election year globally, in which the new availability and sophistication of GAI tools allow bad actors to proliferate deepfaked imagery and disinformation at an exponentially fast rate. Although generative AI, thus far, does not seem to be the source of entirely *new* disinformation narratives, by virtue of its scale and speed, it still boasts the potential to increase the vulnerability of platform users to polarization, manipulation, health risks, market instability, and misrepresentation. We are already seeing how AI-enabled deepfakes and misinformation worsen distrust in the government and may threaten our democratic institutions. Generative AI has also heightened the "liar's dividend," wherein politicians may induce informational uncertainty or encourage oppositional rallying of their supporters by claiming true events are the manifestation of AI.

Platforms are using a variety of methods to identify and moderate AI-generated content. For example, Meta decided it would handle AI-manipulated media by adding "AI info" labels to content on its social media platforms (as well as integrating invisible watermarking and

metadata to help other platforms identify content generated by Meta AI). Similarly, the social publishing platform, Medium, requires any writing created with AI assistance to be clearly labeled. Other platforms' approaches are simply extensions of their existing strategies to mitigate disinformation.

Many policymakers are looking to the proliferation of AI-generated content as a cause for increased content moderation scrutiny and platform regulation. However, some of the harms this content may create are better addressed elsewhere in the ecosystem. For example, there may be regulation that introduces liability for the AI developers or deployers themselves (rather than the platforms that simply distribute such content). (Note that some liability already exists because Section 230, generally speaking, does not and should not protect generative AI). Existing laws can be clarified to ensure the underlying acts (like distribution of child sexual abuse material, or CSAM) are illegal if they are conducted using AI. Or they can be made illegal at the federal level if they are not now (like distribution of synthetic non-consensual intimate imagery, or NCII), which, among other things, would change the platforms' incentives by placing this content outside the liability protections of Section 230.

Researchers and policymakers have also focused on requirements to track "digital provenance" and ensure "content authenticity" from AI developers to distribution platforms. While this is a promising area, these methodologies remain imperfect and least likely to be adopted or retained by bad actors. And some of these methods raise concerns that they may encourage platforms to detect and moderate certain forms of content too aggressively, threatening free expression. This, too, has the potential to damage our democracy and will likely disproportionately impact marginalized communities.

Policy Parameters for Moderation of Synthetic Content

Public Knowledge has detailed the risks associated with GAI-generated digital replicas, and some of the policy guidelines we advocate for apply here, as well. For example, we advocate for narrow, commonsense protections for our elections, leveraging well-established legal doctrines for how to require disclosures in political advertising, crack down on fraud, and protect the integrity of the electoral process. We urge caution on the potential for over-moderation, censorship, and degraded privacy. Any policy proposal for tackling harms stemming from GAI-generated content should be evaluated carefully to ensure that the solutions will not result in over-enforcement or have collateral effects that will damage free expression or result in democratic harms.

Policymakers should consider the authentication and content provenance solutions that do not rely on watermarking synthetic content. Watermarking synthetic content is an often-discussed policy solution that merits additional investigation, but the technology and techniques being developed are not yet up to the task. An alternative is to invest in solutions to confirm and track the authenticity of *genuine* content. Bolstering authentic content builds trust in factuality and truth, rather than fixating on rooting out fake and synthetic content. Such an approach will likely have a high rate of adoption among good actors whereas other methods focused on synthetic

content would amplify the potency of any disinformation bad actors manage to sneak past detection.

In general, though, Public Knowledge advocates for solutions that address the harms associated with disinformation no matter how they originate. The resulting policy solutions would encompass things like requirements for risk assessment frameworks and mitigation strategies; transparency on algorithmic decision-making and its outcomes; access to data for qualified researchers; guarantee of due process in content moderation; impact assessments that show how algorithmic systems perform against tests for bias; and enforcement of accountability for the platform's business model (e.g., paid advertising).

Legislative Proposals for Moderation of Synthetic Content

As noted above, we believe there are certain circumstances where trade offs between free expression and content moderation are necessary, like in the context of elections. For instance, the AI Transparency in Elections Act requires labeling elections-related AI-generated content within 120 days before Election Day, aligning with existing disclaimer requirements for political ads. This bill attempts to balance constitutional concerns with transparency needs by excluding minor AI alterations and potentially parody or satire. However, its time limitations fail to address post-election AI-related disinformation risks, such as those the nation collectively experienced after the 2020 election. Conversely, the Protect Elections from Deceptive AI Act creates a federal cause of action for content involving a candidate's voice or likeness and prohibits distributing AI-generated content for election influence or fundraising. While well-intentioned, this legislation could potentially infringe on political speech because, despite the name of the bill, it lacks any requirement that the content actually be deceptive in intent or in effect, and instead presumes that *anything* AI-generated is deceptive. This would empower candidates to sue over content that is *not* deceptive or harmful. This approach risks incentivizing litigiousness to silence critics and public debate, potentially leading to the censorship of political discourse by candidates and non-candidates alike, including journalists and nonprofits.

Thanks in part to powerful stakeholders in the entertainment industry, an enormous amount of the current focus on content from generative AI has to do with digital replicas, defined most recently by the Copyright Office as "a video, image, or audio recording that has been digitally created or manipulated to realistically but falsely depict an individual." There are a couple of bills – namely the NO FAKES Act and the No AI FRAUD Act – specifically aiming to protect *public figures'* publicity rights against unauthorized AI-generated replicas and digital depictions and to hold platforms liable for hosting those unauthorized replicas. Public Knowledge does not support these bills: they both adopt a flawed and complex intellectual property rights framework, fail to adequately address non-economic harms, and create problematic platform liability issues that could lead to over-moderation. As noted above, we have previously detailed the harms and explored other potential remedies for digital replicas.

Executing the Vision for Free Expression and Content Moderation

Many of the solutions we have framed will call for ongoing enforcement and evolution as technological capabilities develop over time. Given the pace of innovation in digital technology and the need for specific, technical expertise to regulate it, we strongly believe a sector-specific, dedicated digital regulator is required.

The role of government is constitutionally bound in regard to both citizens' free expression and platforms' content moderation. However, there is a strong tradition of promoting positive content (e.g., educational content), public safety (e.g., emergency alert system), and diversity and localism in regulation of electronic media. The fact is, our nation has always used policy to ensure that the civic information needs of communities were met. (Public Knowledge explored this tradition in a white paper explaining how we can combat misinformation through policy uplifting local journalism.) One of the core tenets of this history is that whenever there have been changes in technology, new, evolved, or renewed regulators as well as regulations have episodically been required to ensure that the public interest is protected. Often, this has taken the form of a dedicated, empowered regulator with both the expertise and the agility to understand and address both technological and societal change.

In our view, the same is true today. As we've noted in sections above, the concentration of private power over public discourse is itself a threat to free expression. The key elements and functions of a dedicated regulator, such as fostering competition, requiring interoperability, ensuring strong privacy protections, and prohibiting discrimination, would allow more consumer choice and the selection of platforms aligned with a user's values. The regulator may also have roles more directly related to the theories we have laid out above. For example, it may take on aspects of consumer protection and safety by enforcing requirements for clear terms of service, due process, algorithmic transparency and choice, while also ensuring access to data for researchers. It would also be the appropriate body to determine the role and definition of concepts like fiduciary duties, duties of care, and codes of conduct in regard to content moderation.