common sense®

Submitted via email to criley@rstreet.org.

May 3, 2021

Chris Riley
Senior Fellow, Internet Governance
1212 New York Ave NW, Suite 900
Washington, DC 20005

RE: Opportunities and Challenges in Online Content Management

Dear Chris,

We appreciate the efforts of R Street to identify opportunities and challenges in content moderation and management online. Common Sense offers the following comments to share our perspective on how to improve the digital ecosystem through the lens of how content amplification and negative social experiences can impact children and teens.

The modern internet offers frictionless opportunities for endless consumption, be it of ads, media, or user-generated content. Kids face a dizzying array of digital platforms, some overly powerful, some truly toxic, and others misunderstood or confusing for adults to grasp. Underlying many of these platforms and services, unfortunately, is a business model that is designed to engage and extract kids' attention. Teens understand this: they think these platforms are designed to make them spend more time on their devices and distract them and their friends.[1] Tech companies claim this is not the case, but this claim is increasingly impossible to reconcile with online reality or supported by the information that platforms publicly provide.[2]

Common Sense supports laws and regulation that could improve online experiences, like limits on manipulative design, so-called "dark patterns," and algorithmic amplification of harmful and sensationalist content to kids, but online platforms can make changes now and we encourage them to do so. Simple design changes, or just providing more information about how they're ranking and categorizing people, pages, and content, are something platforms can do right now.

Too often, online platforms have taken a laissez-faire approach to any responsibility to promote a safe and healthy online community.[3] That isn't to say one should not support or enable a wide

[1] Rideout, V., & Robb, M. B. (2018). Social media, social life: Teens reveal their experiences.
[2] Justin Hendrix, Ben Sasse is right: the claims of Big Tech and its critics cannot be reconciled, Tech Policy Press (Apr. 27, 2021), https://techpolicy.press/ben-sasse-is-right-the-claims-of-big-tech-and-its-critics-cannot-be-reconciled/.
[3] Omegle may well be the poster child for this, but this perspective is prevalent across platforms like Parler and 4chan. Historically, even more mainstream social media platforms like Twitter and Reddit prioritized

variety of different online communities and approaches to content moderation -- particularly for adults -- but it is rarely, if ever, appropriate for a platform to disclaim responsibility and put the onus entirely on users to protect themselves.[4] This is particularly the case when children and teens are involved.

First, platforms should know where children and teens are. One lesson that should be taken from ongoing debates about operators' knowledge of kid activity under the Children's Online Privacy Protection Act (COPPA): Platforms should know their audience. Some platforms are clearly directed to children, but all platforms should honestly assess their user base and recognize when they have large populations of kids that use their products. This does not mean either intrusive age-checking mechanisms for each user or accepting that age-gating is a sufficient protection. In the current online environment, ignorance is not a sensible business decision.

Second, one global approach is for platforms to consider the best interests of the child. This concept derives from Article 3 of the United Nations Convention on the Rights of the Child and has been embraced by the UK Information Commissioner's Office (ICO) in its Age Appropriate Design Code, which goes into force this September.[5] This requires companies to consider the specialized needs of child users, take into account the ages and developmental capacities of users, and support that through the design of an online service. Companies should be encouraged to rely on evidence and advice from expert third parties. Among considerations identified by the ICO, platforms should consider how they can:

- Keep kids and teens safe from exploitation, including the risks of commercial or sexual exploitation;
- Protect and support kids' health and wellbeing;
- Protect and support kids' physical, psychological and emotional development;
- Protect and support kids' need to develop their own views and identity;
- Protect and support kids' right to freedom of association and play;
- Recognize the role of parents in protecting and promoting the best interests of the child and support them in this task; and
- Recognize the evolving capacity of the child to form their own view, and give due weight to that view.

We recognize that R Street's priorities are to promote policies that advance free markets and limited government, but the incentives to prioritize kids do not always exist online or off. While content management can be expensive and many, if not all, solutions could hurt an attention-based online business model, it must be acknowledged that the current problem is directly attributable

free speech principles over safety. See also Nicholas Kristof, The Children of Pornhub, N.Y. Times (Dec. 4, 2020), https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html.

[4] See https://www.omegle.com (warning that teens "[u]se Omegle at your own peril.").

[5] Best interests of the child, Age Appropriate Design: A Code of Practice for Online Services, UK Information Commissioner's Office (2020), https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services/1-best-interests-of-the-child/.

to a wildly-profitable online ecosystem that is driven by advertising and engagement. Toxic communities have always existed online, but the current status quo involves platforms that claim a desire to be responsible actors while refusing to acknowledge the core problem of prioritizing growth and user engagement above all else.

Common Sense's view is that any platform or service that is directed to or ultimately used by children and teens should have these costs baked into the product or service. Further, invocation that content moderation does not scale is an excuse that prioritizes free expression over safety, good digital citizenship, and other community values -- a calculation that must be questioned where young people are concerned.

Comments on R Street's Propositions follow, and we look forward to seeing how you categorize opportunities to improve content management online.

Joseph Jerome
Director, Platform Accountability

--

Proposition 1: Down-Ranking and Other Alternatives to Content Removal
We always encourage platforms to provide more clarity about how they moderate and rank content. Lawmakers, as well as the public, frequently view content management as a binary: content is either left up or taken down. Down-ranking and other types of content-suppression receive less attention, and criticism of the bipartisan Platform Accountability and Consumer Transparency (PACT) Act, which Common Sense supports, has highlighted this issue.[6] Platforms should explore creative approaches to content moderation. We are curious to see more about so-called virality circuit breakers,[7] but as discussed below, we also support efforts to embed friction into social media. This includes "time outs"[8] and other temporary limits on sharing of problematic posts and placing comments into a holding pen where they can be approved -- or not -- by the original content poster.

However, the failure of platforms both to (1) sufficiently explain their general method or approach to content moderation and/or (2) adequately communicate why certain content is problematic contributes to a lack of public understanding. It is accurate that people are upset when they *feel*

[6] Mike Masnick, PACT Act Is Back: Bipartisan Section 230 'Reform' Bill Remains Mistargeted And Destructive, Techdirt (Mar. 17, 2021), https://www.techdirt.com/articles/20210317/12015646438/pact-act-is-back-bipartisan-section-230-reform-bill-remains-mistargeted-destructive.shtml.
[7] Erin Simpson & Adam Conner, Fighting Coronavirus Misinformation and Disinformation, Center for American Progress (Aug. 18, 2020), https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation/.
[8] Roblox includes disciplinary actions ranging from warnings to temporary and permanent bans. See https://en.help.roblox.com/hc/en-us/articles/360020870412-Understanding-Moderation-Messages.

like they are being shadowbanned or suppressed online, but this irritation may stem from a foundational lack of understanding of how platforms moderate content and algorithmically boost feeds. For kids in particular, this lack of feedback can be confusing and impair the ability to model better behavior.

In terms of how platforms might communicate or build upon down-ranking systems, Common Sense has supported labeling efforts, both in terms of partisan political content[9] and automated bots that can artificially boost and promote traffic.[10] More recently, Common Sense has called for labeling of retouched or "photoshopped" commercial posts that promote unhealthy body images.[11] The presence of these sorts of labels could be used as signals to facilitate down-ranking or suppression of content that is not appropriate or healthy for children.

Proposition 2: Granular/Individualized Notice to Users of Policy Violations
As mentioned above, the failure of platforms to communicate with content creators about why posts are problematic -- or violate community standards -- contributes to confusion, allegations of bias, and generalized criticism about online content moderation. Platforms should not be put into the position of having to litigate each and every content decision they make with users, but platforms should be encouraged to and perhaps commended for providing more individualized outreach.

We think this could be particularly useful for teen audiences. Granular notice, particularly of initial violations, can help set user expectations and help teens understand what behavior will not be tolerated. One area that may warrant further consideration is the video game space, where game publishers are struggling with not just harassing behavior but the need to monitor and police cheaters.[12] As you identify, one overarching challenge is to what degree companies should be expected to justify every policy violation they take action against. There is no easy way to reconcile calls for digital due process rights, but platforms should be encouraged to experiment with ways to make notice of policy violations a teaching opportunity. To alleviate confusion, platforms could also do more to explain false-negatives and concede some of the underlying subjectivity in content moderation decisions; this sort of transparency should go hand-in-hand with establishing more elaborate takedown/suppression appeals processes.

[9] See Common Sense, 2020 Social Media Voter Scorecard, available at https://www.commonsensemedia.org/social-media-voter-scorecard.

[10] Press Release, Common Sense Supports BOT Act to Identify Bot Accounts on Social Media (Apr. 16, 2018), https://www.commonsensemedia.org/about-us/news/press-releases/common-sense-supports-bot-act-to-identify-bot-accounts-on-social-media.

[11] CA A.B. 613 (Social media: retouched images: disclosure) (2021), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202120220AB613.

[12] See Disruption and Harms in Online Gaming Framework, ADL & Fair Play Alliance (2020), https://www.adl.org/fpa-adl-games-framework. There is also some anecdotal support for the idea that major game publishers have begun to see a need for individualized communication for player bans and other major disciplinary actions.

Another approach is for platforms to provide more individualized and repeated notice of community guidelines. Promoting positive online behavior is important for both children and teens, but well-meaning written guidelines mean little if no one is readily exposed to them. Studies have shown that posting clear and digital rules promotes prosocial behaviors and can deter casual violations online,[13] and social experiences should ensure that kids see top-level behavior expectations when first using a social platform.[14]

Unfortunately, this is not a standard practice -- even among apps directed at children. Roblox and TikTok are illustrative. Both platforms are incredibly popular among kids of all ages, and both companies have recently publicly prioritized the importance of creating a safe and healthy online environment for younger users.[15] Despite Roblox's efforts at policing itself -- including using human moderators, implementing digital-civility rules, and more -- bad actors continue to find ways to share dubious content.[16] Meanwhile, according to TikTok's most recent transparency report, the largest category of videos removed from the platform involve violations of TikTok's minor safety rules.[17] However, neither platform actively encourages younger users to engage with their community guidelines. Children may never be directly taught what either platforms' expectations for their community are. This contrasts with how schools aim to set rules and expectations from the moment that children enter the classroom.[18]

Proposition 3: Use of Automation to Detect and Make Classifications of Policy Content (including filters)

While automation is important both for moderation at scale and functions as a talking point for tech companies, Common Sense's basic position is that more trained and supported human moderation is needed. Human moderation is essential and one primary problem is that platforms are not sufficiently internalizing the costs of human moderation, including the trauma and well-being of moderators. R Street's position is confusing, and pointing to automation as a "cheaper" solution at scale while highlighting automate's upfront and ongoing costs as a challenge.

Proposition 4: Clarity and Specificity in Content Policies to Improve Predictability at the Cost of Flexibility

[13] J. Nathan Matias, Posting Rules in Science Discussions Prevents Problems & Increases Participation (Apr. 29, 2019), https://civilservant.io/moderation_experiment_r_science_rule_posting.html.
[14] Joseph Jerome, Safe and Secure VR: Policy Issues Impacting Kids' Use of Immersive Tech, Common Sense Media (Mar. 2021), available at https://www.commonsensemedia.org/kids-action/blog/vr-irl.
[15] See David Baszucki, Building a Safe and Civil Community, Roblox (Aug. 19, 2020), https://blog.roblox.com/2020/08/building-safe-civil-community/; Tracy Elizabeth & Alexandra Evans, Supporting youth and families on TikTok, TikTok (Nov. 17, 2020) ,https://newsroom.tiktok.com/en-us/supporting-youth-and-families-on-tiktok.
[16] Erin Brereton, Roblox Review, Common Sense Media, https://www.commonsensemedia.org/website-reviews/roblox.
[17] TikTok Transparency Report -- July 1, 2020-Dec. 31, 2020 (Feb. 24, 2021), https://www.tiktok.com/safety/resources/transparency-report-2020-2?lang=en&appLaunch=.
[18] Ben Fenton, Living Codes of Conduct, www.ascd.org/ascd-express/vol5/507-fenton.aspx.

As we have seen with data use and privacy policies, terms of service and community guidelines present a difficult balance between being thorough (and legally compliant in some instances) and clear and understandable to the average person. This is a fundamental problem that contributes to confusion and misunderstanding about how content management is handled. As a baseline measure, platforms should consider ensuring that community guidelines are not just visible but easy to find and eas(ier) to understand.[19]

It should not be discounted that it is difficult for interested stakeholders, let alone parents or kids, to understand where platforms stand on their content policies. One lesson of the 2020 U.S. election was that content policies and community guidelines were evolving rapidly, and often, anyone not deeply in the weeds was left behind as to the current state of thinking.

To keep track of content management practices on Facebook, the Stanford Internet Observatory "scoured Facebook Newsroom blog posts and tweets by executives, and even used the Wayback Machine from the Internet Archive for a side-by-side comparison of changes."[20] This is not acceptable, but too often, community guidelines are treated like legal documentation and privacy policies. Again, there may be lessons that can be learned from how schools implement codes of conduct, which aim to reinforce the responsibility students have to fulfill positive expectations in schools, and often require adults and other authority figures to set a tone for a school's culture.[21]

Rule specificity is a harder issue. We believe that platforms should be accountable for upholding the promises they make to their online communities, but prioritizing specificity can turn online content moderation into a "gotcha" game that may also encourage platforms to narrow what their expectations for good behavior are. One suggestion would be to encourage platforms to better communicate with stakeholders, including the general public, what areas are of particular concern on the platform. For instance, Pinterest very clearly began prioritizing responses to anti-vaccination misinformation on its platform in 2019.[22]

One approach could be to better improve reporting tools, provide additional transparency into what sorts of violative content are being seen (and flagged) on the platform, and adapt and communicate community guidelines in response. This is something that Facebook attempts to do to varying degrees,[23] but this sort of feedback loop and communication appears to be sorely lacking on kid-focused platforms and social VR. Reporting on content decisions based on

[19] Transparency and Accountability, https://www.esafety.gov.au/about-us/safety-by-design.

[20] Carly Miller, Facebook, It's Time to Put the Rules in One Place, Lawfare (Mar. 5, 2021), https://www.lawfareblog.com/facebook-its-time-put-rules-one-place.

[21] Ben Fenton, Living Codes of Conduct, www.ascd.org/ascd-express/vol5/507-fenton.aspx.

[22] Ifeoma Ozoma, Bringing authoritative vaccine results to Pinterest search, Pinterest Newsroom (Aug. 28, 2019), https://newsroom.pinterest.com/en/post/bringing-authoritative-vaccine-results-to-pinterest-search.

[23] See ADL, Facebook's Transparency Reporting Continues to Obscure the Scope and Impact of Hate Speech (Nov. 20, 2020), https://www.adl.org/blog/facebooks-transparency-reporting-continues-to-obscure-the-scope-and-impact-of-hate-speech.

community guidelines remains a newer form of transparency reporting,[24] and one for which standardization and best practices are needed, particularly for platforms targeted to and used by children.

Proposition 5: Friction in the Process of Communication at Varying Stages (or, more broadly, UX design as a way to encourage user thoughtfulness/manage user flagrancy)
We agree with R Street's premise that most of the boldest efforts to improve online communities involve introducing friction onto social media. This makes sense: the fundamental problem is an underlying business model that is designed to engage and extract kids' attention and offer immediacy endlessly. Adding friction, such as requiring individuals to think twice before they post or reducing the volume and flow of content, go directly to counteracting an engagement-based business model.

This is why it is disappointing to see one purported challenge with this proposition is that friction will reduce ad revenue by reducing "impulse-clicking." The reliance on ad-dollars that flow from individuals making impulsive choices online is at the core of the problem today.[25] This is particularly true for kids. Research shows that teens are more likely to share without thinking, focusing on the immediate present and not long-term consequences as their brains prioritize rewards and minimize risks.[26] Given susceptibility to peer pressure, teens will even stay and share in online communities that are no longer enjoyable to them, as that is where their friends are.[27] Social media is designed in a way that is particularly appealing to teenagers and emerging adults, when individuals are oriented toward others, belonging, groups, and acceptance.

This proposition requires R Street to engage in the emerging discussion around so-called "dark patterns" or, more appropriately, the field of manipulative design.[28] Manipulative design takes further advantage of kids by subverting their choices and autonomy and causing them not only to give up more information than otherwise but also to spend more time clicking and scrolling and taking them down rabbit holes. Indeed, almost half of teens report feeling "addicted" to their

[24] See Spandana Singh & Kevin Bankston, The Transparency Reporting Toolkit: Content Takedown Reporting, OTI (Oct. 25, 2018), https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/.
[25] While perhaps less germane to content management, it is worth highlighting bipartisan concern about in-app purchases and how platforms have manipulated kids into paying for products and services online. Regardless, reducing "impulse-clicking" is not a bad thing.
[26] Adriana Galvan et al., Earlier Development of the Accumbens Relative to Orbitofrontal Cortex Might Underlie Risk-Taking Behavior in Adolescents 26 Journal of Neuroscience 25 (2006); Adriana Galván and Kristine M. McGlennen, Enhanced Striatal Sensitivity to Aversive Reinforcement in Adolescents versus Adults, 25 (2) J. of Cognitive Neuroscience 284–296 (2013).
[27] Center for Digital Democracy and the Campaign for a Commercial Free Childhood Comments before the Federal Trade Commission, Competition and Consumer Protection in the 21st Century, Hearing #12: The FTC's Approach to Consumer Privacy (2019), at 12, citing Taylor Lorenz, Teens Are Being Bullied 'Constantly' on Instagram, The Atlantic (Oct. 10, 2018) ("[T]eens stay on Instagram even with cyberbullying because "quitting wasn't an option.").
[28] See Bringing Dark Patterns to Light: An FTC Workshop (Apr. 29, 2021), https://www.ftc.gov/news-events/events-calendar/bringing-dark-patterns-light-ftc-workshop.

phones.[29] Social media platforms offer immediate and variable rewards, just like casino games, and these can lead to compulsion.[30] Infinite scrolls do not offer any visual cues or reminders to young people to stop. "Awards" for repeat use or actions, like Snapchat's "Snapstreaks" for daily communication with friends, encourage unnecessary and excessive engagement. Autoplay videos keep kids glued to the screen even after a show is over. As a result, Common Sense has supported and advocated for legislative reforms specifically targeted toward manipulative design in technology, including the Kids Internet Design and Safety (KIDS) Act and the Deceptive Experiences To Online Users Reduction (DETOUR) Act. Both pieces of legislation have provisions that could inform R Street's thinking on content management.

Proposition 6: Experimentation with, and Transparency in, Weightings in Recommendation Engines

One large tension point here is ensuring some group of outside stakeholders can get insight into how companies are modifying their algorithms. Common Sense, for example, has had some conversations with tech platforms about how to establish "do not amplify" criteria. Platforms like Twitter monitor thousands of accounts and hashtags, and we have seen TikTok invest in systems to reduce and suppress discoverability of disinformation and terms of incitement. Misleading hashtags, including #stopthesteal and other QAnon content, can be redirected to relevant community guidelines rather than receive any search results.[31]

This requires continual updating and monitoring as content and terminology evolves. Another problem is that providing too much public transparency about anti-amplification efforts can let bad actors effectively game the system and undermine these safeguards. On the other hand, there is little reason for civil society to trust that platforms can be responsible for doing this without some oversight by outside stakeholders. Even where platforms offer specialized "hotlines" for civil society, tech companies have still responded slowly and ineffectively to complaints.[32] It is unclear what an optimum solution is, but some formalized mechanism by which trusted outside parties can engage with platforms and flag emerging issues is needed.

Proposition 7: Separate Treatment for Paid or Sponsored Content, such as Reviewing for Heightened Standard

We agree that online advertising warrants a heightened standard of review. Indeed, many proposals in the ongoing debate about reforms to Section 230 have prioritized excluding

[29] Rideout, V., & Robb, M. B. (2018). Social media, social life: Teens reveal their experiences.

[30] Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, p. 448 (2019).

[31] Michael Beckerman, TikTok's H2 2020 Transparency Report, TikTok (Feb. 24, 2021), https://newsroom.tiktok.com/en-us/tiktoks-h-2-2020-transparency-report.

[32] See, e.g., Press Release, Georgetown Law's Civil Rights Clinic and Father of Slain Journalist File FTC Complaint to Remove Violent Murder Videos from YouTube (Feb. 20, 2020), https://www.law.georgetown.edu/news/georgetown-laws-civil-rights-clinic-and-father-of-slain-journalist-file-ftc-complaint-to-remove-violent-murder-videos-from-youtube/ (highlighting examples where YouTube has been unresponsive to trained volunteers that monitor YouTube for death videos clearly in violation of YouTube's Terms of Use).

commercial context from the scope of the statute's liability shield.[33] The Safeguarding Against Fraud, Exploitation, Threats, Extremism and Consumer Harms (SAFE TECH) Act, for example,which Common Sense supports, aims to ensure that Section 230 protections do not apply to ads or other paid content, and other tech policy advocates have begun to discuss the idea of removing liability protections for online advertising.[34] (Much of the criticism about the SAFE TECH Act has focused on the potential breath of its application to companies that have "accepted payment to make the speech available," which could capture hosting platforms and others in the tech stack, but critics have been unwilling to offer more targeted language themselves likely due to their general opposition to the SAFE TECH Act.[35])

Opponents remain concerned that this would sweep in too much speech, particularly in the context of political speech. If that is a concern, we would suggest incorporating some sort of threshold, such as the $500 limit proposed by the Honest Ads Act.[36] But this challenge is not insurmountable and should not distract from the basic premise that advertising and other paid speech warrants different policy considerations from speech writ large and platforms should be expected to have additional requirements imposed upon them in this space. Advertising has long been a primary subject of civil rights laws, and the online advertising ecosystem frequently concerns itself with the importance of combating ad fraud and ensuring ad integrity.

[33] See, e.g., Press Release, Warner, Hirono, Klobuchar Announce the SAFE TECH Act to Reform Section 230 (Feb. 5, 2021), https://www.warner.senate.gov/public/index.cfm/2021/2/warner-hirono-klobuchar-announce-the-safe-tech-act-to-reform-section-230.

[34] Bertram Lee, Where the Rubber Meets the Road: Section 230 and Civil Rights, Public Knowledge (Aug. 12, 2020), https://www.publicknowledge.org/blog/where-the-rubber-meets-the-road-section-230-and-civil-rights/.

[35] Twitter Thread between Jeff Kosseff and Joseph Jerome, beginning at https://twitter.com/jkosseff/status/1358156938165559302.

[36] S.1356 - Honest Ads Act, 116th Congress (2019-2020).