STRIPED

A PUBLIC HEALTH INCUBATOR

Strategic Training Initiative for the Prevention of Eating Disorders

Understanding Algorithms and Social Media Platform Design

A Judicial Primer on Algorithmic Systems and Digital Harm

FEBRUARY 2026

Authors
Nancy Costello, Jill R. Kavanaugh, Caroline Berish, Mallory Kean, Jacob Appel, Marilyn Bromberg, S. Bryn Austin

Funding
This project is supported by the American Association for Justice Robert L. Habush Endowment.

Disclosures
The authors do not have financial conflicts of interest to declare.

About this Publication
This project is a collaboration between scholars with the Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED) based at the Harvard T.H. Chan School of Public Health and Boston Children's Hospital and Michigan State University College of Law.

STRIPED is a research and training program dedicated to bringing public health approaches across disciplines and sectors to prevent eating disorders. Founded and directed by S. Bryn Austin, ScD, STRIPED brings together experts in public health, medicine, law, and policy to advance evidence-based strategies to prevent eating disorders, shape healthier digital environments, and promote body confidence.

Collaborating scholars from Michigan State University College of Law contributed expertise in First Amendment, media, and technology law. Professor Nancy Costello directs the First Amendment Law Clinic and teaches media law, intellectual property, and advocacy in the digital age. Trained as both a journalist and lawyer and with practice experience in commercial litigation, defamation, and e-business law, she plays a key role in STRIPED's legislative work to protect youth from social media harms. Jacob Appel, Chief Strategist at the algorithmic auditing firm O'Neil Risk Consulting and Algorithm Auditing (ORCAA), adds complementary expertise in evaluating platform design and algorithmic systems. ORCAA investigates these systems and helps regulators and enforcement agencies develop rules and tests to ensure they are safe and fair.

Together, we bring complementary strengths in research, law, technology, and mental health. Our collaboration bridges public health, technology, and legal scholarship to advance evidence-informed policy and foster systems that protect youth well-being.

Acknowledgements
We would like to thank the many experts, spanning the fields of law and technology, who we consulted with for our research and in the creation of this white paper. Their invaluable insights and perspectives were instrumental in informing our research and we are grateful for the generous contributions of their time and expertise. We are especially grateful to Dr. Meg Salvia for her thoughtful feedback and support throughout the development of this work.

Suggested Citation
Costello N, Kavanaugh JR, Berish C, Keane M, Appel J, Bromberg M, Austin SB. *Understanding Algorithms and Platform Design: A Judicial Primer on Algorithmic Systems and Digital Harm.* Boston, MA: STRIPED (Strategic Training Initiative for the Prevention of Eating Disorders); February 23, 2026. http://hsph.me/algorithm-auditing

Table of Contents

Executive Summary

Engagement-based algorithms shape what people, especially youth, see and experience online. On major advertising-supported platforms, engagement-based algorithms determine which posts appear, in what order, and to whom. These systems are not neutral hosts of user speech; they are product designs built to maximize time and advertising revenue. For courts, distinguishing between user content and platform design is essential to evaluating foreseeability of harm, product safety, and the reach of the First Amendment and Section 230.

This white paper has two goals. First, it explains how engagement-based algorithms work, how they differ from search and chronological feeds, and why their design is consequential for minors. Second, it connects these technical realities to civil law frameworks, highlighting case law that distinguishes editorial choices protected by the First Amendment from potentially actionable product design.

Engagement-based algorithms operate as adaptive prediction loops: Social media platforms collect behavioral and demographic data, predict what will capture each user's attention, assemble a personalized feed in real time, and adjust continuously based on micro-behaviors such as scrolls, pauses, and clicks. Design features like infinite scroll, autoplay, streaks, and variable notifications reduce natural stopping points and encourage repeated use. These functions operate independently of any particular message or viewpoint; they are optimized for attention and data, not truth, accuracy, or developmental suitability for younger users who are still developing cognitively and emotionally. For minors, these systems create predictable categories of harm:

1. **Exposure harms:** Repeated presentation of violent, sexual, appearance-based, or otherwise harmful content youth did not seek out.
2. **Escalation harms:** Rapid progression from general-interest material to more extreme content, including dangerous physical challenges or self-harm.
3. **Engagement harms:** Compulsive use, sleep disruption, and downstream effects on learning, mood, and functioning.

These harms arise from the mechanics of algorithmic curation at scale—automated systems that select and promote content across vast numbers of users—not from isolated posts. Research links design features to increased time, frequent checking, and greater exposure to extreme material. The legal analysis focuses on whether a platform's design choices create foreseeable risks of harm distinct from user-generated content.

Recent Supreme Court and appellate decisions suggest a growing willingness to treat some algorithmic functions as a product design rather than expressive activity, especially where a platform's technology responds solely to a user's behavioral signals rather than the platform's editorial judgment. This distinction opens space for legislation and litigation that could challenge and regulate structural design choices by social media platforms and do so constitutionally. One option could be requiring independent algorithm risk audits that would measure harm caused by engagement-based algorithms.

With looming lawsuits and legislation, judges and clerks will increasingly be asked to decide disputes that turn on how these social media systems work, what they are optimized to do, and whether resulting harms were foreseeable. This paper aims to provide a foundation for that task and suggest tools, including doctrinal legal frameworks, risk audits, expert testimony, and amicus input, to help courts address design-driven algorithmic harms without undermining protected free expression on social media.

Understanding the Role and Impact of Algorithms

In today's digital environment, algorithms shape a substantial portion of what people, minors, see, engage with, and spend time on. By determining how information flows across social media platforms, these systems influence behavior, attention, and well-being. Courts now encounter cases involving online harms, consumer protection, and platform responsibility, and judges increasingly must evaluate whether and how the design of these systems creates foreseeable risks of harm. This paper explains how engagement-based algorithms on social media function, why their design matters for minors, and how these mechanisms differ from editorial choices and protected speech.

Defining Key Terms: "Content" and "Design"
For clarity, this paper uses these terms in a specific way:

- "*Content*" refers to the words, images, videos, and other material that users or advertisers upload to a social media platform.
- "*Design*" refers to how a social media platform organizes, displays, ranks, recommends, or withholds content on social media. For example, design includes feed-ranking algorithms, autoplay, infinite scroll, and notification systems.

This paper focuses on engagement-based algorithms, including recommendation algorithms, personalized feed algorithms, and engagement-based ranking systems used in advertising-supported platforms. Chronological lists or user-initiated search tools are discussed for contrast but are not the primary focus.

In this section, the distinction between content and design on social media platforms is focused solely on a platform's technological function. The legal significance of that distinction for the First Amendment and Section 230 is addressed in the second half of the paper (*Intersection with Civil Law*).

How Social Media Algorithms Determine What Users See

Many civil disputes involving social media do not focus on a single post, video, or piece of "speech." Instead, they center on how a platform's product is designed—specifically, how the system sorts, ranks, and amplifies content in ways that influence user behavior.

What Is an Algorithm? A Plain-Language Definition

In this paper, an algorithm is a set of instructions a computer follows to perform a task or make a decision. On social media, engagement-based algorithms determine:

- Which posts or videos are shown
- In what order they appear
- How prominently and for how long they remain visible
- To which users they are shown

These decisions are typically based on predictions about what will keep each user watching, scrolling, tapping, or clicking on social media platforms. The following section describes how these engagement-based algorithms operate in social media feeds in practice.

How Prediction Works in Social Media Feeds

Social media feeds operate as adaptive prediction systems rather than chronological lists.[1] Each time a user opens an app, engagement-based algorithms estimate what content is most likely to capture their attention, display those items first, and then revise their estimation based on how the user reacts. This cycle runs continuously, producing a highly personalized environment, tailored to the user's observed behavior. In broad terms, engagement-based algorithms operate as iterative prediction loops:[2]

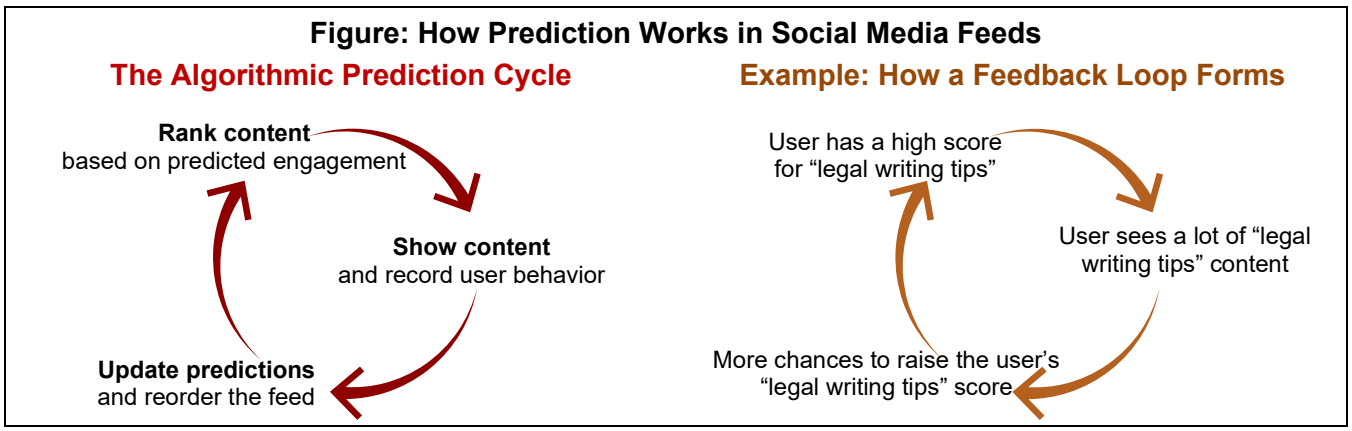
1. **Input**: The platform gathers data about user behavior, demographics, and inferred interests.
2. **Modeling**: Machine-learning models compare these data to patterns across millions of users to predict likely engagement.
3. **Output**: The feed is assembled in real time, ordered according to predicted engagement.
4. **Feedback**: The user's interactions become new data, prompting immediate recalibration.

This process is driven by several mutually reinforcing mechanisms:[3]

- **Engagement-Based Ranking**: Determines which posts appear and in what order, fundamentally shaping what the user sees.
- **Engagement Signals**: Micro-behaviors—likes, comments, scroll speed, pauses, rewatches—are interpreted as preferences, even when they are incidental or fleeting.
- **Feedback Loops**: Small or incidental engagement signals shift the feed's predictions, prompting the system to show more content in that direction and creating additional opportunities for engagement.

These automated feedback loops, rather than discrete editorial decisions made by social media administrators, determine most of what users encounter online. The graphic below (Figure: How Prediction Works in Social Media Feeds) shows how the prediction cycle works and how a "high score" simply reflects the system's estimate that a user is likely to engage with a certain type of content, reinforcing a feedback loop.

Figure: How Prediction Works in Social Media Feeds



[1] Arvind Narayanan, *Understanding Social Media Recommendation Algorithms*, KNIGHT FIRST AMEND. INST. AT COLUMBIA UNIV. (Mar. 9, 2023), https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

[2] Chris Meserole, *How Do Recommender Systems Work on Digital Platforms?* BROOKINGS INST. (Sep. 21, 2022) https://www.brookings.edu/articles/how-do-recommender-systems-work-on-digital-platforms-social-media-recommendation-algorithms/.

[3] Hannah Metzler & David Garcia, *Social Drivers and Algorithmic Mechanisms on Digital Media*, 19 PERSP. ON PSYCH. SCI. 735, 735–48 (2024).

Analogy: The Grocery Store

A familiar real-world analogy helps illustrate how design, rather than the product itself, drives outcomes. Traditional media and retail environments also use design to attract attention: Newspapers choose headlines and front-page placement, television schedules programs in particular time slots, and stores place items on shelves at eye level or near the checkout to attract shoppers. What is distinctive about social media is the speed, scale, and personalization with which social media feeds reorganize themselves in response to each user's behavior.

In this analogy, the product represents content, or expressive material, while the shelving system represents design choices that influence which speech is most visible; this affects exposure without altering content or adopting a viewpoint.

One way to see the difference between content and design, and why it matters, is to imagine a grocery store:

- Food companies create the food products (the speech).
- The store decides where items sit on the shelves (the design): cereal at eye level, flour on high or low shelves, candy near the checkout.
- By choosing what is most prominent and what is harder to reach, the store—though it does not create the cereal, flour, or candy—influences what customers notice and buy, including items they might otherwise avoid.

Now imagine a second version of the store that doesn't just have a fixed layout, but reacts to you as you shop, using automated predictions:

- Shelves reorganize themselves automatically based on what you glanced at.
- Items the system predicts you'll want are pulled closer as you move, even if you were trying to avoid them.
- Products you didn't pick up are moved out of sight to prioritize predicted preferences.
- The layout rearranges continuously as the system updates its predicted preferences, moment by moment, based on your behavior as you walk through the aisles.

These dynamic adjustments create self-reinforcing loops: If you linger in the candy aisle, the aisle grows longer, with more candy appearing in front of you, while exits become harder to see.

And crucially, imagine the store doing this differently for every customer at the same time:

- Two shoppers walk down the chip aisle; one sees mostly potato chips (her favorite), while the other sees mostly tortilla chips (his favorite). This difference is not because the store has a view about chips, but because it cares only about what each individual person will reach for.

The layout is personalized, reactive, and optimized moment-to-moment, based on individual behavior and predicted engagement, not on editorial preference or message. This mirrors engagement-based ranking systems online:

- The content (the "product") remains constant.
- But the arrangement (the design) shifts endlessly to maximize attention and consumption, tailored to each user.

The analogy illustrates how design—not content—determines visibility. Traditional media and retail environments use design to draw attention. Online, engagement-based algorithms do this at scale—that is, automatically and continuously across enormous volumes of content and large user populations—and in real time, reorganizing content based on predicted preferences.

For minors, whose attention and reward systems and self-regulation capacities are still developing,[4] these shifting arrangements can be especially powerful. Over time, this design creates self-reinforcing incentives to engage more and disengage less, regardless of content or viewpoint.

Distinguishing User-Driven Features From Design-Driven Features

Not all features function the same way.[5] Courts often must distinguish aspects of a platform shaped by user choice from those shaped by platform design, especially engagement-based algorithms. The following table summarizes common examples.

Categorizing Platform Algorithms for Legal Assessment			
Algorithm Type	**What It Does**	**User Choice Involved?**	**Why It Matters**
Search-Based	Returns information a user explicitly asks for	Yes, entirely user-initiated	Primarily user-initiated retrieval; design plays a smaller role in determining what appears
Chronological Feed	Displays posts in the order they were created	Yes, reflects who the user follows	Displays content in time order based on who the user follows, without using a predictive engagement model
Recommendation and Engagement-Based Ranking Systems	Surfaces and ranks unrequested content based on predicted engagement	Partial/No	Shapes what users see and how long they stay

For the legal analysis that follows, the central focus is on recommendation and engagement-based ranking systems, as defined earlier. Search-based tools and chronological feeds are discussed mainly as points of comparison.

How Automated, Engagement-Based Design Replaced Human Judgment

For most of the twentieth century, the flow of information was shaped by human judgment. Newspaper editors, television producers, and radio hosts decided which stories to highlight, how to frame them, and what audiences needed to know. These decisions were imperfect and sometimes biased, but they were rooted in identifiable professional standards: Accuracy, relevance, public interest, and timeliness. Individuals could reasonably expect that the information they encountered reflected some degree of human deliberation.

---

[4] OFF. OF THE SURGEON GEN., *Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory*, U.S. DEP'T OF HEALTH & HUM. SERVS. (2023), https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf.

[5] Alex Moehring, et al., *Better Feeds: Algorithms That Put People First*, KNIGHT–GEORGETOWN INST. (Mar. 4, 2025), https://kgi.georgetown.edu/research-and-commentary/better-feeds/.

Social media platforms operate differently. Many functions once performed by human editors are now carried out by automated prediction models that prioritize whatever is most likely to keep a user engaged on a platform. The shift from human editorial decision-making to algorithmic optimization has fundamentally altered the information environment:[6]

- The algorithmic system interprets micro-behaviors as signals of preference
- It elevates material that performs well among similar users
- It recalibrates constantly to increase time spent online

This means two users, even within the same household, can encounter entirely different content ecosystems online very quickly. These divergent paths emerge from the way the social media platform is designed to personalize and optimize content delivery, rather than from a single, uniform editorial judgment about what all users should see.

This section is descriptive: It explains how automated systems now perform many of the functions once carried out by human editors and how those systems prioritize engagement. The legal implications of whether these design choices should be treated as speech, editorial judgment, or product design are developed later in this white paper.

Why Social Media Platforms Optimize for Engagement: The Business Model

Most large advertising-supported social media platforms are engineered around a straightforward commercial reality: More time on the platform translates directly into more advertising impressions (the number of times an ad is displayed to a user) and more behavioral data to monetize.[7] Advertising revenue depends on two primary inputs—time spent and behavioral data—and engagement-based algorithms are the mechanism that maximizes both.[8-9]

Unlike many traditional media models, where advertisers buy a fixed placement in advance (for example, a half-page newspaper ad, a 30-second TV spot), most social media advertising operates through instant ad auctions run billions of times per day.[10] When a user opens the app and scrolls, recommendation algorithms conduct a real-time auction to determine which ad to show that specific user at that moment. Advertisers bid for access to that user's attention, and the winning bid is selected by the social media platform's ad-auction algorithm, which weighs both the bid amount and the predicted likelihood that the user will respond.

This auction model creates two incentives:[11]

---

[6] Hastuti et al., *Algorithmic Influence and Media Legitimacy: A Systematic Review of Social Media's Impact on News Production*, 10 FRONTIERS IN COMMC'NS 1, 1–19 (2025).

[7] MediaSmarts, *The Social Media Industry*, MEDIASMARTS, https://mediasmarts.ca/digital-media-literacy/general-information/interactive-media/social-media/social-media-industry (last visited Dec. 30, 2025).

[8] Narayanan, *Understanding Social Media Recommendation Algorithms*, supra note 1.

[9] Christian Montag & Jon D. Elhai, *On Social Media Design, (Online-) Time Well-spent and Addictive Behaviors in the Age of Surveillance Capitalism*, 10 CURRENT ADDITION REPS. 610, 610–16 (2023).

[10] Matthew Johnston, *How Does Facebook (Meta) Make Money?*, INVESTOPEDIA (June 29, 2024), https://www.investopedia.com/ask/answers/120114/how-does-facebook-fb-make-money.asp.

[11] Jerrold Nadler & David N. Cicilline, *Investigation of Competition in Digital Markets: Majority Staff Report and Recommendations*, U.S. HOUSE OF REPRESENTATIVES (2020), https://democrats-judiciary.house.gov/sites/evo-subsites/democrats-judiciary.house.gov/files/migrated/UploadedFiles/Competition_In_Digital_Markets.pdf.

1. **Improve targeting accuracy to raise price per ad.** More precise predictions of user interests make each ad impression (a single showing of an ad to a user) more valuable, incentivizing platforms to collect more data and monitor behavior more closely, so an ad can be more precisely customized for the user.
2. **Increase total time online to expand supply of ads.** More user-minutes spent scrolling means more advertising auctions, more ad impressions, and more ad revenue, incentivizing platforms to optimize the feed for engagement.

Engagement-based algorithms are a direct response to these incentives:[12]

- Predicting what each user will find compelling
- Reorganizing the feed to keep the user active
- Generating continuous behavioral data to strengthen future predictions

As a result, this creates a self-reinforcing cycle: More time online produces more data, more data produces more refined predictions, and refined predictions produce even more time online.[13] These mechanics reflect commercial product-design choices about how to structure and optimize the service (and thus the profits) of social media platforms, rather than case-by-case editorial judgments by the platforms about the merits of specific posts.

The Role of Youth Engagement

A 2022 economic analysis found that major social media platforms generated approximately US $11 billion in advertising revenue from users under age 18 in the United States.[14] This revenue stream depends on keeping minors engaged for long sessions and returning frequently. Research shows that adolescents:

- Spend significant time on short-form video feeds, meaning feeds composed of very short, rapidly cycling video clips.[15]
- Respond strongly to emotionally charged or appearance-based content focused on physical appearance and comparison.[16]
- Have less-developed self-regulation and reward-processing systems.[17-18]

These developmental factors make youth engagement both highly profitable and highly predictable.

---

[12] Hunt Allcott et al., *Digital Addition*, 112 AM. ECON. REV. 2424, 2424–63 (2022).

[13] Smitha Milli et al., *Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media*, KNIGHT FIRST AMEND. INST. AT COLUMBIA UNIV. (Jan. 3, 2024), https://knightcolumbia.org/content/engagement-user-satisfaction-and-the-amplification-of-divisive-content-on-social-media.

[14] Amanda Raffoul et al., *Social Media Platforms Generate Billions of Dollars in Revenue from U.S. Youth: Findings From a Simulated Revenue Model*, PLOS ONE (Dec. 27, 2023), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0295337.

[15] Michelle Faverio, *Teens, Social Media and AI Chatbots*, PEW RSCH. CTR. (Dec. 9, 2025), https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/.

[16] Bohee So & Ki Han Kwon, *The Impact of Thin-Ideal Internalization, Appearance Comparison, Social Media Use on Body Image and Eating Disorders; A Literature Review*, 20 J. OF EVIDENCE-BASED SOC. WORK 55, 55–71 (2023).

[17] D. Albert & L. Steinberg, *Judgment & Decision Making in Adolescence*, 21 J. OF RSCH. ON ADOLESCENCE 211, 211–24 (2011).

[18] Leah H. Somerville, *Special Issue on the Teenage Brain: Sensitivity to Social Evaluation*, 22 CURRENT DIRECTIONS IN PSYCH. SCI.121, 121–27 (2013).

Design Features Built to Maximize Use

To support this model, platforms deploy design features specifically engineered to increase engagement and prolong use, regardless of the particular content being shown, including:

- **Infinite scroll.** The feed never ends; as the user nears the bottom, new posts automatically appear instead of requiring a click to load a new page, eliminating natural stopping points.[19]
- **Autoplay**. When one video ends, the next one starts automatically without the user pressing "play," reducing the need for any deliberate decision to continue.[20] The user must take action to *stop* watching.
- **Variable or intermittent notification cues.** The app sends alerts at unpredictable intervals, and some notifications contain especially rewarding content (such as a new "like" or comment), which encourages frequent checking.[21-22]
- **Streaks and reward mechanisms.** The platform tracks and displays "streaks"—counts of consecutive days a user engages in a particular activity (such as logging in, posting, or exchanging messages with a specific friend)—and offers badges or other rewards for keeping these streaks going, making it feel costly to skip a day.[23]
- **Seamless transitions between videos or posts.** Content appears instantly from one item to the next, minimizing natural pauses that might prompt a user to stop.[24-25]

These features operate independently of the content they deliver. Their primary function is to increase user engagement and time spent on the platform, not to assess accuracy, quality, or developmental suitability of specific content.

Why This Matters for Legal Analysis

The commercial incentives underlying social media explain why many harms stem from design, not content. Engagement-driven features shape what minors see, how quickly content escalates, and how difficult it is for them to disengage.[26] These outcomes flow directly from the product's architecture and revenue model, not from editorial decisions protected by the First Amendment or Section 230. This distinction, between choices made to optimize engagement and choices made to convey a message, becomes central in the legal analysis that follows.

---

[19] Jan Rixen, et al., The Loop and Reasons to Break It: Investigating Infinite Scrolling Behaviour in Social Media Applications and Reasons to Stop, 7 Proc. of the ACM on Human-Comput. Interaction 1, 1–22 (2023).

[20] Kai Lukoff et al., How the Design of YouTube Influences User Sense of Agency, 368 Proc. of the 2021 CHI Conf. on Human Factors in Computing Sys. 1, 1–17 (2021).

[21] Nicholas Fitz et al., *Batching Smartphone Notifications Can Improve Well-Being*, 101 COMPUTS. IN HUM. BEHAV. 84, 84–94 (2019).

[22] Kostadin Kushlev et al., *"Silence Your Phones": Smartphone Notifications Increase Inattention and Hyperactivity Symptoms*. 1 PROC. OF THE 2016 CHI CONF. ON HUMAN FACTORS IN COMPUTING SYS. 1011, 1011–20 (2016).

[23] Christina M. van Essen & Joris Van Ouytsel, *Snapchat Streaks–How are These Forms of Gamified Interactions Associated with Problematic Smartphone Use and Fear of Missing Out Among Early Adolescents?*, 11 TELEMATICS & INFORMATICS REPS. 1, 1–6 (2023).

[24] Rixen et al., *The Loop and Reasons to Break It*, supra note 19.

[25] Jeroen Stragier et al., *Lifting the Veil on Smartphone Screen Time: The Role of Notifications and Specific App Activities in Explaining Session Length*, INT'L COMMC'N ASS'N (May 2021), https://www.researchgate.net/publication/352374227_Lifting_the_veil_on_smartphone_screen_time_The_role_of_notifications_and_specific_app_activities_in_explaining_session_length.

[26] Sandro Galea, et al., *Social Media and Adolescent Health*, NAT'L ACADS. (Mar. 25, 2024), https://nap.nationalacademies.org/initiative/committee-on-the-impact-of-social-media-on-adolescent-health.

How Algorithmic Design Creates Predictable Patterns of Harm

Harms associated with social media do not typically originate from one harmful post or viewpoint. They emerge from ranking logic, preference prediction cycles, and design features that shape what users encounter over time. These effects accumulate not because users seek out harmful material, but because the system repeatedly reinforces and amplifies types of content that generate strong engagement signals, regardless of accuracy or suitability. For young users, this can result in rapid exposure to content that is not developmentally appropriate.

For clarity, this white paper uses three descriptive categories to organize these design-driven harms:

1. **Exposure harms:** Being served content that makes young people feel ugly, unsafe, worthless, or scared, because the system keeps showing them violence, sexualization, thin-ideal imagery, extremist or hateful content, or contact from unknown adults they did not ask for, with documented associations with negative impacts on health and functioning.
2. **Escalation harms:** Getting pulled into spirals of increasingly extreme content—fitness to starvation, stress memes to self-harm tips, pranks to choking challenges—that can push youth toward physical injury, disordered eating, or suicidal ideation, with predictable harmful effects on emotional well-being at a population level.
3. **Engagement harms:** Being hooked into hours-long scrolling loops that interfere with sleep, undermine academic performance, fuel anxiety, and make kids feel out-of-control of their own behavior, even when they are frightened by how they feel, with real consequences for safety, development, and learning.

These risks stem from the way platforms curate and prioritize content across millions of posts and videos, not from isolated pieces of content. While each user's experience is unique, the design features create predictable patterns of exposure and harm when deployed at scale.

1. Exposure Harms: How the Design Shapes What Users Encounter

Engagement-based ranking systems surface material that generates strong behavioral signals, not material that is accurate, balanced, or developmentally appropriate. Because the system optimizes for predicted attention, youth may repeatedly encounter content they did not seek out, including violent, sexual, biased, or extreme material.

Youth describe feeds that showed them things they weren't looking for,[27-28] including self-harm videos and hate speech. In this design, the ranking system's logic, rather than individual speakers' choices, determines what appears repeatedly in a minor's feed. For purposes of legal analysis, this is significant: when harmful exposures result from the mechanical operation of ranking logic, rather than from the content of individual speech acts, those harms can be understood as foreseeable consequences of product design choices, rather than as inherent to the speech itself.

---

[27] Nancy Costello et al., *Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth From the Dangers of Social Media*, 49 AM. J. OF LAW & MED. 135, 135–72 (2023).
[28] Valdeep Gill et al., *Qualitative Research Project to Investigate the Impact of Online Harms on Children*, ECORYS (2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1167838/Online_Harms_Study_Final_report_updated_51222_updated_290623.pdf.

- **Algorithmic bias and discrimination.** Algorithms trained on historical data and user behavior can replicate and amplify unequal treatment across large numbers of users. Research shows that posts from different racial, ethnic, or gender groups do not appear with equal frequency or reach.[29-30] For youth developing identity and self-concept, repeated exposure to exclusionary content can undermine feelings of belonging, self-worth, and safety. Because these effects arise from how predictive systems rank and deliver content—not from any single post—they may be reasonably viewed as foreseeable outcomes of product design, relevant to assessing unreasonable risk of harm.

- **Misinformation and Echo Chambers.** Sensational or emotionally charged content generates strong engagement signals, leading the algorithmic system to elevate similar material. Over time, the system elevates and repeats such materials, creating *feedback loops* or *echo chambers* in political, health, or social contexts, including content that normalizes hate or extremism,[31,32,33] even without intentional searching. For example, a teen briefly pausing on a diet-related video may then see increasingly restrictive eating content for days; if a teen pauses on a stress meme, the system may interpret it as interest and begin recommending anxiety or self-harm content. This occurs because the system is designed to optimize for engagement—such as attention, clicks, and watch time—not for truth, accuracy, or user well-being.

- **Developmentally Inappropriate Exposures.** Recommendation systems heavily weight peer-group engagement. As a result, minors may be shown violent, sexual, or high-risk viral challenges that they never searched for.[34,35,36] Exposure occurs because the behavior of similar users signals "relevance," not because the youth sought out the material.

- **Data Extraction and Youth Privacy.** Ongoing collection of behavioral data from users is not incidental—it is the mechanism that fuels prediction and recommendation. Youth behavioral and psychological data are harvested to model future engagement, which in turn shapes what content they are shown and how long they remain engaged. Because minors cannot meaningfully understand or consent to this, the practice exposes them to both privacy violations and downstream harms linked to exposure and engagement.[37]

---

[29] Jack Brandy & Tomo Lazovich, *Exposure to Marginally Abusive Content on Twitter*, 17 PROC. OF THE INT'L AAAI CONF. ON WEB & SOC. MEDIA 24, 24–33 (2023).

[30] Fundacja Panoptykon et al., *Fixing Recommender Systems*, PANOPTYKON (Aug. 25, 2023), https://panoptykon.org/sites/default/files/2023-08/Panoptykon_ICCL_PvsBT_Fixing-recommender-systems_Aug%202023.pdf.

[31] WSJ Staff, *Inside TikTok's Algorithm: A WSJ Video Investigation*, WALL STREET J. (July 21, 2021), https://www.wsj.com/tech/tiktok-algorithm-video-investigation-11626877477.

[32] Milli et al., *Engagement, User Satisfaction*, supra note 13.

[33] Fabrizio Germano et al., *Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization*, SSRN (June 23, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5316506.

[34] Fatmaelzahraa Eltaher et al., *Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms*, ARXIV (May 16, 2025), https://arxiv.org/abs/2505.11160.

[35] Family Kids & Youth, *Understanding Pathways to Online Violent Content Among Children*, OFCOM (Mar. 2024), https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/keeping-children-safe-online/experiences-of-children/understanding-pathways-to-online-violent-content-among-children.pdf?v=368021.

[36] V. Jaynes, *Risky-By-Design: Recommendation Systems*, 5RIGHTS FOUND. (2022), https://riskybydesign.5rightsfoundation.com/recommendation-systems.

[37] Integrity Inst., *On Risk Assessment and Mitigation for Algorithmic Systems*, INTEGRITY INST. (Feb. 29, 2024), https://www.integrityinstitute.org/news/institute-news/risk-assessment.

2. Escalation Harms: How Design Intensifies What the User Sees

Engagement-based systems use real-time feedback loops that update feeds in response to micro-behaviors such as pauses, partial views, scroll speed, or rewatches. Even fleeting interaction with a social media post is interpreted as interest. Thus, feeds can escalate quickly from general-interest material to more intense, emotionally charged, or harmful content, sometimes within a single session.[38] [39] This escalation occurs automatically through ranking logic, not a platform's editorial judgment or user intent.

- **Algorithmic Rabbit Holes.** Small or accidental signals—pausing on a video, lingering for a few seconds, hovering over a post—are treated as positive engagement. The algorithmic system will then immediately increase the ranking weight of similar material, which can lead to quick escalation.[40] For example:
    - A teen who watches a few clips of school fights or "prank gone wrong" videos may quickly be shown more violent or humiliating content, including recordings of assaults or bullying, even if they were initially watching out of mere curiosity or discomfort.
    - Liking or replaying videos with cosmetic filters or makeover challenges can escalate into content promoting cosmetic procedures, filler injections, or "fixing flaws," reinforcing appearance anxiety and body modification pressures.

  Because emotion regulation and risk evaluation capacities are still developing in adolescents, they are especially susceptible to the swift escalation from benign content to more extreme or risky materials. What feels like harmless scrolling is treated by the system as meaningful feedback. That feedback, in turn, prompts the system to serve progressively more intense versions of the content the adolescent hesitated over or watched.
- **Viral Amplification.** When a trend generates strong engagement among a broad set of users, the algorithmic system assigns it higher relevance scores and distributes it more broadly—even to minors who never expressed interest. This mechanism has surfaced choking challenges, dangerous dares, and self-harm–related trends to youth through automated recommendations rather than deliberate user searching.[41-42] The amplification is a byproduct of engagement-based ranking, not a judgment about the content it elevates.
- **Case Examples.** Real-world incidents illustrate that algorithmically driven recommendation systems can lead to serious, sometimes fatal consequences when they amplify dangerous "challenge" content. Below are several documented instances.
    - In 2025, a trauma surgeon at a burn center in Northeast Ohio reported serious burns, and in some cases deaths, among children attempting "fire challenges" on social media. These stunts involve pouring alcohol on the body and igniting it or similar hazardous acts.[43]

---

[38] WSJ Staff, *Inside TikTok's Algorithm*, supra note 31.

[39] Mozilla Found.,*YouTube Regrets: A Crowdsourced Investigation into YouTube's Recommendation Algorithm*, MOZILLA FOUND. (July 2021), https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

[40] Dimitris Kalimeris et al., *Preference Amplification in Recommender Systems*, 1 PROC. OF THE 27TH ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 805, 805–15 (2021).

[41] OFF. OF THE SURGEON GEN., *Social Media and Youth Mental Health*, *supra* note 4, at 8

[42] Ofcom, *Tech Firms Must Tame Toxic Algorithms to Protect Children Online*, OFCOM (May 8, 2024), https://www.ofcom.org.uk/online-safety/protecting-children/tech-firms-must-tame-toxic-algorithms-to-protect-children-online.

[43] Mike Mason, *Northeast Ohio Kids Severely Burned, Even Killed, Attempting TikTok Fire Challenges*, CLEVELAND 19 NEWS (Nov. 4, 2025), https://www.cleveland19.com/2025/11/04/northeast-ohio-kids-severely-burned-even-killed-attempting-tiktok-fire-challenges/

- In 2025, prosecutors in Pennsylvania charged two teenagers after a "social-media challenge" stunt resulted in a 17-year-old's death and another young adult suffering permanent catastrophic injury. According to the official account, a folding table was tied to the back of a car — a "table-surfing" stunt filmed for social media — and the rider was thrown from the table when the vehicle collided with a parked car.[44]
- The so-called "blackout challenge" (also known as a form of the "choking game"), popularized on social media platforms, has been linked to the deaths of multiple minors worldwide. One report details a 12-year-old boy who died after allegedly attempting the challenge after exposure on a social-media platform.[45]
- The so-called "Tide Pod challenge" involved people, particularly youth, filming themselves ingesting Tide Pods, a well-known laundry detergent product designed to dissolve in washing machines. When consumed, the pods can cause serious harm, including neurological symptoms such as loss of consciousness and respiratory problems that may be fatal. [46] In response, Tide and the U.S. Consumer Product Safety Commission issued warnings on social media about the dangers of ingestion.[47]

3. Engagement Harms: Difficulty Disengaging and Health Consequences of Design

Engagement-driven design does more than keep users online longer—it also shapes emotional, cognitive, and behavioral patterns in ways that can negatively impact youth health.[48] Previously described design features (e.g., infinite scroll, autoplay, streak counters, and intermittent notifications) limit natural stopping points and stimulate repeated checking.[49] Because adolescents are in a developmental stage marked by heightened reward sensitivity, strong peer orientation, and ongoing maturation of self-regulation systems,[50] these features can produce predictable patterns of compulsive use and emotional strain.

- **Social Comparison and Body-Image Harms.** Algorithms often respond to brief interactions with appearance-related or fitness content by surfacing more of the same.[51] This can heighten social comparison, reinforce narrow beauty ideals, and increase body dissatisfaction—patterns

---

[44] Hayden Mitman, *Teens Charged in 'TikTok Challenge' Crashes That Led to Death, Permanent Injury*, NBC PHILADELPHIA (Feb. 11, 2025), https://www.nbcphiladelphia.com/news/local/live-teens-charged-in-tiktok-challenge-crashes-that-led-to-death-permanent-injury/4273962/.video

[45] Anne Marie D. Lee, *Child Deaths Blamed on TikTok "Blackout Challenge" Spark Outcry*, CBS NEWS (Aug. 9, 2021), https://www.cbsnews.com/news/tik-tok-blackout-challenge-child-deaths/

[46] Jamie Ducharme, *Here's How Common the Tide Pod Challenge Really Is*, TIME (Jan. 16, 2018), https://time.com/5104225/tide-pod-challenge/.

[47] Sarang Koushik, *Why Internet Craze the Tide Pod Challenge Is Dangerous, Potentially Deadly*, ABC NEWS (Jan. 17, 2018), https://abcnews.go.com/Health/internet-craze-tide-pod-challenge-dangerous-potentially-deadly/story?id=52379523.

[48] OFF. OF THE SURGEON GEN., *Social Media and Youth Mental Health*, *supra* note 4, at 8.

[49] Am. Psych. Ass'n, *Health Advisory on Social Media Use in Adolescence*, AM. PSYCH. ASS'N (May 9, 2023), https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use.pdf.

[50] Laurence Steinberg, *A Dual Systems Model of Adolescent Risk-Taking*, 52 DEVELOPMENTAL PSYCHOBIOLOGY 216, 216–24 (2010).

[51] Renee Engeln, *Compared to Facebook, Instagram Use Causes More Appearance Comparison & Lower Body Satisfaction in College Women*, 34 BODY IMAGE 38, 38–45 (2020).

well documented in adolescent development research.[52-53] For some youth, these exposure trajectories raise the risk of disordered eating behaviors. The mechanism is design-driven: Small signals feed a ranking system that elevates high-performing content (material that draws strong engagement such as likes, views, or watch time) much of which is idealized or extreme.

- **Compulsive Use and Reduced Self-Regulation.** Design elements that function like intermittent-reward systems—seamless feeds, unpredictable notifications, and rapid content transitions—make it difficult for youth to disengage from using social media.[54] Adolescents describe scrolling for hours at night, skipping sleep, and not being able to stop, which can interfere with sleep, attention, and daily functioning.[55-56] The difficulty stopping is therefore not simply a matter of preference or "willpower," but a foreseeable effect of product design on developing self-regulation.
- **Mental Health.** Because emotionally charged or negative content generates high engagement signals, algorithms may repeatedly surface material tied to stress, sadness, anger, or fear.[57] Over time, these patterns can intensify rumination, increase feelings of isolation, and contribute to elevated rates of anxiety and depressive symptoms among adolescents.[58-59] These outcomes reflect how ranking models interprets user engagement, treating emotional arousal as a sign of relevance, without distinguishing whether the user experiences the content as helpful or harmful.
- **Sleep Disruption and Constant Alerts.** Design features intended to drive re-engagement—such as "streak" counters that reward consecutive days of use and variable or unpredictable notifications that appear at irregular intervals—can interfere with adolescents' sleep patterns.[60] Many youth respond to these cues at night, checking apps when a streak is at risk or when a notification suggests something socially important has happened, leading to shortened sleep duration, reduced sleep quality, and next-day fatigue—all of which are strongly associated with worsening mood and attention problems.[61] These harms emerge because the product is built to pull users back into the social media feed, regardless of developmental needs.
- **Escalation to Disordered Eating or Self-Harm Content.** Recommendation systems can escalate quickly from general wellness or emotional-support posts to more extreme dieting, appearance-focused, or self-harm-related material—often within a short window and without any

---

[52] Ciera Elaine Kirkpatrick & Sungkyoung Lee, *Effects of Instagram Body Portrayals on Attention, State Body Dissatisfaction, and Appearance Management Behavioral Intention*, 38 HEALTH COMMC'N 1430, 1430–41 (2023).
[53] Hannah K Jarman, *Direct & Indirect Relationships Between Social Media Use and Body Satisfaction: A Prospective Study Among Adolescent Boys and Girls*, 26 NEW MEDIA & SOC'Y 292, 292–312 (2024).
[54] Rasan Burhan & Jalal Moradzadeh, *Neurotransmitter Dopamine (DA) and Its Role in the Dev. of Soc. Media Addiction*, 11 J. OF NEUROLOGY & NEUROPHYSIOLOGY 1, 1-2 (2020).
[55] OFF. OF THE SURGEON GEN., *Social Media & Youth Mental Health*, *supra* note 4, at 8.
[56] Jessica Leigh Hamilton & Woanjun Lee, *Associations Between Social Media, Bedtime Technology Use Rules, & Daytime Sleepiness Among Adolescents: Cross-Sectional Findings From a Nationally Representative Sample*, 8 JMIR MENTAL HEALTH 1, 1 (2021).
[57] Ric G. Steele et al*., Conceptualizing Digital Stress in Adolescents & Young Adults: Toward the Development of an Empirically Based Model*, 23 CLINICAL CHILD & FAMILY PSYCH. 15, 15–26 (2020).
[58] Sarah M Coyne et al., *Suicide Risk in Emerging Adulthood: Associations with Screen Time Over 10 Years*, 50 J. OF YOUTH & ADOLESCENCE 2324, 2324–2338 (2021).
[59] Amy Orben et al., *Windows of Developmental Sensitivity to Social Media*, 13 NATURE COMMC'N 1649, 1649 (2022).
[60] Am. Psych. Ass'n, *Health Advisory on Social Media Use in Adolescence*, *supra* note 49, at 16.
[61] Hamilton & Lee, *supra* note 56, at 16.

deliberate searching—because such content performs well among similar users.[62,63,64] Research links these trajectories to increases in body dissatisfaction, emotional dysregulation, and exposure to high-risk behaviors.[65] Again, the driver is algorithmic prediction—not a platform editorial decision about the value of the speech. The next section addresses what the emerging research shows about causal links between these design features and youth outcomes.

Causation Versus Correlation

Platforms often argue that any negative effects linked to social media are merely correlational: young people who are already struggling simply tend to use social media more, and the platform itself is not contributing to the problem. An increasing body of research, however, has moved beyond correlation and now tests causal effects.

In plain terms, **correlation** means two things happen at the same time but we cannot tell whether one preceded or caused the other. **Causation** means a feature or design choice directly contributes to a change in behavior or outcome.

The strongest studies that help determine causation are experimental studies and longitudinal studies. In experimental studies, participants are randomized usually into two groups, where one would be exposed to a set of social media design elements and the other would not be exposed, to examine how the two groups differ in symptoms or behaviors after the first group is exposed. Longitudinal studies follow participants over time and track how their social media use and exposure to different design elements on the platforms change over time and how these changes influence changes in symptoms and behaviors.[66] Studies of specific design features—such as infinite scroll, autoplay, notification cues, and engagement-based ranking—show that these elements cause predictable changes in behavior. They reliably increase users' time spent online, prompt them to repeatedly check their social media accounts throughout the day (and often late at night), and elevate content that generates strong reactions, including harmful or extreme material.[67] Evidence connecting design to longer-term mental health outcomes is still developing, but the short-term, design-driven effects on behavior are well established. Longitudinal research indicates that appearance-focused or emotionally charged feeds are

---

[62] Ctr. for Countering Digital Hate, *Deadly by Design: TikTok's Recommendation Algorithm and Eating Disorder Content*, CTR. FOR COUNTERING DIGITAL HATE (Dec. 15, 2022), https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf.

[63] Ctr. for Countering Digital Hate, *Digital Hate. YouTube's Anorexia Algorithm: How YouTube Recommends Eating Disorder Videos to Young Girls*, CTR. FOR COUNTERING DIGITAL HATE (Dec. 10, 2024), https://counterhate.com/research/youtube-anorexia-algorithm/

[64] Tawnell D. Hobbs et al., *The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos*, WALL STREET J. (Dec. 17, 2021), https://www.wsj.com/tech/how-tiktok-inundates-teens-with-eating-disorder-videos-11639754848.

[65] Rubinia Celeste Bonfanti et al., *The Association Between Social Comparison in Social Media, Body Image Concerns & Eating Disorder Symptoms: A Systematic Review & Meta-Analysis*, 52 BODY IMAGE 101816, 101841 (2025).

[66] TIMOTHY L. LASH, TYLER J. VANDERWEELE, SEBASTIEN HANEUSE & KENNETH J. ROTHMAN, MODERN EPIDEMIOLOGY (4th ed. 2020).

[67] Kaitlyn Regehr et al., *Safer Scrolling: How Algorithms Popularise and Gamify Online Hate and Misogyny for Young People*, ASS'N OF SCH. & COLL. LEADERS (2024), https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf.

associated over time with worsened body-image outcomes and increased anxiety or depressive symptoms among youth.[68,69,70,71,72,73]

The legal analysis that follows does not depend on resolving all scientific debates about the magnitude of harm to youth. It focuses on whether particular platform design choices create foreseeable risks of harm, distinct from harms caused by expressive online content, and may therefore be legally regulated.

Implications for Civil Law

The risks outlined in the previous section do not arise from individual pieces of content but from the design and operation of engagement-based algorithmic systems. Understanding these predictable patterns of exposure, escalation of content, and harm is essential to evaluating how algorithmic design intersects with questions of foreseeability of harm, product safety, and consumer protection. The next section applies these concepts to civil law frameworks, including product liability, negligence, and consumer protection theories.

Intersection with Civil Law

Despite the apparent risks algorithms pose for youth, formulating laws that can effectively tackle these harms while remaining within the permissible confines of the Constitution has proven to be challenging. Courts in the United States have little legal precedent to guide them when assessing laws that moderate algorithms. This dearth of legal guidance poses a formidable difficulty for courts.

First and foremost, courts must distinguish whether newly enacted laws unconstitutionally target content (expression posted on a social media platform) or whether they target the product design of algorithms in social media. Laws that moderate content–expressive speech–are unlikely to survive constitutional scrutiny under the First Amendment. This may be the most substantial hurdle for legislation focused on the regulation of social media algorithms. Section 230 of the Communications Decency Act ("Section 230")[74] also provides a significant obstacle. Section 230 immunizes social media platforms and other internet service providers from being liable for third-party content that is posted on that company's platform whether it is defamatory, threatening violence, or otherwise illegal. Algorithms do not create content, but moderate the third-party content that is posted on social media platforms. Thus, it may appear that social media platforms should be immunized from liability for their algorithms that merely promote harmful third-party content. Yet key questions remain: Do all algorithms operate the same way? Are the social media companies that use them immune from liability regardless of what kind of algorithm they use?

---

[68] Orben et al., *supra* note 59, at 17.
[69] Guilia Conte et al., *Scrolling Through Adolescence: A Systematic Review of the Impact of TikTok on Adolescent Mental Health*, 34 EUR. CHILD & ADOLESCENT PSYCHIATRY 1511, 1511–27 (2025).
[70] Jarman, *supra* note 53, at 16.
[71] Coyne et al., *supra* note 58, at 17.
[72] Russell Viner et al., *Roles of Cyberbullying, Sleep, and Physical Activity in Mediating the Effects of Social Media Use on Mental Health and Wellbeing Among Young People in England: A Secondary Analysis of Longitudinal Data*, 3 LANCET CHILD & ADOLESCENT HEALTH 685 (2019).
[73] ANNE J. MAHEUX ET AL., *Longitudinal Associations Between Appearance-Related Social Media Consciousness and Adolescents' Depressive Symptoms*, 94 J. ADOLESCENCE 264 (2022).
[74] *See* 47 U.S.C. § 230.

Legal Gaps and Current Challenges

Several obstacles stand in the way of legislation that targets the regulation of social media algorithms. The most significant obstacles are the First Amendment and Section 230.

The First Amendment

The First Amendment protects the speech and expression of those who host internet sites, those who post on internet sites, and people's ability to access others' expressions on internet sites. As a result, some laws trying to regulate social media have been struck down because they were content-based, or, in other words, focused on limiting expression of social media platforms and those who post on them. Other laws trying to measure and mitigate the harm caused by social media companies have been struck down because they compelled speech by the social media platforms. For example, in *Alario v. Knudsen*[75], a Montana law that banned TikTok statewide was struck down by Montana's federal district court as an unconstitutional content-based restriction. The law singled out one platform, depriving users of their preferred mode of expression and infringing on TikTok's First Amendment right to select, arrange, and curate third-party speech.[76] The court held the law ultimately targeted speech itself rather than regulating TikTok as a product.[77]

NetChoice LLC v. Reyes[78] provides a helpful illustration. There, Utah enacted the Minor Protection in Social Media Act which required platforms to implement age-verification systems and impose special restrictions on accounts belonging to minors.[79] Utah's asserted interest was the protection of the well-being and privacy of Utah minors who use social media.[80] NetChoice, representing social media companies, and several individual users sued Utah.[81] Utah's federal district court struck down the Act, ruling that it violated the First Amendment because it was a content-based restriction.[82] Utah's interest in protecting minors was not compelling, and the Act was not narrowly tailored.[83] The law was drafted too broadly because it directly interfered with platforms' expressive editorial choices in organizing "their own distinctive compilations of expression."[84]

Similarly, in *NetChoice v. Bonta* ("*Bonta #1*")[85], the court struck down part of California's Age Appropriate Design Code. The Code contained a Data Protection Impact Assessment requirement for social media companies.[86] The Assessment required social media platforms to identify how their platform designs or data collection practices could harm children.[87] Platforms were then required to file a private report of those findings with the state and develop a plan to mitigate the harm.[88] NetChoice, representing social media companies, sued California, claiming that the First Amendment speech rights

---

[75] 704 F. Supp. 3d 1061 (D. Mont. 2023).
[76] *Id.* at 1074.
[77] *Id.*
[78] 748 F. Supp. 3d 1105 (D. Utah 2024); *see also* Brief of Plaintiff-Appellant at 41–42, *Florida v. Snap, Inc.*, No. 25-12814 (11th Cir. Sep. 24, 2025) (arguing that a Florida bill HB3, which prohibits Snap from contracting with children under 14 years old to make Snapchat accounts and prohibits Snap from contracting with children without parental consent, is consistent with the First Amendment because the bill targets nonexpressive activity that does implicate free speech rights).
[79] *Id.* at 1111.
[80] *Id.* at 1113.
[81] *Id.* at 1111-12.
[82] *Id.* at 1119-22.
[83] *Id.*
[84] *Id.* at 1120.
[85] *See NetChoice, LLC v. Bonta*, 113 F.4th 1101 (9th Cir. 2024).
[86] *Id.* at 1109-11.
[87] *Id.*
[88] *Id.* at 1110.

of the social media companies were violated by the entire act.[89] The court struck down the Data Protection Impact Assessment requirement, ruling that it compelled speech by forcing platforms to identify how their platform design or data collection could harm children.[90] The court noted less restrictive alternatives existed such as educating users and parents about the dangers of social media, relying on existing criminal laws, and offering voluntary content filter blockers.[91] In sum, these cases are helpful in understanding how the First Amendment has been used to successfully oppose laws that regulate social media platforms.

Section 230

Section 230 also poses a significant obstacle for social media legislation and civil lawsuits. Because it immunizes social media platforms and other internet service providers from being liable for third-party content that is posted on a company's platform,[92] Section 230 has repeatedly prevented regulation of social media and been used successfully as a defense in lawsuits. Analysis of laws under Section 230 frequently focuses on whether the social media platform is engaged in an editorial or publication function. An editorial function involves the traditional choice of a publisher to publish, withdraw, postpone, or alter content.[93]

Jones v. Dirty World Entertainment Recordings[94] provides a helpful illustration of how Section 230 is applied to online platforms. There, a teacher and former NFL cheerleader sued a gossip website and its operator for defamation after users posted false statements about her.[95] The court ruled that Section 230 immunity applied to the website because a website is immune as long as it does not materially contribute to the illegality of third-party content.[96] Mere encouragement, commentary, or failure by the website operator to remove posts is not enough to strip immunity.[97] The gossip website exercised traditional editorial functions (selecting and publishing third-party submissions) but did not create or materially alter the defamatory nature of the posts.[98]

Estate of Bride v. Yolo Technologies Inc.[99] better illustrates how Section 230 has been used more recently with social media. There, the estate of a teenager who died by suicide after being tormented by anonymous posters sued Yolo and LMK, apps integrated within Snapchat, and Snapchat itself.[100] The estate alleged the applications' anonymous messaging features encouraged bullying and harassment.[101] The Ninth Circuit dismissed the case under Section 230, finding that anonymous messaging design was the same thing as the decision to publish third-party speech without disclosing a speaker's identity, which falls squarely within a traditional publisher function.[102] The court also rejected the plaintiff's products liability argument, finding that allowing, transmitting, and failing to remove harmful messages was still an editorial function by the app.[103] Overall, Section 230 provides a

---

[89] *Id.* at 1112.
[90] *Id.* at 1117.
[91] *Id.* at 1121.
[92] *See* 47 U.S.C. § 230.
[93] *E.g.*, *Jones v. Dirty World Ent. Recordings*, 755 F.3d 398, 416 (6th Cir. 2014).
[94] *Id.*
[95] *Id.* at 401-05.
[96] *Id.* at 410.
[97] *Id.* at 414-15.
[98] *Id.* at 415-16.
[99] 112 F.4th 1168 (9th Cir. 2024).
[100] *Id.* at 1173-74.
[101] *Id.* at 11-74-75.
[102] *Id.* at 1180-81.
[103] *Id.* at 1179-81.

significant obstacle for plaintiffs in civil lawsuits and legislation that seeks to regulate social media because it immunizes platforms from illegal third-party content.

Regulating Algorithmic Design Outside the United States

Outside of the United States, many countries treat engagement-based algorithmic systems as design features subject to safety regulation. For instance, the European Union Digital Services Act requires platforms to assess and mitigate systemic risks created by algorithms, particularly for minors.[104] The UK Online Safety Act similarly mandates that platforms address risks arising from recommendation systems and virality.[105] These European laws explicitly recognize recommendation algorithms as product design rather than as content itself, and require social media platforms to measure and reduce associated harms. Free speech protections in Europe are far less stringent than in the United States. In the U.S., the First Amendment and Section 230 complicate direct regulation, meaning questions about design-based risk often arise in civil litigation. Courts must therefore determine whether a platform's challenged practice reflects protected editorial judgment, or a commercial product design that may create an unreasonable risk of harm.

Australia has taken a more prescriptive approach: as of December 2025, a federal law requires major social media platforms to take reasonable steps to prevent users under 16 from holding accounts, reflecting deep concern about the risks social media design poses to minors.[106]

New Developments

Multiple courts in the United States have begun to differentiate between what is protected expression on social media versus the structural product design of a social media platform, which is a design that is solely implemented to generate revenue for the platform. Instead of analyzing platform design choices as content-based decisions that implicate the First Amendment or Section 230, courts are beginning to apply products liability law that examines harmful design of a product and not expression.

First Amendment

The Supreme Court considered in *Moody v. NetChoice, LLC*, 603 U.S. 707 (2024) whether two Texas and Florida laws that prohibited social media platforms from engaging in certain forms of content moderation violated the First Amendment.[107] Examples of prohibited content moderation under the Florida law included willfully "deplatforming" or otherwise removing political candidates from the social media site.[108] In Texas, social media platforms could be penalized for engaging in any viewpoint-based banning or de-platforming.[109] NetChoice argued that these laws interfered with the companies' First Amendment right to editorial discretion.[110] The Court reasoned that these regulations that relate to content were not permitted under the First Amendment because it would cause social media platforms to display content that they might not agree with.[111]

---

[104] Eur. Comm'n, *The Digital Services* Act, EUR. COMM'N (Dec. 12, 2025), https://digital-strategy.ec.europa.eu/en/policies/digital-services-act.
[105] U.K. Dep't for Sci., Innovation & Tech., *Online Safety Act: Explainer*, U.K. GOV'T (Apr. 24, 2025), https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer.
[106] Australian eSafety Comm'r, *Social Media Age Restrictions in Australia.* AUSTRALIAN GOV'T (2025), https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions.
[107] *See Moody v. NetChoice, LLC*, 603 U.S. 707, 717 (2024).
[108] *See id.* at 720.
[109] *See id.* at 710.
[110] *See id.* at 722.
[111] *See id.* at 717.

However, in *Moody*, Justice Kagan stated that platforms exercising authority "to remove, label or demote messages they disfavor" was not the same as engagement-based "algorithms [that] respond solely to how users act online–giving them the content they appear to want, without any regard to independent content standards" such as the social media platform's own Community Standards or Community Guidelines.[112] The Court acknowledged that social media companies could potentially be held responsible for engagement-based algorithms that do not focus on content, but instead prioritize user interaction and time spent on the app to generate the highest amount of advertising revenue possible.[113]

In her concurring opinion in *Moody*, Justice Barrett asked "what if a platform's algorithm just presents automatically to each user whatever the algorithm thinks the user will like – e.g., content similar to posts with which the user previously engaged? The First Amendment implications of the Florida and Texas laws might be different for that kind of algorithm."[114] Justice Alito also stated that algorithms are not always inherently expressive, suggesting that certain types of algorithms may not have anything to do with regulating content and thus would not get First Amendment protections.[115]

In *NetChoice v. Bonta*, 152 F.4th 1002 (9th Cir. 2025) ("Bonta #2"), the Ninth Circuit Court of Appeals addressed an appeal regarding California's Protecting Our Kids from Social Media Addiction Act and whether it violated the First Amendment.[116] The Act aimed to reduce risk of harm to children and teenagers posed by algorithmic delivery of content on online platforms.[117] One of the ways it did this was through restricting minors' access to algorithmic feeds by prohibiting personalized-feeds that would display posts based on user information (i.e. age, gender, location, scrolling history).[118] In addition, the act required default-settings for minors on social media platforms. Two default-setting provisions were at issue.[119]

The first default-setting regulated the default private mode that would set a child's account to "private" (allowing only the child's friends and followers to see the child's account) and allowed it to be set to "public" (anyone could see the child's account) only with parental consent.[120] The court held that this default private mode requirement was allowable because it did not reflect content preferences and was narrowly tailored to protect minors' mental health.[121] Setting accounts to private did not require the company to regulate or censor messages to minors.[122] Young users were still able to view and follow whoever they would like, and could display their own posts to those that follow them.[123]

The second default-setting required a like-modified mode that limited a child's ability to view the number of likes or other forms of feedback given to posts within an addictive feed.[124] It required the like-

---

[112] *See id.* at 736 n.5
[113] *See id.*
[114] *Id.* at 746 (Barrett, J., concurring).
[115] *Id.* at 766–767 (Alito, J., concurring).
[116] *See NetChoice, LLC v Bonta*, 152 F.4th 1002, 1009 (9th Cir. 2025).
[117] *See id.*
[118] *See id.* at 1010.
[119] *See id.*
[120] *See id.* at 1010–11.
[121] *See id.* at 1017.
[122] *See id.* at 1016.
[123] *See id.*
[124] *See* 2024 Cal. Stats. Ch. 321, SB 976, 2023-2024 Reg. Sess. (Cal. 2024).

modified mode to be set to "on" by default unless bypassed by parental consent.[125] By being set to "on," a child user would not be able to see the number of likes on other users' posts in addictive social media feeds.[126] The court held that regulation of "like" counts was inherently content based because "like counts are speech with particular content" and forcing a social media company to restrict a minor from viewing likes would require the company to restrict the message of those likes, i.e. expressing popularity, trending topics, etc.[127] Essentially, requiring a like-modified mode would restrict social media platforms from showing posts based on the speech expressed, thus violating the First Amendment.[128]

In reaching its decision, the Ninth Circuit relied on the reasoning in *Moody*, highlighting again the difference between certain types of algorithms.[129] The court recognized that "an algorithm that promotes a platform's own message to users is likely to be protected speech. Such an algorithm, after all, is not unlike traditional media curated by human editors."[130] However, the court also acknowledged that "on the other hand, an algorithm that responds solely to how users act online, merely giving them the content they appear to want, probably is *not expressive*.[131] Personalized algorithms might express a platform's own editorial messages to users (and therefore would be expressive), or they might reflect users' revealed preferences to them (non-expressive)."[132]

As an increasing number of states begin to enact legislation to protect users from algorithmic harm, courts are having to differentiate between different types of algorithms, and as shown by cases in the Supreme Court and Ninth Circuit, judges are beginning to do just that. Engagement-based or recommendation-based algorithms are being increasingly distinguished as product design choices that harm users to generate advertisement revenue rather than protected expression under the First Amendment.

Section 230

In *Lemmon v. Snap*, 995 F.3d 1085 (9th Cir. 2021), two teenagers and a 20-year old man were killed in a 123 mph car crash after using Snapchat's speed filter that would display a filter that indicated how fast the teenagers were moving in real time.[133] At the time of the accident, young Snapchat users believed that Snapchat would reward them for recording the speed filter at over 100 mph with recognition on the platform in the form of badges or points.[134] The parents brought a negligent product design claim against Snap for the speed filter's design that encouraged young users to speed.[135]

Snap attempted to invoke Section 230 immunity under the Communications Decency Act for the parents' claim.[136] The court held that Snap could not invoke Section 230 immunity because the negligent design claim "treats Snap as a products manufacturer, accusing it of negligently designing a

---

[125] *See id.*
[126] *See id.*
[127] *See id.* at 1016.
[128] *Id.*
[129] *See id.* at 1014.
[130] *See id.* at 1014.
[131] *See id.*
[132] *Id.* (citation modified).
[133] *See Lemmon v. Snap*, 995 F.3d 1085, 1087 (9th Cir. 2021). *See generally* Complaint at 2–4, Platkin v. Tiktok, ESX-C-000228-24 (N.J. Super. Ct. Ch. Div. Oct.8, 2024) (describing how Tiktok harms young users in New Jersey by providing addictive feeds to generate profit).
[134] *See Lemmon*, 995 F.3d at 1088.
[135] *See id.* at 1090.
[136] *See id.*

product with a defect" rather than "hold[ing] Snap liable for its conduct as a publisher or speaker."[137] Because the negligent design claim was concerned with the design of Snap's speed filter and could be resolved by designing a safer product rather than editing, monitoring, or removing user generated content, Snap was not granted immunity under Section 230.[138] The parents of the teenagers won the case.[139]

In *TikTok v Anderson,* 116 F.4th 180 (3rd Cir. 2024), the Third Circuit Court of Appeals addressed the "Blackout Challenge," a dangerous social media trend posted by third parties encouraging children to self-asphyxiate.[140] Videos of the challenge were delivered to young users by TikTok's engagement-based algorithm.[141] It resulted in the death of a ten-year-old Pennsylvania girl, and her mother sued TikTok.[142]

The Third Circuit decided that TikTok's algorithm was the platform's own first-party speech.[143] Because the algorithm actively decided what third-party speech would be included or excluded from users' pages, TikTok's expressive choices were deemed to be speech rather than neutral hosting of third-party expression.[144] Section 230 immunity did not apply, because that protection only shields platforms from liability for independent third-party speech, not for speech or expression created by the platform itself.[145] As an increasing number of courts have begun to differentiate between what is protected expression on social media versus a social media platform's structural design, more plaintiffs may be successful in challenging First Amendment and Section 230 protections of social media platforms that they claim have caused harm.

Potential Solutions

Courts have already begun to adapt to the age of social media. As outlined above, recent developments in the law may provide viable routes to address harms caused by social media companies and survive First Amendment and Section 230 challenges. For example, laws that focus on the harm of social media *product design* rather than the harmful *speech* posted on the platforms may withstand First Amendment or Section 230 challenges. The negligent design claim in *Lemmon v. Snap*[146] provides an illustrative example because that claim focused on the negligent design of the Snapchat app rather than objecting to any kind of third-party or user expression. Interpreting laws that regulate social media as regulations of product design could also be applied to non-content related algorithms. For example, engagement-based algorithms are used to measure and display whatever the algorithm thinks a user is most likely to engage in for long periods of time, rather than regulating content itself. Algorithms that are designed to maximize user engagement rather than promote specific content may not implicate the First Amendment or Section 230.

Requiring social media companies to do algorithm risk audits would also provide a viable path to constitutional regulation of social media algorithms. These audits would be conducted by independent

---

[137] *See id.* at 1092.
[138] *See id.* at 1092–93.
[139] *See id.* at 1095.
[140] *See TikTok v Anderson,* 116 F.4th 180, 181 (3rd Cir. 2024).
[141] *See id.* at 182.
[142] *Id.*
[143] *Id.* at 184.
[144] *Id.*
[145] *Id.*
[146] 995 F.3d at 1090.

third-party auditors. The audits would analyze anonymous user data that platforms already collect, to measure actual harms—such as how quickly an engagement-based algorithm shifts a user's feed from general-interest content to extreme content. Such measurements can be made for specific vulnerable groups of users, like youth or young women. A regulatory framework could establish limits for the acceptable level of harm measured in these audits, and consequences should a platform exceed the limits. Importantly, these audits would not regulate the speech and expression of third-party content on the platforms. Rather, platforms could limit the harms arising from their product designs however they see fit.

Lawmakers in several jurisdictions, including Massachusetts, have begun proposing legislation that adopts this model. In Massachusetts, Senate Bill S.51, *An Act Relative to Social Media, Algorithm Accountability, and Transparency*, would require platforms to undergo independent algorithm risk audits and publicly disclose the results. The bill is currently before the Senate Committee on Ways and Means. If enacted, it would create one of the first state-level regulatory frameworks focused specifically on assessing and mitigating harms arising from engagement-based algorithmic systems.

The results of the audits would be given to law enforcement authorities, including state attorneys general, who could use evidence of the harm to users to bring legal action against social media companies. These legal actions could include claims of unfair business practices and deceptive advertising. However, if a social media company should challenge claims brought against them, courts must be able to differentiate what is an unconstitutional content-based restriction versus regulation of a harmful product design.

Recommendations for Judges

General Educational Requirements

As discussed in *Moody v. NetChoice* and *Bonta 2*, different types of algorithms may influence whether or not Section 230 or the First Amendment are implicated in algorithm-regulating legislation.[147] Social media platforms use a variety of different algorithms. Thus, being able to differentiate between different types of algorithms is vital to understanding how they impact online users. The proposed Federal Rule of Evidence 707 makes this particularly timely for the judiciary. If the proposed rule is adopted, judges will be required to make admissibility determinations on machine-generated evidence that is offered by non-expert witnesses.[148] Such evidence could potentially include algorithms and data collected from algorithms. As a result, if the rule is passed, it is vital for judges to have background knowledge of such technology to be able to make accurate admissibility determinations.[149]

Algorithms that would implicate Section 230 or First Amendment Protection
Both search-based algorithms and chronological algorithms would implicate Section 230 or First Amendment protection because they support the editorial function of a social media platform in

[147] *See generally Moody v. NetChoice, LLC*, 603 U.S. 707 (2024); *NetChoice, LLC v Bonta*, 152 F.4th 1002 (9th Cir. 2025).
[148] *See* Fed. R. Evid. 707 (proposed June 10, 2025) (proposed rule addressing AI and other machine-generated outputs, directing judges to evaluate such evidence under the Rule 702(a)-(d) reliability factors when it is offered through the testimony of a non-expert. The rule states: "*When machine-generated evidence is offered without an expert witness and would be subject to Rule 702 if testified to by a witness, the court may admit the evidence only if it satisfies the requirements of Rule 702 (a)-(d). This rule does not apply to the output of simple scientific instruments*").
[149] If proposed FRE 707 goes into effect, it will be extremely important for the judiciary to be educated about it, although there is concern that the proposed rule could be premature as AI's development and use in the legal field and litigation as of this writing, March 2026, is yet to be fully understood.

compiling content for users, including content that users have requested to see.[150] With search-based algorithms, users search for what they are looking for, and the algorithm compiles results based on search terms posited by the user. Search-based algorithms employ technological advancements like Google search and search bars on social media platforms. Another example of an algorithm that implicates Section 230 or First Amendment protections would be chronological algorithms. These algorithms show newest content first and display other content based on when it was posted with no regard to what the content actually conveys. Chronological algorithms are used for features like 'Most Recent' feeds on social media platforms including Threads, Facebook, and Instagram, which display posts in the order they are created or shared—showing the newest content first—rather than re-ranking them based on predicted engagement or user-specific interests.

Algorithms that would not implicate Section 230 or First Amendment protection

In contrast, there are algorithms that would not implicate Section 230 or First Amendment protection because their function is not for editorial purposes but rather to maximize platform advertising revenue by keeping users' eyes on the screen.[151] Engagement-based and recommendation-based algorithms display what the social media platform thinks will maximize likes, shares, and time spent on the platform by users. They also display information that has received substantial viewing time and likes from other users. These algorithms predict what to show users based on what has kept similar people on the app, when considering criteria like age or gender of the users. An example of this would be Tiktok's For You Page. Engagement-based algorithms also can be structured to show users content similar to what they have liked or interacted with in the past. For example, TikTok's "For You" feed and Instagram's "Suggested for you" or "People like you also liked" recommendations continually serve new videos and posts that resemble a user's prior engagement. The content shown to users is often accompanied by advertisements. The main goal of engagement-based algorithms, therefore, is not to display expressive speech, but to keep users' eyes on the computer screen, which often contain advertisements and thereby earn more money in ad revenues for the social media company.

Judicial Roles

It is well within the purview of courts to evaluate the dangers caused by algorithms when adjudicating a case that claims harm was caused by a social media platform or assessing the constitutionality of a law attempting to regulate social media. Courts can evaluate algorithm evidence produced by uninterested third-party algorithm risk audits, question technical experts, and assess causation versus correlation of harm. However, to make more informed decisions on deeply technological questions concerning algorithms, judges could ask the parties to brief specific topics, such as how the technology works, how to define the technology, what the purpose of the legislation is, or whether the legislation targets speech and expression versus harmful product design.

Moreover, judicial clerks could aid in the research of algorithms and assist judges in navigating changing trends and technological advances in social media. Additionally, judges could seek out amicus briefs to understand the technology at issue and its effects. Engineers, technology experts (including those specializing in the creation and operation of algorithms), public health experts, legal experts, and more are willing to write amicus briefs to help courts parse dense technological subjects and nuanced constitutional issues. Topics for amicus briefs could include basic algorithm functions, possible guidelines for evaluating algorithms, potential of algorithm auditing as a practice, or a more in-depth analysis of the technology underpinning social media.

---

[150] *See Moody v. NetChoice, LLC,* 603 U.S. at 746 (Barrett, J., concurring).
[151] *See NetChoice LLC v. Reyes,* 748 F. Supp. 3d at 1120.

Conclusion

This white paper has argued that the central legal question in many social media cases is not simply what was expressed online, but how the platform was designed to deliver and amplify that speech. Engagement-based algorithms now function as the operating system of modern social media. They collect behavioral and demographic data, predict what will most effectively capture attention, and continuously reorganize each user's feed to maximize time on platform and advertising revenue. For minors, whose developmental stage makes them particularly sensitive to reward cues, social comparison, and peer approval, these designs predictably generate exposure, escalation, and engagement harms.

The evidence summarized here shows that these harms arise from structural features of product design, not from isolated pieces of user content. Infinite scroll, autoplay, notification patterns, streaks, and personalized ranking systems operate regardless of viewpoint or message. They are commercial choices about how to shape an environment so that young people stay online longer, check their feeds more often, and provide more data. When these choices result in foreseeable risks to safety, health, and development, they fall squarely within the kinds of product design issues that courts routinely evaluate in other product liability contexts.

Recent case law demonstrates that courts are beginning to draw a line between algorithms that function as expressive editorial tools and algorithms that operate as non-expressive engagement systems. That distinction matters for both the First Amendment and Section 230. Search-based and chronological algorithms that compile or order content in ways that reflect user requests or platform editorial judgment generally warrant traditional speech protections. Engagement-based and recommendation-based systems that operate solely to maximize user interaction and advertising revenue can be analyzed as product design, particularly when they drive harmful outcomes for minors regardless of any particular viewpoint.

For judges, this emerging distinction creates both an opportunity and a responsibility. Courts can uphold core free expression principles while still recognizing that certain platform features may be defectively designed, deceptively marketed, or unreasonably dangerous, especially for youth. Doing so requires a baseline understanding of how algorithms function, active engagement with expert testimony and independent algorithm audits, and careful attention to whether a challenged feature is expressing speech or structuring a product environment.

The paper has outlined practical tools to support that work. Independent algorithm risk audits can document design-driven harms without compelling platforms to adopt or suppress specific viewpoints. Product liability, negligence, and consumer protection theories can be used to address unsafe design choices by social media companies where safer alternatives exist. Judicial education, technologically literate clerks, and targeted amicus briefs can help courts interpret complex technical evidence and distinguish between product designs that correlate with harm or cause harm.

Looking forward, the same questions will arise in new forms as generative artificial intelligence, adaptive recommender systems, and other emerging technologies become embedded in online platforms and everyday life. The core principles set out here remain relevant: the judiciary must distinguish content from design, ask what the technological system is optimized to do, examine whether harms are reasonably foreseeable, and identify whether a particular feature operates as protected expression or as a commercial product choice.

If courts adopt this structured approach, they can play a meaningful role in ensuring that powerful algorithmic online systems, especially those shaping the lives of minors, evolve in ways that are consistent with both constitutional protections and basic expectations of safety, fairness, and transparency.

Appendix: Glossary of Key Terms

Advertising-supported social media platform. An online service that is free for users to join and use, but earns revenue by selling advertising. The platform's business model depends on keeping users' attention and collecting behavioral data to target ads.

Algorithm. A set of instructions a computer follows to perform a task or make a decision. On social media, algorithms decide which posts to show, in what order, for how long, and to which users.

Algorithmic design / product design. The way a platform's technical systems are built and configured, including how algorithms rank, recommend, and display content. In this paper, "algorithmic design" is treated as a product design choice, separate from the content users post.

Algorithmic rabbit hole. A pattern in which small or accidental signals of interest (for example, pausing on a video for a few seconds) lead the system to show more of the same type of content, escalating quickly to more extreme or harmful material without deliberate searching.

Algorithmic recommendation system. A system that predicts what a user is most likely to watch, click, or share, and then recommends or pushes that content into the user's feed. It operates on the basis of behavioral and demographic data rather than individual editorial review.

Algorithm risk audit. An independent, structured evaluation of how a platform's algorithms and design features affect users, including whether they create foreseeable patterns of harm. Audits focus on system behavior and outcomes, not on judging the truth or value of particular messages.

Behavioral data. Information about how users interact with a platform, such as what they click, how long they pause on a post, what they share, and when they log in. This data is used to train and update algorithms and to target advertising.

Chronological feed. A feed that displays posts in time order, usually with the newest content first, without reordering based on predicted engagement. It is closer to a user-initiated list than to an engagement-optimized recommendation system.

Causation versus correlation. A distinction between evidence showing that one factor actually contributes to a particular outcome (causation) and evidence showing that two things happen together without proving that one causes the other (correlation). In this context, it describes whether design features reliably change user behavior rather than simply appearing alongside harms.

Default setting. The configuration that applies automatically unless the user or a parent actively changes it. For example, setting a minor's account to "private" by default or turning off visible "like" counts unless a parent opts out.

Design (platform design). The structural choices that determine how content is organized, displayed, ranked, recommended, or withheld. Examples include feed-ranking algorithms, infinite scroll, autoplay, and notification patterns. Design affects which speech is most visible without changing the underlying words or images.

Engagement. A general term for how users interact with content, including clicks, likes, comments, shares, watch time, and pauses. Platforms treat engagement as a key measure of what content is "successful."

Engagement-based algorithm / engagement-based ranking. An algorithm that orders and recommends content based on predictions about what will maximize user engagement and time on platform, rather than on accuracy, quality, or developmental suitability.

Engagement-based feed / personalized feed. The stream of content a platform curates for each user

based on predicted engagement. Two users may see very different feeds, even if they follow similar accounts, because the system is personalized to their past behavior.

Engagement harms. Harms that arise when design features make it difficult for users, especially minors, to disengage. Examples include hours-long scrolling, sleep disruption, impaired attention, and feelings of being unable to stop using the platform.

Escalation harms. Harms that occur when algorithms quickly intensify content, moving from general or benign material to more extreme, risky, or harmful content. The escalation is triggered by micro-behaviors and ranking logic rather than deliberate searching for extreme material.

Exposure harms. Harms that result from being repeatedly shown violent, sexual, appearance-based, or otherwise harmful content that the user did not actively seek out. The harm flows from repeated exposure produced by ranking and recommendation systems.

First Amendment (in this context). The constitutional protection for speech and expressive conduct, including the expressive editorial choices of platforms that select, arrange, and present content. The paper focuses on when platform behavior is expressive, and when it is more like non-expressive product design.

Generative artificial intelligence (generative AI). Computer systems that can create new text, images, audio, or video in response to prompts. The paper notes that similar questions about content versus design will arise as generative AI tools are used in platforms.

Infinite scroll. A design feature in which the feed continuously loads new content as the user nears the bottom of the screen, without requiring a click to move to the next page. This removes natural stopping points and extends time on platform.

Machine-learning model. A type of algorithm that “learns” patterns from data rather than following a simple, fixed rule. On social media, machine-learning models analyze behavioral and demographic data to predict what content will generate engagement.

Notification cues / variable notifications. Alerts sent by a platform to draw users back to the app, often at unpredictable intervals and with mixed types of rewards (for example, a like, a comment, or a “streak” warning). This variability encourages repeated checking.

Optimization loop. A feedback process in which the system tests content, measures how users respond, and then adjusts what it shows next in order to improve a goal such as engagement. The loop runs continuously and reshapes each user’s feed over time.

Product liability (in this context). A legal framework that treats certain aspects of a platform, including algorithmic design features, as product design choices that can be defective or unreasonably dangerous, separate from the content posted by users.

Publication function / editorial function. Activities that involve deciding whether to publish, withhold, prioritize, or alter content as a matter of editorial judgment. Courts often treat these functions as expressive and protected under the First Amendment and Section 230.

Recommendation-based algorithm. An algorithm that selects and pushes content to users based on predicted interest, often in the form of “For You” pages or suggested videos, rather than waiting for the user to search or request specific material.

Section 230 of the Communications Decency Act. A federal statute that generally shields online platforms from liability for content created by third parties. It protects platforms when they act as publishers of user content, but it does not automatically apply when claims focus on the platform’s own product design or first-party speech.

Search-based algorithm. An algorithm that compiles and orders results in response to a user's search query. The user initiates the request, and the algorithm retrieves content based on the terms entered, rather than pushing content that the system predicts will maximize engagement.

Streaks and reward mechanisms. Design features that track and display consecutive days of use or messaging, or that award badges and other rewards for continued engagement. These mechanisms increase the perceived cost of skipping a day.

Youth or minor (in this context). Children and adolescents under 18 years of age. The paper focuses on youth because their attention, reward processing, and self-regulation systems are still developing, which makes them more vulnerable to design-driven harms.