

How to Hold Social Media Platforms Accountable

A Roadmap for State Policymakers & Advocates for Legislation to Require Independent Algorithm Risk Audits

Authors

Nancy Costello, Mackenzie Almassian, Rebecca Sutton, Madeline Jones, Stephanie Diamond, Oluwadunni Ojumu, Amanda Raffoul, Meg Salvia, Jill R. Kavanaugh, S. Bryn Austin

Funding

This publication was supported by the Becca Schmill Foundation and the Strategic Training Initiative for the Prevention of Eating Disorders. Amanda Raffoul was supported by the Canadian Institutes of Health Research Institute of Population and Public Health grant MFE-171217.

Disclosures

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

Acknowledgments

We would like to thank the many experts, spanning the fields of law, technology, youth mental health, and policy advocacy, who we consulted with for our research and in the creation of this Roadmap. Their invaluable insights and perspectives were instrumental in informing our research and we are grateful for the generous contributions of their time and expertise.

Suggested Citation

Costello N, Almassian M, Sutton R, Jones M, Diamond S, Ojumu O, Raffoul A, Salvia M, Kavanaugh JR, Austin SB. (2024). How to Hold Social Media Platforms Accountable: A Roadmap for State Policymakers & Advocates for Legislation to Require Independent Algorithm Risk Audits. Strategic Training Initiative for the Prevention of Eating Disorders, Boston, MA. <https://www.hsppharvard.edu/stripped/social-media-algorithm-auditing/>

Copyright

Copyright © 2024 The President and Fellows of Harvard College

Table of Contents

Executive summary	4
Introduction	5
What do insights from public health and neuroscience reveal about the impact of social media on adolescents?	9
What economic incentives drive social media companies to compromise on safety for children online?	14
What legal obstacles prevent the regulation of harm caused by social media?	16
What legal strategies exist that could reduce the harm caused by social media algorithms?	18
Why focus on algorithm risk audits?	21
How to respond to commonly asked questions	32
Conclusion	34
Fact sheets & model legislation	35
References	50

Executive Summary

Social media platforms use engagement-based algorithms and deceptive design techniques that manipulate user behavior and promote unhealthy usage, harming young users' mental health and wellbeing. Due to a lack of transparency and accountability, platforms prioritize profits over children's safety. To create safer online environments for youth and hold social media platforms accountable, our research team at the Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED) developed model legislation, along with a supporting Roadmap, that pulls expertise from public health, neuroscience, economics, and legal studies.

How to Hold Social Media Platforms Accountable: A Roadmap for State Policymakers & Advocates for Legislation to Require Independent Algorithm Risk Audits provides guidance for policymakers and community advocates on reasonable actions that U.S. states can take to mandate third-party audits of social media platforms. Our Roadmap offers strategies for lawmakers and advocates to champion our model legislation *Social Media Algorithm Accountability Act* to create a healthier digital environment for young people.

This Roadmap also includes evidence of youth mental health harms due to social media, research on ad revenue platforms generate off of young users, legal strategies for accountability, responses to common questions, and printable fact sheets, including:

- **Fact Sheet: Science Says** on the science behind why third-party risk audits are necessary safeguards.
- **Fact Sheet: Economic Drivers** on the results of our simulation study estimating significant annual U.S. youth social media user ad revenue.
- **Fact Sheet: Legal Arguments** on the legal arguments in support of third-party risk audits to increase transparency and accountability of social media algorithms' harmful effects on youth users.
- **News Summaries** offering news stories on the negative impacts social media has had on youth mental health and well-being.

Youth deserve a safer digital environment. Our model legislation will bring needed transparency and accountability to protect youth mental health. Lawmakers must take action to curb the exploitation of young users on social media by championing independent algorithm risk audits.

Introduction

Social media has emerged over the last several decades as both an invaluable resource and a source of harm to youth. With the rise in social media use among youth and the increase in social media platforms' use of deceptive design techniques, there is growing concern about the harms of these platforms to the mental health and wellbeing of young people. Profit-driven design techniques manipulate user behavior and promote high usage patterns that trap users in loops of engagement, keeping them online for minutes or even hours longer each day than they would otherwise. Why would platforms do this? The answer is simple: The more time users spend online, the more platforms can charge advertisers for a chance to grab their attention. So, platforms have come up with an arsenal of deceptive design techniques to keep people glued to their screens, regardless of the consequences for young people, their families, and their communities. Without legal avenues in place to hold platforms accountable, young people are especially exploited by social media platforms' deceptive design and made more susceptible to anxiety, depression, eating disorders, suicidal thoughts, and other harmful mental health effects.

To create safer online environments for youth and hold social media platforms accountable, our research team at the Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED) developed model legislation, the [Social Media Algorithm Accountability Act](#), along with this supporting Roadmap, that pulls expertise from public health, neuroscience, economics, and legal studies. We created this Roadmap to guide policymakers and community advocates in championing common-sense steps that U.S. states can take to require independent, third-party algorithm risk audits for social media platforms and to make the results of those audits public. Our Roadmap offers easy-to-use legal and message framing strategies for lawmakers and community advocates to champion legislation that creates a healthier digital environment for all young people.

This Roadmap distills two years of painstaking legal research by our team to identify the most viable legal options states have to strengthen protections for young people on social media. Our full legal research article, which was published in the *American Journal of Law & Medicine*, can be accessed [here](#).

Primer on Key Terms

- **Social media:** A broad term for websites and applications that emphasize social interaction and information sharing among users through social connections established via user profiles. Social media is different from traditional media in that social media platforms typically do not charge users to access content. Instead, platforms charge advertisers to place their content based on the number, type, and time online of the users the advertisers aim to reach.
- **Algorithm:** Algorithms are sophisticated computer programs that social media platforms use to tailor what content to send to whom and when. There are different types of algorithms. Our Roadmap and model legislation focus on recommendation or engagement-based algorithms, and not algorithms used to respond to search engine requests. It is the algorithm that drives what appears in a social media feed on a smartphone or computer in a way that is uniquely tailored to the user. Algorithms work by feeding content to users automatically without users purposely searching for that content. They tailor a social media feed based on hundreds of different bits of information that platforms are

continually gathering about users, including their personal demographics (e.g., age, gender), location at the time they are online, other websites they visit, how many minutes or hours they've been on the platform at any one time, and much more.

The Invisible Influence of Social Media Algorithms

Because algorithms operate in the background, they are essentially invisible to users, and social media platforms keep their algorithms a trade secret. As users, we can see only our own social media feed, so we are often unaware that what others see can be entirely different as a consequence of uniquely tailored algorithms. While social media algorithms may be complicated, their purpose is simple: To draw users in and maximize how much time we spend online so that our attention, calculated in minutes and hours, can be sold to advertisers.

- **Algorithm risk audit:** An algorithm risk audit is a technique to make algorithms and their effects on content feeds visible. Audits provide an objective way to compare how recommendation or engagement-based algorithms may be automatically pushing out different content recommendations to different user groups or demographics in an unfair or unbalanced way, without those users purposely seeking out that content. Think of audits as a safety check that allows for the specific identification of how algorithms are pushing out content in biased ways. With this knowledge, advocates can pressure platforms to take steps to remedy their unfair practices. Algorithm risk audits conducted by independent third parties will make sure platforms cannot keep their algorithms and their biased effects secret any longer. These types of audits will provide a way to hold platforms accountable for keeping their digital environments safe for young people.

- **Deceptive design:** Deceptive design includes features in the design of social media apps and websites that are intentionally designed to drive users to spend more time on the platform, engage more, and share more of their personal data. These features often are based on extensive psychological research and are tested by platforms to maximize their effects (1-2).

Examples include:

- Limitless scroll and persistent notifications
- Increasing shock value of content and images over time to counteract naturally decreasing interest the longer a person stays online
- Misleading users to click on prompts that allow platforms to gather data on their contacts, emails and text, websites they've been to, their photo gallery, and location without the user knowing



- Default settings set to minimum rather than maximum privacy, with confusing information making it difficult to change settings to increase privacy
- Impossible or near-impossible cancelation procedures requiring multiple clicks without end, preventing users from deleting their accounts
- **Design-related harms:** Design-related harms refer to a social media company's product, service, or design feature that would result in a foreseeable risk of harm to children including mental health disorders, addiction, physical violence or online bullying, sexual abuse, illegal marketing of drugs, alcohol, or tobacco, and harms caused by predatory advertising.

What do insights from public health and neuroscience reveal about the impact of social media on adolescents?

The past decade has seen a marked increase in mental health concerns among young people in the United States (3-4). Many adolescents are struggling with anxiety, depression, suicide-related thoughts or behaviors, eating disorders, and bullying (4-6). One potential driver of increased mental health concerns among youth lies in the business practices of social media companies and the algorithms they employ, which can encourage excessive and harmful usage.

Why is social media a concern?

The vast majority of young people use social media platforms, and one third of U.S. teens say they are on social media “almost constantly” (7). Even younger children have access to social media platforms, despite the fact that most social media platforms – including TikTok, Instagram, Facebook, YouTube, and X (formerly known as Twitter) – have age limits built into their terms of use stating that people must be 13 years old before they can create an account. The age limit is in place because of the Congress in the Children’s Online

Privacy Protection Act (COPPA), which prohibits the collection of personal information online from a child under 13 years of age without parental permission. However, a national survey by Common Sense Media found that among 8 to 12 year olds in the United States, 38% report having used social media (8).

Researchers are zeroing in on why and how social media platform business practices may affect adolescents’ mental health, as the harmful effects on body image and risk of eating disorders are emerging as top concerns among youth. As a result of algorithms’ decision-making process, users are inundated with idealized images and messages about what is valued and popular in youth culture. During adolescence, it is a normal part of development for adolescents to make comparisons with their peers, assessing their appearance and experience against others, but how this otherwise normal process gets hijacked by social media platforms’ deceptive designs has emerged as a focal point of

concern (9-10). Adolescents are trying to figure out where they fit in and how they measure up, and they are making comparisons between themselves, their peers, and the influences they see online (11). For example, images promoting excessive and unrealistic thinness (which often are edited or altered to make people look thinner than they are) can lead to negative body image and distort adolescents' sense of what is realistic (10,12).

Qualitative research, which uses focus groups or interviews to explore individuals' experience in greater detail, has shed some light on adolescents' experiences. Adolescents compare themselves to others on Instagram based on the images they see, and many acknowledge the negative impact of edited photos on their self-esteem (13).



The type of social media platform matters. Highly visual social media platforms like Instagram and TikTok, which rely on algorithms to determine what content users see in their feeds, appear to have a bigger influence on poor body image compared to mostly text-based platforms like X (formerly Twitter) (14-15). One cross-sectional study found that adolescent girls looking at pictures of, and comparing themselves to, social media influencers had worse body image compared to when they looked at friends' or peers' posts (9). One year long study with teen boys and girls found that being highly focused on appearance on social media at the beginning of the study was linked with worsening depressive symptoms over the year of follow up (16). While social media content itself may advance unrealistic standards of beauty, it is the platforms' recommendation algorithms that promote what content dominates users' feeds and how much time they spend passively consuming it.

Eating disorders affect people of all ages, but adolescence is a time when the risk of developing an eating disorder is high (17). Eating disorders are also one of the deadliest mental health conditions (18-20). Research has found links between social media use and eating disorder symptoms (21-22). For example, using social media sites is linked with disordered eating behavior

like meal skipping, and mounting evidence shows that being highly focused on physical appearance in social media posts, including idealized images and unrealistically thin bodies, are key to this connection (23-24). Platform algorithms play a role by promoting the types of content users are exposed to most frequently. One study found that eating disorder content on TikTok alone has received over 13 billion views, demonstrating the potentially massive reach and impact of algorithms on young users (25).

What about the adolescent brain puts young people at risk due to social media platform practices?

There are several aspects of the adolescent brain that social media platforms deliberately exploit, placing young minds at a greater risk for mental health concerns related to social media use compared to the risk to adults or younger children.

Adolescents are highly sensitive to peer feedback and social cues.

Between ages 10 and 22 years, the adolescent brain is at a unique stage of development. Adolescents in this age range are particularly tuned in to social information about themselves in relation to their peers (26-27). When adolescents receive positive social cues through social media platform features such as “likes” on a post, they use that



feedback to shape their understanding of social norms and values. This can take a negative turn when highly liked posts display risk-taking behaviors like substance use or bodies that have been edited to be thinner (28-29).

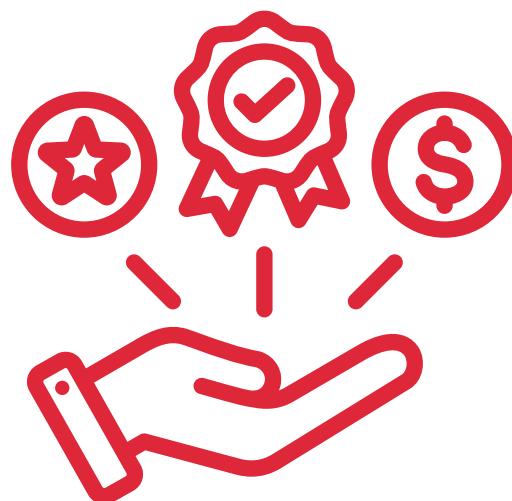
The adolescent brain is naturally highly emotional, making emotional responses to social media content more exaggerated. The frontal regions of the brain, which are responsible for reasoning, judgment, and decision-making, are not fully developed in adolescence. On the other hand, regions of the brain that handle emotion are more advanced in development during this period (30). The combination of heightened emotionality and underdeveloped judgment capacity makes adolescents prone to having more exaggerated emotional responses to social media than adults or younger children would have. Due to this perfectly normal but

uneven trajectory of development in different parts of the adolescent brain, an adolescent viewing an image of a digitally edited body on social media would likely experience body dissatisfaction and self-esteem concerns more than an adult or young child would.

The adolescent brain is particularly captivated by social media features that offer them a form of social reward. The adolescent brain, much like the adult brain, is motivated by reward. However, adolescents are more sensitive to rewards than adults or younger children due to developmental differences in the activation of the ventral striatum, a brain region involved in identifying, evaluating, predicting, and responding to rewards. This region has also been implicated in risk-taking behaviors and substance use disorders (31-33). Moreover, *social* rewards, such as those related to receiving positive feedback from peers, are more motivating for adolescents than adults (34). As such, platform features such as “likes” engage reward processing parts of the brain that motivate continued engagement with social media for adolescents who already have a high sensitivity to rewards. Paired with platform deceptive design features such as limitless scroll and persistent

notifications, adolescents can become trapped in cycles of continuous scrolling.

For adolescents, the urge to scroll incessantly can feel like an uncontrollable itch they cannot seem to scratch enough—the more they scratch, the more they want to keep scratching, even if they start to bleed. But make no mistake, the trove of internal Facebook memos and reports made public by the Facebook whistleblower makes clear that platform designers know exactly what effect their deceptive design features are having (35). Due to deceptive design features such as limitless scroll, the more adolescents engage on social media platforms, the more their brains motivate them to continue engaging, even if they are experiencing damaging mental health effects.



Is all social media bad?

No. Adolescents, like adults, use these platforms for entertainment and to connect with friends or communities. What we are concerned about is the deliberate design practices of social media platforms that take unfair advantage of young users' need for social connection to boost their profits with little regard for young people's mental health. Platforms have gotten a free pass for far too long, allowed by weak regulation to drive their products into the hands of 9 out 10 American teens without transparency or accountability for how their algorithms are working or the consequences on young people's mental health. Internal memos from Meta reveal that platform designers are aware of the effects of their platform's features on adolescents' engagement and mental health (36). We are not proposing to ban social media. It is the lack of accountability that is bad about social media practices today, and, together, that is what we can change.

What economic incentives drive social media companies to compromise on safety for children online?

Social media platforms thrive on substantial advertising revenue generated from young users, yet the extent of this economic benefit has not been well-documented. Under current U.S. law, these companies are not obligated to disclose the types of content young users are exposed to nor the impacts of such content (37-38). As a result, social media companies lack incentives to self-regulate and curb online harms affecting youth (39). To shed light on how much revenue social media platforms generate from minors, our STRIPED research team applied a rigorous, mathematical simulation modeling method to provide the first publicly disclosed estimates of the: (1) number of users and (2) annual advertising revenue generated from U.S.-based users aged 0-12 and 13-17 years for the six major social media platforms.

Platform	Projected 2022 U.S. Ad Revenue from U.S. Users Ages 0 to 12	Projected 2022 U.S. Ad Revenue from U.S. Users Ages 13 to 17
 YouTube	\$959 million	\$1.2 billion
 Instagram	\$802 million	\$4.0 billion
 Facebook	\$137 million	\$356 million
 TikTok	\$102 million	\$2.0 billion
  Twitter/X	\$19 million	\$40 million
 Snapchat	\$123 million	\$1.0 billion

Total projected 2022 U.S. ad revenue from Facebook, Instagram, Snapchat, TikTok, Twitter/X, and YouTube from U.S. users...

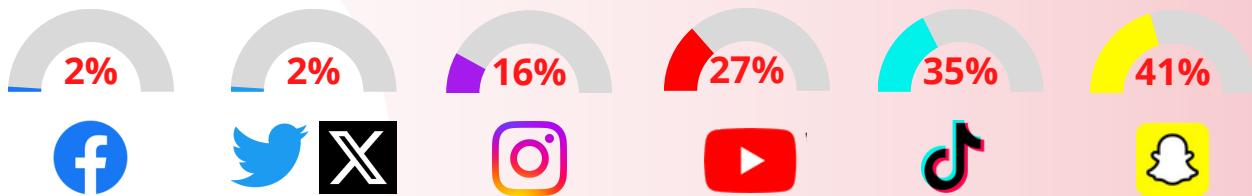
AGES 0-12 YEARS

\$2.1 billion

AGES 13-17 YEARS

\$8.6 billion

**% of total 2022 U.S. ad revenue from U.S. users
under 18 years old**



The study estimated annual advertising revenue from U.S. children ages 0-12 years to be over \$2 billion in U.S. dollars and from all children ages 0-17 years to be nearly \$11 billion across the 6 major platforms (40). On most social media platforms, children under the age of 13 years are not even supposed to create their own account, yet for several platforms, we estimated nearly 30 to 40 percent of their annual advertising revenue is generated from users ages 0-17 years. Despite Meta publicly stating that they did not consider profitability when designing products for teens, their internal documents showed their product teams assigned a "lifetime value" of \$270 to each teen user of the platform, and took profitability into consideration when making decisions, raising issues around transparency, potential exploitation of vulnerable youth, and ethics (41). The massive revenue generated from young users discourages social media platforms from self-regulation and further demonstrates the need for greater transparency and legislative intervention to curb harms. Our full study details can be [accessed here](#).

What legal obstacles prevent the regulation of harm caused by social media?

First Amendment

The First Amendment of the U.S. Constitution protects a wide array of speech, ranging from highly protected speech such as political speech, to less protected speech such as commercial speech and sexually explicit speech. Certain speech, such as defamation, incitement, and fighting words are not protected by the First Amendment. Legislators trying to curb harms resulting from online activity must avoid infringing on First Amendment protections. Laws attempting to restrict or prohibit speech based on its content, such as hate speech, are unconstitutional, thus, the content placed by the algorithm cannot be targeted by legislation, but the harmful design of an algorithm can be.

An algorithm is a set of step-by-step instructions that tells computer software exactly what to do to perform a specific task, achieve a certain result, or solve a problem. Algorithms are used widely in all areas of technology, including internet search engines, social media platforms, cybersecurity, and analyzing large sets of data. A search

engine algorithm is actively directed by the user (i.e., the user is entering a search term), while a recommendation or engagement based algorithm passively populates a user's page without any direct action by the user. Historically, algorithms have been considered protected speech under the First Amendment because they are compilations of expression generated by computer engineers and others. This notion is currently being contested in court challenges claiming that recommendation algorithms are product designs, and not speech, causing harm to children. The U.S. Supreme Court has not decided this issue. As long as this uncertainty remains, the First Amendment appears to be the greatest hurdle to curbing harms caused by social media. The model legislation focuses on measuring the design-related harms caused by recommendation algorithms.

Section 230

Another major obstacle to regulating the harm caused by social media platforms has been Section 230 of the Communications Decency Act. Section 230 grants immunity to online services, protecting such service providers from being sued for the harmful speech of third parties on their platforms. Section 230 has become a major roadblock to legislation aimed at protecting youth from online harms. Challenges to Section 230 have been filed in courts in more recent years, arguing that it has been applied in an overbroad manner and wrongly granted immunity to a social media company engaged in bad conduct.

In two recent decisions, the U.S. Supreme Court declined to decide the issue of broad Section 230 immunity in the context of platforms promoting harmful (terrorist) content published by third parties on the platform. However, lower courts have recognized that Section 230 does not protect social media platforms in some instances of negligence, failure to warn, fraud, and products liability. These courts have determined that Section 230 immunity is not unlimited and should not protect a social media company for everything that transpires on its platform.



What legal strategies exist that could reduce the harm caused by social media algorithms?

Federal Trade Commission Act

One approach to regulate design-related harms on social media platforms is applying Section 5 of the Federal Trade Commission Act, which declares “unfair or deceptive acts or practices in or affecting commerce” are unlawful. The Federal Trade Commission (FTC) finds that an act or practice is unfair where “[1] the act or practice causes or is likely to cause substantial injury to consumers which [2] is not reasonably avoidable by consumers themselves and [3] not outweighed by countervailing benefits to consumers or to competition.” The FTC finds that an act or business practice is deceptive when (1) a representation, omission, or practice misleads or is likely to mislead the consumer; (2) a consumer’s interpretation of the representation, omission, or practice is considered reasonable under the circumstances; and (3) the misleading representation, omission, or practice is material.

The FTC may sue under Section 5 of the FTC Act if it can prove that the persistent pushing of harmful algorithmic content, such as eating



disorder content shown to a user through the social media platform’s algorithm, meets the definition of an “unfair or deceptive business practice,” regardless of the platform’s intent to harm the user. A claim of unfair or deceptive business practices may also be brought by states’ attorneys general offices against social media companies, similar to a recent legal action brought against META by 41 states’ and the District Columbia attorneys general offices. However, to bring a robust, successful FTC claim, or claim by a state attorney general, both entities should have specific data showing design-related harm caused to adolescents and linking that harm to business practices of a social media platform.

Age Appropriate Design Code

The California Age-Appropriate Design Code Act was signed into law in September 2022 and was set to go into effect on July 1, 2024; however, a court ruling in September 2023 put implementation on hold. The Act provides a series of standards that online goods, services, or products must comply with if they are likely to be accessed by a child. This includes that an online good, service, or product must act in the best interests of the child, maintain the highest level of privacy for children by default, and avoid collecting geolocation information. Additionally, under this law, social media platforms, goods, or services must avoid using the personal information of a child user in a way that is harmful to the physical health, mental health, or well-being of a child.

The California law is modeled after the Age Appropriate Design Code in the U.K., which came into force on September 2, 2020. The U.K. law appears to have inspired some meaningful change by platforms. For example, TikTok has introduced that it will restrict sharing options for children and adolescent users, and it has disabled notifications from the app after bedtime for users under the age of 18 years. Google has also announced changes, including that it will disable the "location history" service for children.

Additionally, YouTube has updated its default privacy settings and it has also turned off the autoplay option by default for users under 18 years.

The success of the Age Appropriate Design Code in the U.K. and the changes platforms have already taken to meet the criteria of the U.K. Code indicate that the California law could offer some help to protect youth online if the court allows it to be implemented.

Data Protection Impact Assessments

The California Age Appropriate Design Code requires businesses to complete a Data Protection Impact Assessment before a new online service, product, or feature is offered to the public and to maintain documentation of this assessment. A Data Protection Impact Assessment is a survey that assesses and mitigates risks that arise from business practices when that product, design, or feature is likely to be accessed by a child.



This assessment will address whether a product, service, or feature could harm children or expose children to harmful content, could lead children to experiencing harmful contacts, whether it could permit children to witness or participate in harmful conduct, whether algorithms used could harm the child, and whether targeted advertisements could harm the child.

The California law's Data Protection Impact Assessments are confidential and will not be publicly disclosed. The assessments need only be made available to the California Attorney General. Further, the assessments will be conducted internally by the company, instead of by a third-party, meaning that Facebook, for example,

would conduct an assessment on itself. These two factors likely would render the Data Protection Impact Assessments ineffective. A stronger more transparent approach would be to require public disclosure of the assessment and an external investigation to hold social media platforms more accountable for the harms it causes to minors as result of its algorithmic design. Implementing an algorithm risk audit is an essential enforcement mechanism both for proving FTC Section 5 claims and enforcing the California Age Appropriate Design Code.

Other states have passed or proposed similar laws requiring data protection impact assessments, including Florida, but they are equally likely to be ineffective because the social media platforms, and not an independent third party, would conduct the assessments.

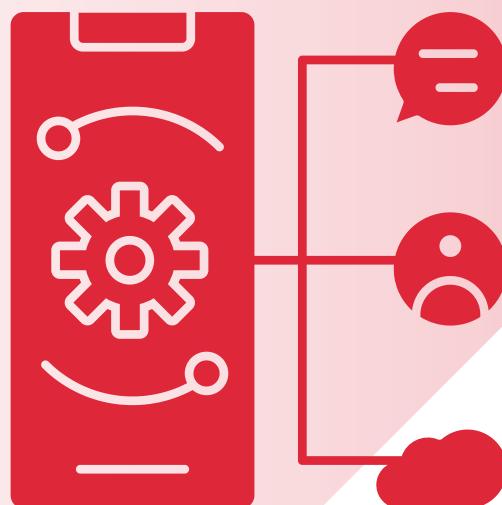


Why focus on algorithm risk audits?

In a world where technology is constantly evolving, lawmakers must enact protections for child and adolescent users for the potential harms that social media use can cause through platform features, designs, and programming decisions. To best draft such legislation, policymakers must understand what design-related harms social media causes and the effect these harms have on adolescents. Any law aimed at protecting adolescents from the harmful effects of social media must address how social media platforms employ algorithms in the function and design of their product. Further, for any law in this area to be effective, it should incorporate an enhanced means of enforcement, rather than just mere prohibitions of particular acts. This dual task could best be accomplished through the use of algorithm risk audits.

An algorithm risk audit is a technique to make algorithms and their effects on content feeds visible. Audits provide an objective way to compare how algorithms may be automatically pushing out different content to different users in an unfair or unbalanced way without those users purposely seeking out that content. Audits are a safety check that identify specifically how algorithms are

pushing out content in biased ways, and with this knowledge, platforms can be pressured to take steps to remedy their unfair practices. Algorithm risk audits that are conducted by independent third parties at regular intervals would be publicly disclosed to make sure platforms cannot keep their algorithms and their biased effects secret. The audits will help law enforcement hold platforms accountable for keeping digital environments safe for young people.



Legally mandating algorithm risk audits is a relatively new strategy that is gaining traction both nationally and globally. New York City was among the first jurisdictions to mandate these types of audits, passing a law on December 11, 2021, requiring annual audits assessing bias in automated employment decision tools, which use algorithms to screen applicants for employment positions (42). By requiring these audits, known as bias audits, this law helps identify when an algorithm might intentionally or unintentionally weed out applicants based on certain demographics, such as race and gender, regardless of whether they meet the job qualifications (42). The New York City law, which took effect on April 15, 2023, requires an impartial evaluation by an independent auditor, the results of which must be made publicly available (43).^{Note 1}

Another real-world example to look to for guidance on how an algorithm risk audit might work is the recent settlement between Meta Platforms, Inc. (Meta) and the U.S. Department of Justice (DOJ). On June 21, 2022, the DOJ announced it entered into a settlement agreement that resolved allegations that Meta engaged in discriminatory advertising in violation of the Fair Housing Act (FHA) (44). The agreement resolved a lawsuit filed against Meta by the United States, which alleged that

“Meta’s housing advertising system discriminated against Facebook users based on their race, color, religion, [gender], disability, familial status, and national origin” (44). Meta was charged with unevenly displaying housing ads to Facebook users of certain FHA-protected demographics, such as gender and race. The settlement between Meta and the DOJ required Meta to develop a new system to make the display of housing ads more even across race and gender groups, and therefore address the discrimination caused by its algorithms (44).

Note 1: For a deeper dive

The bias audits will measure the disparate impact the use of algorithms has on a specific demographic group by comparing the number of applicants, who meet the job qualifications, from a specific demographic selected to move forward in the hiring process to the number of those in the most highly selected demographic. See CITY N.Y. RULES, tit 6, §§ 5-300–5-301 (2023). For additional information on the New York City law, please see reference #47.

Details of the Meta / U.S. Department of Justice Settlement on Housing Ads

The settlement in the Meta housing ad case took a three-step approach: (1) identify the specific harm, (2) determine how to measure the extent of harm, and (3) agree on reporting periods and benchmarks to reduce harm. The first step was to identify the specific harm, which was the discrimination caused by housing ads being unevenly displayed to Meta users of certain demographics, namely gender and race, in violation of the Fair Housing Act.

The second step—to determine how to measure the extent of harm—required Meta and tech experts to figure out how to examine Meta's data to assess the extent of the discriminatory harm. The discriminatory harm is shown through differences between the eligible and actual audiences for housing ads. The eligible audience includes all users who (1) fit the targeting options selected by an advertiser for an ad, and (2) were shown one or more of any type of ad on a Meta platform over the past 30 days (44). The actual audience includes all users in the eligible audience who actually viewed the specific ad (44). Once these audiences are identified, a measurement is taken to determine the difference between the number of users of each demographic group in the eligible versus actual audience using a measure called the Earth Mover's

Distance, one of many possible measures used by algorithm risk auditors (43).^{Note 2}

Under the final step of the settlement, Meta and the DOJ have to agree on reporting periods and benchmarks to reduce harm. This requires that Meta meet “certain [benchmarks] within a specific period of time” to reduce the variance between the eligible and actual audience for housing ads (45).

Note 2: For a deeper dive

Earth Mover's Distance is basically a measure of fairness that can determine the difference between the eligible audience of content and the actual audience who views the content. For a detailed explanation of how the Earth Mover's Distance functions please see reference #44, pages 167-68.

To meet these benchmarks, Meta had to develop a system to reduce variances between the eligible and actual audiences for housing ads (43).^{Note 3}

Additionally, under the settlement, Meta must prepare a report every four months confirming that it has met the benchmarks for the previous four-month period (46). Importantly, Meta and the DOJ selected an *independent, third-party reviewer* “to investigate and verify on an ongoing basis” whether the benchmarks are being met (44). The third-party reviewer, therefore, serves as an objective check on Meta’s compliance with the agreement.

This settlement agreement marked the first time Meta was subject to court oversight for its ad targeting and delivery system (44). The settlement required Meta to alter the way its algorithms target and deliver housing ads to ensure compliance with the Fair Housing Act. This three-step approach to monitoring and measuring design-related harm caused by algorithms can be adapted to assess harm caused by social media platforms to adolescent users in the form of an algorithm risk audit.

Note 3: For a deeper dive

Once a variance is detected between the eligible and actual audiences using the Earth Mover’s Distance measurement, Meta can use the Variance Reduction System (VRS) to help reduce that variance. Think of the two working in tandem with one another, similar to how a radar and auto-pilot work with a plane. The radar identifies when there is a hazard ahead and the auto-pilot shifts the plane’s speed or altitude to avoid the hazard. Likewise, the Earth Mover’s Distance identifies the variance between the audiences and the VRS works to shrink that variance. See Settlement Agreement at 6, United States v. Meta Platforms, Inc., No. 1:22-cv-05187 (S.D.N.Y. June 21, 2022). For further explanation please see reference #44.

Algorithm Risk Audit Proposed

Legislation: Three-Step Approach

A legally mandated risk audit of engagement-based algorithms would measure the harms caused to youth and adolescents by the design and programming of engagement-based algorithms used by social media platforms. The risk audit would mirror the three steps used in the Meta/DOJ settlement: (1) identify the specific harm(s), (2) determine how to measure the extent of each harm, and (3) agree on reporting periods and benchmarks to reduce harm. Our model legislation, the Social Media Algorithm Accountability Act, does provide a specific set of design-related harms to be measured, but lawmakers could customize it to determine what kind of design-related harms they want to address.

To understand how an algorithm risk audit would work, think about the specific harm adolescent users experience when confronted with pro-eating disorder content. Pro-eating disorder content may include very restrictive dieting plans, extreme exercise regimens, and images of very thin bodies with protruding bones that intend to serve as “inspiration” for users who are seeing the content (47). A risk audit could be used to measure the extent of this harm and to whom the content is directed by engagement-based algorithms, which could lend government and consumers the leverage to pressure social media platforms to agree to alter the way their algorithms function to reduce the design-related harm.

Algorithm Risk Audits



Step 1: Identify Harms

Using the audit's three-step approach, the first step would be to identify the specific design-related harm. The specific harm might be described as "eating disorder rabbit holes," (48) such as when adolescent social media users begin engaging with content related to mental health and body image and then are progressively shown more and more pro-eating disorder related content (50).^{Note 4}

Step 2: Measure Harms

The second step would be to determine how to measure eating disorder rabbit-holes. For this step, a social media platform might be required to measure the number of users who have been pushed by algorithms from viewing mental health and body image-related content into viewing pro-eating disorder related content (e.g., an extremely restrictive dieting plan) within a certain number of minutes, hours, or days. The social media platform could measure the users who plunge into eating disorder rabbit holes and compare the demographics of these users. If the specific concern is adolescent users, the social media platform could compare the number of all users, of any age and gender, who enter eating disorder rabbit holes to that of adolescent users who do. Comparing the difference between these numbers would show whether adolescent users are disproportionately likely to be propelled down eating disorder rabbit holes by engagement-based algorithms.

Note 4: For a deeper dive

"[TikTok] starts recommending content tied to eating disorders and self-harm to 13-year-olds within 30 minutes of their joining the platform, and sometimes in as little as three minutes . . ."

Step 3: Report Harms

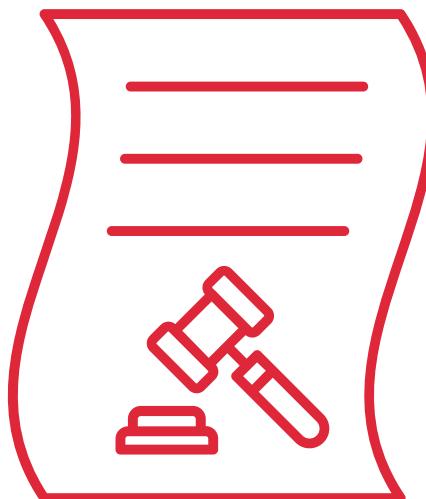
The social media platform and the governmental body that enacted a law requiring an algorithm risk audit would then move to the third step—agreeing on reporting periods and the benchmarks to reduce harm. The legislative body could determine that the social media platform needs to implement a new system, similar to Meta's development and implementation of the VRS, to alter its current algorithm to address the disparate impact it has on adolescent users. The legislative body might alternatively assign responsibility for determining benchmarks to a state administrative agency or governmental office with particular knowledge of issues relating to adolescent mental health and technology. In implementing such a change, the parties would need to determine benchmarks for improvement and reporting periods to ensure compliance with those benchmarks. Reporting periods could be required at any reasonable rate, such as on a quarterly, monthly, or even weekly basis. Similar to the Meta/DOJ settlement, a law requiring algorithm risk audits would require that the reports be evaluated by a third-party, independent reviewer to ensure compliance with the benchmarks agreed upon by the parties.

Confronting a Possible Constitutional Challenge

If a social media company challenges the constitutionality of the required algorithm risk audit, the first two requirements (identifying the design-related harms and measuring those harms) would not violate the First Amendment because these steps do not regulate content on a platform. Instead, these requirements mandate that the harms to be measured are identified and the incidents of harm caused by the platform's algorithmic design are measured. However, a platform may claim the third step, mandating that the platforms reduce the design-related harms by agreed upon reporting periods and

benchmarks, illegally regulates content because the platform might need to change its algorithm to cause less harm. In the Meta/DOJ settlement this third step did not violate the First Amendment because it corrected Meta's violation of the Fair Housing Act. In contrast, the design-related harms to be measured by our model legislation are not illegal. For example, although it may be disturbing for viewers, a video depicting an extremely thin adolescent girl with protruding bones is not violating any law. Thus, a platform may claim that the third step, which may regulate a platform's algorithm, could be considered censoring content and run afoul of the First Amendment.

Advocates, however, for our model legislation would be on solid ground to persuasively assert that requiring a platform to modify its algorithmic design to reduce identified design-related harms, does not shut down content but rather imposes a "time, place, or manner" restriction by which the speech can be presented. A time, place, or manner restriction is frequently permissible under the First Amendment because it does not shut down all speech. It merely regulates the manner in which the speech can be presented to prevent proven harm.



Here, the third-step provided in our model legislation would likely survive as a time, place, or manner restriction. To impose such a restriction, the speech must be content-neutral, serve a substantial government interest and be narrowly tailored, and allow for alternative channels of communication. Modification of a platform's algorithm required by the third step could identify certain speech as harmful, but it is motivated by a content-neutral purpose: protecting youth and adolescent social media users. Lessening the design-related harms caused to youth and adolescent users on social media platforms is a substantial governmental interest and requiring algorithmic changes is narrowly tailored. This third step does not ban this speech altogether, it merely requires that such content is not algorithmically pushed unevenly to youth and adolescent users. A social media user can deliberately and independently still search for the content, or specifically request the content within the platform, using mechanisms such as the platform's search bar or search page. These searching mechanisms provide a sufficient channel of alternative communication of the speech. Therefore, the third-step would likely survive a First Amendment challenge.

However, to discourage a First Amendment challenge, the model legislation could make the third step, mandating reporting periods and benchmarks to reduce harm, optional. Although this might seem to dilute the purpose of the law, it does not because the first two steps of the law still produce the evidence of harm needed for an attorney general's office to bring a claim of deceptive advertising or unfair business practices against a social media company. The first two steps require that (1) the design-related harms to be measured are identified, and (2) that the risk audits, conducted by an independent auditor, measure the incidents, and the results of the audits be publicly disclosed.



To avoid an attorney general claim, the legislation allows a social media company to be proactive and take a compliant, affirmative step to alter the design of its platform so it does not hurt youth and adolescents. If the platform does not choose this option, then an attorney general can use the evidence of design-related harm produced by the algorithm risk audit to bring its cause of action, and the social media company will face negative publicity, especially because the platform had the option of mitigating the harm, but chose not to do so. To further ensure that the model legislation could withstand a constitutional challenge, a severability clause should be added so that if any portion of the legislation, such as the third step, should be found to violate the First Amendment, the first two steps, mandating the identification and measuring of design-related harm would still be required.

Public Disclosure of Social Media's Algorithmic Harm Must be Required

Beyond this three-step approach, a law mandating algorithm risk audits must require public disclosure of a social media platform's compliance with the agreed upon benchmarks. The compliance reports developed by the platform, and reviewed by a third-party, should be made publicly available.^{Note 5}

This level of transparency would encourage social media platforms to be diligent in curbing harms caused by algorithms. Even more importantly, algorithm risk audits could provide proof of design-related harm that could significantly add to the mounting evidence that shows there is a causal link between social media platforms' business practices and harm to adolescents.

Note 5: For a deeper dive

In recent legal challenges, social media companies have argued that requiring the platforms to disclose such information is a form of compelled speech, and thus, unconstitutional. However, legal analysts do not consider this argument to be robust because reporting requirements generally do not trigger a First Amendment analysis. In cases where a First Amendment analysis does apply, courts have often ruled in favor of government regulation, requiring commercial disclosure of factual information by companies that is connected to an important public interest. See *Zauderer v. Office of Disciplinary Counsel*, 471 U.S. 626 (1985).

Indeed, if social media platforms are able to alter their practices to comply with benchmarks required under a law of this kind, it would indicate that these platforms have some control over the harms their algorithms cause. Armed with evidence of harm, policymakers, state attorney general offices, and state administrative agencies could pursue legal action to hold social media platforms accountable for the design-related harm caused to adolescent users that they negligently create and ignore.

How to respond to commonly asked questions

How will this bill help reduce design-related harms caused by social media platforms?

This bill will mandate that social media platforms conduct independent third-party risk audits of engagement-based algorithms to assess whether harmful mental health-related content is being unfairly targeted to youth by a platform's design practices. If the results of an audit reveal that a platform's algorithms unfairly target some youth more than others, the platforms could agree to correct that biased, harmful distribution. If the platforms fail to make such an agreement, audit results could be provided to law enforcement agencies, enabling them to bring claims against social media platforms.

Who might oppose this bill?

Until recently, social media platforms in the U.S. have faced very few regulations. Representatives from the tech industry have come out again and again to oppose any type of regulation, no matter how reasonable or how urgent the need, and instead they typically argue that they should be left to regulate themselves without any accountability to the public. Unfortunately, self-regulation without accountability to the public does not work. Social media companies generate an immense amount of revenue from underage users, so they will not make meaningful changes to increase safety unless they are held accountable for the harm they are doing to children and adolescents.

This legislation does not take away any free speech rights. It will (1) assess algorithmic design in promoting harmful mental health content and (2) hold social media companies accountable for design-related harms.

How to respond to commonly asked questions

Can the Social Media Accountability Act be used to harm vulnerable populations?

Our model legislation cannot be used to target specific marginalized or vulnerable populations, nor take away users' rights to the online spaces they visit, nor take away users' rights to see the results of any search they choose to do. Instead, our model legislation will mandate accountability and transparency for social media platforms to reveal the potential harms of the algorithms that automatically drive content to users without their actively searching for it. It will not reveal individual user information, nor will it publicize the types of content that any individual young people see online.

In fact, our model legislation is anti-bias, as it aims to reveal biases that are currently hidden by social media platforms. Social media platforms can provide a safe space for youth from marginalized communities, and this legislation provides the additional benefit of ensuring that they will not be unfairly targeted by biased algorithms.

Conclusion

Social media is here to stay, and in fact many young users have found ways to make the best of what social media offers to make meaningful connections with their friends and community. But as the research shows too well, U.S. regulation on social media platforms is weak and woefully out of step with today's social media, failing to require of platforms any semblance of accountability or transparency. The consequence? This weak regulatory state of affairs is allowing platforms to put young users' mental health in harm's way just to satisfy the platforms' insatiable greed for astronomical profits. But communities across the country are starting to speak out and push back on platforms as more people come to realize that it is within their power to create a healthier digital environment for all young people.

This Roadmap offers easy-to-use legal and message framing strategies for lawmakers and community advocates to champion common-sense legislation for a safer future for youth on social media. The *Social Media Algorithm Accountability Act* provides lawmakers and advocates with the tools to make social media platforms accountable for protecting the mental health and wellbeing of young users. Using the resources provided in this Roadmap, we can bring together policymakers, community advocates, parents, and young people themselves to shape a safer digital world that all young people deserve.

Action steps for community advocates

Contact your state representative about championing the *Social Media Algorithm Accountability Act*

Action steps for state lawmakers and policy staff

File the *Social Media Algorithm Accountability Act* and work with your colleagues to pass it into law

Fact Sheets & Model Legislation

Science Says Third-Party Risk Audits are Necessary	36
Safeguards for Youth Mental Health	
Legal Arguments in Support of Third-Party Risk Audits	38
How Economic Drivers Undermine Child Safety Online	39
News Summaries	40
Model Legislation: Social Media Algorithm Accountability Act	43

Science Says Third-Party Risk Audits are Necessary Safeguards for Youth Mental Health

Social media's enduring presence in our society demands attention to the negative impact its deliberate design strategies and algorithms can have on youth mental health.

Why is social media a concern?

Mental health concerns among young people in the U.S. have been worsening in recent years, with social media as a key potential driver (1-2). Many adolescents are struggling with anxiety, depression, suicide-related thoughts or behaviors, eating disorders, and cyberbullying (2-4). Social media only worsens these concerns by bombarding youth with idealized images and videos on highly visual platforms like Instagram and TikTok, influencing their perceptions of what is valuable and popular in youth culture. While self-comparisons are a natural part of teen development, unchecked algorithms exploit this process, leading to heightened body image and self-esteem concerns (5-6).

What about the adolescent brain puts young people at risk?

- The adolescent brain is sensitive to peer feedback and social rewards, often relying on "likes" to shape their understanding of social norms.
- Their emotional responses to their social world online are intense due to underdeveloped reasoning and judgment capacity.

- Deceptive design features, combined with heightened sensitivity to rewards, can trap them in excessive scrolling, even if it harms their mental health.

What can be done to foster online safety?

To create safer online environments and hold platforms accountable, the Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED) has developed the model legislation [Social Media Algorithm Accountability Act](#) and an accompanying [Roadmap](#), providing policymakers and community advocates a strategic blueprint for championing policy change to create a healthier digital environment for all youth.

Action Steps

Community Members: Propose to your state representatives to champion the [Social Media Algorithm Accountability Act](#).

State Policymakers: File the [Social Media Algorithm Accountability Act](#) and work with your colleagues to pass it into law.

Have questions about our Roadmap or model legislation? **Contact us:**
striped@hsp.harvard.edu

References

1. Keyes KM, Gary D, O'Malley PM, Hamilton A, Schulenberg J. Recent increases in depressive symptoms among U.S. adolescents: Trends from 1991 to 2018. *Social Psychiatry and Psychiatric Epidemiology*. 2019; 54(8): 987-996.
2. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*. 2018; 6(1): 3-17.
3. US Department of Health and Human Services/Centers for Disease Control and Prevention. *CDC Morbidity and Mortality Weekly Report*. 71(1): 230-282.
4. American Academy of Pediatrics. AAP-AACAP-CHA declaration of a national emergency in child and adolescent mental health. Oct 19, 2021. Accessed Mar 2, 2024
<https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/>.
5. Jiotsa B, Naccache B, Duval M, Rocher B, Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness. *International Journal of Environmental Research and Public Health*. 2021;18(6):2880.
6. Papageorgiou A, Fisher C, Cross D. "Why don't I look like her?" How adolescent girls view social media and its connection to body image. *BMC Women's Health*. 2022; 22: 261.

Legal Arguments in Support of Third-Party Risk Audits

In an era of advancing technology, protecting young users from potential harms arising from social media use is paramount. Effective laws require understanding the role algorithms play in perpetuating harms on social media and should incorporate enforcement like algorithm risk audits.

What is an algorithm risk audit (ARA)?

An algorithm risk audit (ARA) helps us see how recommendation, or engagement-based, algorithms are designed to tailor a social media feed to automatically show different content to different users. The audits act as a safety check to:

- Provide transparency about the effects of algorithms on content feeds,
- Objectively compare how algorithms are designed to distribute content to users,
- Identify biased content distribution practices to hold platforms accountable.

What are some real-world examples of ARAs?

In a settlement between Meta and the US Department of Justice, Meta must address the discriminatory aspects in its algorithm design to promote housing ads equitably across race and gender groups.

Does an ARA violate the First Amendment?

Steps 1 and 2 of an ARA (identifying and measuring harms) do not infringe on the First Amendment because they do not regulate content posted on social media. To address potential constitutional concerns, the third step (reducing harms) can be optional while requiring harm identification and measurement remains mandatory. Results of the audits can be used by law enforcement to bring deceptive advertising or unfair business practice claims against social media companies.

How does an ARA work? It will:

- Identify Specific Harms: Define the harms to be measured, e.g., pro-eating disorder content.
- Measure Extent of Harm: Choose how to measure the harm caused by a biased distribution of content., e.g., number of young users (versus adult users) within a week who are prompted by algorithms to view content promoting extremely restrictive dieting plans.
- Reduce Harm: Legislative body and social media platform agree on benchmarks and reporting periods to mitigate identified harms, e.g., quarterly or monthly report.

Algorithm risk audits would be conducted by independent third-party reviewers at regular intervals to ensure impartiality. Results of these audits must be made publicly available to discourage the business practice of designing algorithms that promote extreme content in a biased way. Review the full legal rationale here.

Action Steps

Community Members: Propose to your state representatives to champion the [Social Media Algorithm Accountability Act](#).

State Policymakers: File the [Social Media Algorithm Accountability Act](#) and work with your colleagues to pass it into law.

Have questions about our Roadmap or model legislation? **Contact us:** striped@hsph.harvard.edu

How Economic Drivers Undermine Child Safety Online

Social media companies make a lot of money from advertising to young users, but they do not have to reveal how it affects youth. The combination of astronomical profits and a lack of transparency means social media companies have little incentive to protect young people online or adopt meaningful self-regulation, highlighting the need for government regulation.

What type of study was conducted?

Our research team with the Strategic Training Initiative for the Prevention of Eating Disorders (STRIPED) conducted a simulation study using rigorous, state-of-the-art mathematical methods to estimate annual ad revenue generated from users under the age of 18 years in the United States for six social media platforms: TikTok, Facebook, Instagram, YouTube, Twitter (now X), and Snapchat. Data were sourced from Insider Intelligence's eMarketer database, which contains estimates and historical data forecasts and analyses. We also used public survey data, including Pew Research, Common Sense Media, and Qustodio. Note that ad revenue per user by age group is based on the assumption that all users are targeted equally by ads.

Why was the study conducted?

Under current U.S. law, social media platforms have no legal obligation to release data on the types of content youth are exposed to, the impacts of content, the number of youth on the platform, nor how much revenue they generate.

What were the results?

See graphic for the total projected 2022 U.S. ad revenue from Facebook, Instagram, Snapchat, TikTok, Twitter/X, and YouTube from U.S. users.

*Have questions about our Roadmap or model legislation? **Contact us:** striped@hsp.harvard.edu*

Platform	Projected 2022 U.S. Ad Revenue from U.S. Users Ages 0 to 12	Projected 2022 U.S. Ad Revenue from U.S. Users Ages 13 to 17
 YouTube	\$959 million	\$1.2 billion
 Instagram	\$802 million	\$4.0 billion
 Facebook	\$137 million	\$356 million
 TikTok	\$102 million	\$2.0 billion
  Twitter/X	\$19 million	\$40 million
 Snapchat	\$123 million	\$1.0 billion

Why are these results important?

To our knowledge, this is the first study to offer estimates of the number of youth users on these platforms and how much social media platforms generate in ad revenue based on child users on the platforms. The massive revenue generated from young users discourages social media platforms from self-regulation and further demonstrates the need for greater transparency and legislative intervention to curb harms. [Our full study details can be accessed here.](#)

Raffoul A, Ward ZJ, Santoso M, Kavanagh JR, Austin SB. Social media platforms generate billions of dollars in revenue from U.S. youth: Findings from a simulated revenue model. PLOS ONE. 2023;18(12): e0295337.

News Summaries

Mental Health Harms and Loss of Life Experienced by Young People Due to Social Media Platform Design Practices

1. In 2017, Molly Russell, 14, died by suicide after a months-long struggle with distressing online content. Unbeknownst to her family, Molly engaged with thousands of pieces of self-harm and suicide-related material on Instagram, Pinterest, Twitter, and YouTube. <https://www.theguardian.com/technology/2022/sep/30/how-molly-russell-fell-into-a-vortex-of-despair-on-social-media>
2. Alexis Spence, who began using Instagram at 11, suffered anorexia, self-harm, and suicidal thoughts due to the platform's "addictive" nature. Spence's engagement with harmful content led to her hospitalization at 19. In 2022, a lawsuit against Meta was filed on her behalf, as the company was aware of harm to teenage girls caused by their algorithms. <https://www.nbcnews.com/tech/social-media/meta-lawsuit-instagram-caused-eating-disorder-self-harm-rcna32221>
3. The mother of 11-year-old Selena Rodriguez, who died by suicide, is suing Meta and Snapchat. For over two years, Selena struggled with severe addiction to Instagram and Snapchat, battling mental health concerns and pressure to share sexually explicit content. Her time on these platforms resulted in her hospitalization for low self-esteem, eating disorders, and self-harm before her tragic death. <https://www.washingtonpost.com/nation/2022/01/22/selena-rodriguez-suicide-meta-snap-lawsuit/>
4. In 2021, Snapchat discontinued its "speed filter," which allowed users to record and share their speed while driving, following widespread criticism and its link to numerous car accidents and fatalities. Among the tragic incidents is the 2017 story of three youth from Wisconsin, aged 17, 17, and 20, who reached a speed of 123 mph on the feature before fatally colliding with a tree. <https://www.npr.org/2021/06/17/1007385955/snapchat-ends-speed-filter-that-critics-say-encouraged-reckless-driving>
5. Laura Thornton's 13-year-old daughter returned from a summer away, deeply affected by anorexia, leading to her immediate hospitalization. Engaging in online "meal plan contests" where girls boast about minimal eating, she fell prey to extreme diet culture perpetuated on social media. <https://www.washingtonpost.com/opinions/2023/10/04/eating-disorders-social-media-anorexia-democracy-disinformation/>

News Summaries

Mental Health Harms and Loss of Life Experienced by Young People Due to Social Media Platform Design Practices

6. In 2020, Carson Bride, a 16-year-old from Lake Oswego, Oregon, tragically died by suicide after receiving hundreds of harassing messages on the Yolo app, integrated into Snapchat, which allowed anonymous communication. His mother, Kristin Bride, discovered Carson's distressing online interactions and his desperate attempts to stop the harassment through his search history in the days leading up to his death. <https://www.theguardian.com/lifeandstyle/2024/jan/16/online-harms-social-media-lawsuits>
7. Samuel Chapman of Los Angeles is urging California lawmakers to protect kids online after his 16-year-old son, Sammy, died from a fentanyl overdose from drugs purchased through Snapchat. Sammy fell victim to a drug dealer who used the platform's disappearing messages feature to present him with illicit "drug menus." Chapman emphasizes that Snapchat's features, such as location sharing and friend recommendations, enabled drug dealers to specifically target his son. <https://www.latimes.com/politics/story/2023-08-09/meta-instagram-twitter-tiktok-social-media-onlinesafety>
8. Alexandra Martin, a 19-year-old from Kentucky was introduced to Instagram at just 12 years old. Alexandra was exposed to harmful content that contributed to the development of anxiety, depression, and ultimately anorexia, resulting in hospitalization and suicide attempts. Despite being underage, the lawsuit claims Instagram actively encouraged her to open multiple accounts, compounding the harmful effects of its algorithms. <https://abcnews.go.com/GMA/News/instagram-eating-disorders-depression-young-girls-lawsuits-claim/story?id=87418473>
9. Lalani Erika Renee Walton, 8, of Temple, Texas, and Arriani Jaileen Arroyo, 9, of Milwaukee died while participating in the dangerous "blackout challenge" on TikTok, which encouraged users to choke themselves until they lost consciousness. The parents are suing TikTok, alleging that the platform enticed young users, failed to warn them of the risks, and did not do enough to prevent the dissemination of dangerous challenges. <https://www.nbcnews.com/tech/parents-sue-tiktok-deaths-two-girls-blackout-challenge-rcna37100> & <https://www.nytimes.com/2022/07/06/technology/tiktok-blackout-challenge-deaths.html>

News Summaries

Mental Health Harms and Loss of Life Experienced by Young People Due to Social Media Platform Design Practices

10. Erik Robinson, aged 12, and Garrett Pope, aged 11, tragically passed away from accidental asphyxiation while attempting to get high playing the "Choking Game." The challenge, which is connected to the death of dozens of young people, was widespread and easily accessible to youth through video tutorials on platforms like YouTube. <https://time.com/5189584/choking-game-pass-out-challenge/>

11. In 2015, Christopher James (CJ) Dawley, died by suicide at the age of 17 after struggling with social media addiction. CJ's parents allege that his excessive time on Facebook, Instagram, and Snapchat led to sleep deprivation, obsession with body image, and ultimately, his untimely death. They have filed a lawsuit against Snap and Meta. <https://www.cnn.com/2022/04/19/tech/social-media-lawsuits-teen-suicide/index.html>

12. In 2020, Annalee Schott tragically took her own life at the age of 18. Digging through Annalee's journals and TikTok account, her mother Lori discovered numerous videos where young individuals glamorized self-harm and self-hate, many of which garnered significant likes. Motivated by Annalee's passing, Lori became a strong advocate for the Kids Online Safety Act (KOSA). <https://www.denver7.com/news/360/our-children-deserve-better-colorado-mom-pushes-for-online-regulation-after-losing-daughter-to-suicide>

LEGISLATURE

HB/SB No. XXXX

Model Legislation: Social Media Algorithm Accountability Act

Model Legislation to promote social media platform transparency and accountability with regard to how use of these platforms affects the mental and physical health of child users in this state.

Referred to Committee On: _____

Introduced by:_____

Section 1: Purpose.

Sections 1 to 8, inclusive, shall be known, and may be cited, as the “(State Name) Social Media Algorithm Accountability Act.” The purpose of this Act is to promote social media platform transparency and accountability with regard to how use of these platforms affects the mental and physical health of child users in this state.

Section 2: Definitions.

As used in this Chapter, the following words and terms shall have the following meanings:

- (a) “Algorithm” means a computational process that uses machine learning, natural language processing, artificial intelligence techniques, or other computational processing techniques of similar or greater complexity and that makes a decision or facilitates human decision-making with respect to users’ personal information, including to determine the provision of products or services or to rank, order, promote, recommend, amplify, or similarly determine the delivery or display of information to an individual. For purposes of this Act, an algorithm will refer to recommendation algorithms, also known as engagement-based algorithms, which passively populate a social media user’s feed with content without any direct action or request by the user.
- (b) “Child” or “children,” means a consumer or consumers who are under 18 years of age.
- (c) “Covered platform” means a social media platform that conducts business in this state or that produces products or services that are targeted to residents of this state and that during the preceding calendar year: (1) Controlled or processed the personal information of not less than one

Model Legislation: Social Media Algorithm Accountability Act

hundred thousand consumers, excluding personal information controlled or processed solely for the purpose of completing a payment transaction; or (2) controlled or processed the personal information of not less than twenty-five thousand consumers and derived more than twenty-five per cent of their gross revenue from the sale of personal information.

(d) "Consumer" means a natural person who is a (State Name) resident, however identified, including by any unique identifier.

(e) "Design-related Harms to Children" means a covered platform's product, service, or feature design that would result in a reasonably foreseeable risk of:

1. Consistent with evidence-informed medical information, the following mental health disorders: anxiety, depression, eating disorders, substance abuse disorders, and suicidal behaviors.
2. Patterns of use that indicate or encourage addiction-like behaviors in children.
3. Physical violence, online bullying, and harassment of children.
4. Sexual exploitation and abuse of children.
5. Promotion and marketing of narcotic drugs (as defined in section 102 of the Controlled Substances Act (21 U.S.C. 802)), tobacco products, gambling, or alcohol to children.
6. Predatory, unfair or deceptive marketing practices, or other financial harms to children.

(f) "Experts in the mental health and public policy fields" means:

1. academic experts, health professionals, and members of civil society with expertise in mental health, substance use disorders, and the prevention of harms to minors;
2. representatives in academia and civil society with specific expertise in privacy and civil liberties;
3. youth representation;
4. representatives of the National Telecommunications and Information Administration, the National Institute of Standards and Technology, the Federal Trade Commission, the Department of Justice, and the Department of Health and Human Services;
5. State attorneys general or their designees acting in State or local government; and
6. representatives of communities of socially disadvantaged individuals (as defined in section 8 of the Small Business Act (15 U.S.C. 637)).

(g) "Independent third-party auditor" means an auditing firm that has no affiliation with a covered platform as defined by this Chapter.

(h) "Likely to be accessed" means it is reasonable to expect, based on the following factors, that a covered platform would be accessed by children:

Model Legislation: Social Media Algorithm Accountability Act

1. The covered platform is directed to children as defined by the Children's Online Privacy Protection Act (15 U.S.C. Sec. 6501 et seq.).
 2. The covered platform is determined based on audience composition where children comprise at least 8% of its audience.
 3. The covered platform is paid for by advertisements on its platform that are marketed to children.
 4. The covered platform is substantially similar or the same as a covered platform that satisfies paragraph (2).
 5. A significant amount of the audience of the covered platform, 8% or more, is determined, based on internal company research, to be children.
- (i) "Process" or "processing" means any operation or set of operations performed, whether by manual or automated means, on personal information or on sets of personal information, such as the collection, use, storage, disclosure, analysis, deletion or modification of personal information.
- (j) "Personal information" means any information that is linked or reasonably linkable to an identified or identifiable individual.
- (k) "Social media platform" means a public or semipublic internet-based service or application that has users in (State Name) and that meets both of the following criteria:
1. A substantial function of the service or application is to connect users in order to allow users to interact socially with each other within the service or application.
 2. The service or application uses recommendation algorithms to disseminate content to users.
 - A. A service or application that provides email or direct messaging services shall not be considered to meet this criterion on the basis of that function alone.
 - B. A service or application that is an internet search engine or a website whose primary purpose is e-commerce, which would include the buying, selling, or exchange of goods or services over the internet, including business-to-business, business-to-consumer, and consumer-to-consumer transactions, shall not be considered to meet this criterion on the basis of that function alone.
 3. The service or application allows users to do all of the following:
 - A. Construct a profile for purposes of signing into and using the service or application.
 - B. Populate a list of other users with whom an individual shares a social connection within the system.
 - C. Create or post content viewable by other users, including, but not limited to, on message boards, in chat rooms, or through a landing page or main feed that presents the user with content generated by other users.

Model Legislation: Social Media Algorithm Accountability Act

Section 3: Office of Social Media Transparency and Accountability.

- (a) The Office of Social Media Transparency and Accountability (hereinafter "Office") shall be created within the Office of the Attorney General to receive, review, and maintain the reports from covered platforms, to enforce the requirements of this Chapter, and to adopt regulations to clarify the requirements of this Chapter.
- (b) On or before January 31 following each year in which a social media platform meets the definition of a covered platform, as provided in this Chapter, the social media platform shall register with the Office by providing the following:
 - (1) A registration fee in an amount determined by the Office of the Attorney General, not to exceed the reasonable costs of establishing and maintaining the Office; and
 - (2) The name of the social media platform and its primary physical, email, and internet website addresses.
- (c) The Office shall by July X, 202X empanel an Advisory Council of experts in the mental health and public policy fields as defined in section 2(f) to identify the ways covered platforms' design practices potentially cause design-related harms to children.
- (d) By July X, 202X, the Office must promulgate regulations based on the cumulation of the potential design-related harms identified by the processes of subsection (c) that set forth the specific design-related harms that must be examined by the algorithm risk audits required under this Chapter.
- (e) The Office shall compile a list of approved, independent third-party auditors and be charged with assigning independent third-party auditors to conduct algorithm risk audits of covered platforms.

Section 4: Transparency Reports.

- (a) Beginning on January X, 202X covered platforms shall annually generate and submit a transparency report to the Office that contains all of the following:
 - 1. An assessment of whether the platform is likely to be accessed by children;
 - 2. A description of the covered platform's commercial interests in use of the platform by children;
 - 3. The number of individuals using the covered platform reasonably believed, based on existing data, to be children in the United States, disaggregated by the age ranges of 0-5, 6-9, 10-12, 13-15, and 16-18;
 - 4. The median and mean amounts of time spent on the platform by children in the United States who have accessed the platform during the reporting year on a daily, weekly, and monthly basis, disaggregated by the age ranges of 0-5, 6-9, 10-12, 13-15, and 16-18;

Model Legislation: Social Media Algorithm Accountability Act

5. A description of each system design feature covered platforms use to increase, sustain, or extend use of a product or service by users, including automatic playing of media, rewards for time spent, and notification delivery, and how each feature increases, sustains, or extends use;
 6. A description of each product, service, or feature of a covered platform that collects or processes personal information, for what purpose the product, service, or feature collects or processes information, and whether and how the data collection or processing may cause reasonably foreseeable risk of design-related harms to children;
 7. The total number of complaints received regarding the design-related harms described in section 2(e), disaggregated by category of harm; and
 8. A description of the mechanism by which the public may submit complaints, the internal processes for handling complaints, and any automated detection mechanisms for design-related harms to children, including the rate, timeliness, and effectiveness of responses.
- (b) The Office and the records generated by the requirements of section 4 are subject to the (State Name) Public Records Law. However, to the extent any information contained within a report required by this section is trade secret, proprietary or privileged, covered platforms may request such information be redacted from the copy of the report that is obtainable under the public records law. The Office will conduct a confidential, in-camera review of requested redactions to determine whether the information is trade secret, proprietary or privileged information that should not be made accessible for public review. All information from the copy of the report submitted to the Office, including redactions, will be maintained by a covered platform in their internal records.

Section 5: Algorithm Risk Audits.

- (a) By July X, 202X, all covered platforms must submit a preliminary report to the Office.
1. The preliminary report must be prepared by an independent third-party auditor identified in section 3(e).
 2. The Office must consult with independent third-party auditors and covered platforms to determine what data covered platforms must provide to independent third-party auditors to produce the preliminary reports.
 3. The preliminary report must describe each product, service, or feature that uses an algorithm to curate content displayed to users, the purpose for which the product, service, or feature uses an algorithm, and measure whether, how, and to what extent the covered platform's algorithmic design may cause reasonably foreseeable risk of design-related harms to children as identified in section 2(e) of this Chapter.
- (b) After a covered platform has submitted a preliminary report, the covered platform may agree that the Office will consult with independent third-party auditors and the covered platform to set benchmarks the covered platform must meet to reduce the design-related harms, identified in

Model Legislation: Social Media Algorithm Accountability Act

section 2(e) of this Chapter, on its platform as indicated in the preliminary report required under subsection (a) of this section.

- 1.Upon agreement, each covered platform shall thereafter use an independent third-party auditor to produce biannual reports detailing the following:
 - A. Steps taken to mitigate design-related harm on its platform, including implementation of any systems used to meet benchmarks; and
 - B. Measurements indicating the reduction in design-related harm as a result of these systems.
 2. In the case the covered platform has failed to meet the benchmarks, upon agreement its biannual report must also include:
 - A. A mitigation plan detailing changes the platform intends to take to ensure future compliance with benchmarks; and
 - B. A written explanation regarding the reasons the benchmarks were not met.
- (c) The Office and the records generated by the requirements of section 4 are subject to the (State Name) Public Records Law. However, to the extent any information contained within a report required by this section is trade secret, proprietary or privileged, covered platforms may request such information be redacted from the copy of the report that is obtainable under the public records law. The Office will conduct a confidential, in-camera review of requested redactions to determine whether the information is trade secret, proprietary or privileged information that should not be made accessible for public review. All information from the copy of the report submitted to the Office, including redactions, will be maintained by a covered platform in their internal records.

Section 6: Other Remedies.

If a covered platform should choose not to consult with independent third-party auditors to set benchmarks it must meet to reduce the design-related harms, identified in section 2(e) of this Act, on its platform as indicated in the preliminary reports required under section 5(a), an attorney general is not precluded from pursuing any other legal remedy available at law to mitigate harms.

Section 7: Enforcement.

- (a) A covered platform that violates the provisions of this Chapter shall be subject to an injunction and liable for a civil penalty not to exceed twenty-five thousand dollars (\$25,000) per violation, which shall be assessed and recovered in a civil action brought in the name of the people of the State of (State Name) by the Attorney General.
- (b) A covered platform shall be considered in violation of the provisions of this Chapter for any of the following:

Model Legislation: Social Media Algorithm Accountability Act

1. Fails to register with the Office as required by Section 3.
 2. Materially omits or misrepresents required information in a report submitted pursuant to Sections 4 and 5.
 3. Fails to timely submit to the Office a report required pursuant to Sections 4 and 5.
- (c) In assessing the amount of a civil penalty pursuant to this section, the court shall consider whether the covered platform made a reasonable, good faith attempt to comply with the provisions of this Chapter.
- (d) Any penalties, fees, and expenses recovered in an action brought under this Chapter shall be collected by the Office of the Attorney General with the intent that they be used to fully offset costs in connection with the enforcement of this Chapter and to promote the positive mental health outcomes of the children of (State Name).

Section 8: Severability.

If any provision of this Act, or any application of such provision to any person or circumstance, is held to be unconstitutional, the remainder of this Act and the application of this Act to any other person or circumstance shall not be affected.

References

1. Brignull H, Leiser M, Santos C, Doshi K. *Deceptive patterns – user interfaces designed to trick you.* deceptive.design. 2023. Accessed Apr. 25, 2023. <https://www.deceptive.design/>
2. European Data Protection Board. *Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: How to recognise and avoid them.* Feb. 14, 2023. Accessed Nov. 1, 2023. https://edpb.europa.eu/system/files/2023-02/edpb_03-2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf
3. Keyes KM, Gary D, O'Malley PM, Hamilton A, Schulenberg J. Recent increases in depressive symptoms among U.S. adolescents: Trends from 1991 to 2018. *Social Psychiatry and Psychiatric Epidemiology.* 2019; 54(8): 987-996.
4. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science.* 2018; 6(1): 3-17.
5. US Department of Health and Human Services/Centers for Disease Control and Prevention. *CDC Morbidity and Mortality Weekly Report.* 71(1): 230-282.
6. American Academy of Pediatrics. *AAP-AACAP-CHA declaration of a national emergency in child and adolescent mental health.* Oct. 19, 2021. Accessed Mar. 2, 2024. <https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/>
7. Anderson M, Faverio M, Gottfried J. Teens, social media and technology 2023. Pew Research Center. Dec. 11, 2023. Accessed Mar. 2, 2024. <https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/>
8. Rideout V, Peebles A, Mann S, Robb MB. Common Sense census: Media use by tweens and teens, 2021. Common Sense Media. Mar. 9, 2021. Accessed Mar. 2, 2024. <https://www.commonSenseMedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021>
9. Pedalino F, Camerini AL. Instagram use and body dissatisfaction: The mediating role of upward social comparison with peers and influences among young females. *International Journal of Environmental Research and Public Health.* 2022; 19(3):1543.
10. Brown Z, Tiggemann M. Attractive celebrity and peer images on Instagram: Effect on women's mood and body image. *Body Image.* 2016; 19:37-43.
11. Jiotsa B, Naccache B, Duval M, Rocher B, Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness. *International Journal of Environmental Research and Public Health.* 2021;18(6):2880.
12. Tiggemann M. Digital modification and body image on social media: Disclaimer labels, captions, hashtags, and comments. *Body Image.* 2022; 41:172-180.

13. Papageorgiou A, Fisher C, Cross D. "Why don't I look like her?" How adolescent girls view social media and its connection to body image. *BMC Women's Health*. 2022; 22(1): 261.
14. Vandenbosch L, Fardouly J, Tiggemann M. Social media and body image: Recent trends and future directions. *Current Opinion in Psychology*. 2022; 45:101289.
15. Herriman Z, Taylor AM, Roberts RM. Interventions to Reduce the Negative Impact of Highly Visual Social networking site use on mental health outcomes: A scoping review. *Psychology of Popular Media*. 2024; 13(1):111-139.
16. Maheux AJ, Roberts SR, Nesi J, Widman L, Choukas-Bradley S. Longitudinal associations between appearance-related social media consciousness and adolescents' depressive symptoms. *Journal of Adolescence*. 2022; 94(2):264-269.
17. Volpe U, Tortorella A, Manchia M, Monteleone AM, Albert U, Monleone P. Eating disorders: What age at onset? *Psychiatry Research*. 2016; 238:225-227.
18. van Eeden AE, van Hoeken D, Hoek HW. Incidence, prevalence and mortality of anorexia nervosa and bulimia nervosa. *Current Opinion in Psychiatry*. 2021; 34(6):515-524.
19. Smink FR, van Hoeken D, Hoek HW. Epidemiology of eating disorders: incidence, prevalence and mortality rates. *Current Psychiatry Reports*. 2012; 14(4): 406-414.
20. Chesney E, Goodwin GM, Fazel S. Risks of all-cause and suicide mortality in mental disorders: A meta-review. *World Psychiatry*. 2014; 13(2): 153-60.
21. Wilksch SM, O'Shea A, Ho P, Byrne S, Wade TD. The relationship between social media use and disordered eating in adolescents. *Journal of Eating Disorders*. 2020; 53(1): 96-106.
22. Mabe AG, Forney KJ, Keel PK. Do you "like" my photo? Facebook use maintains eating disorder risk. *International Journal of Eating Disorders*. 2024; 47(5): 516-523.
23. Choukas-Bradley S, Roberts RS, Maheux AJ, Nesi J. The perfect storm: A developmental-sociocultural framework for the role of social media in adolescent girls' body image concerns and mental health. *Clinical Child and Family Psychology Review*. 2022; 25(4): 681-701.
24. Kleemans M, Daalmans S, Carbaat I, Anschutz D. Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls. *Media Psychology*. 2018; 21(1): 93-110.
25. Center for Countering Digital Hate. Deadly by design. Dec. 15, 2022. Accessed Mar. 2, 2024. https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf
26. Steinberg L. A social neuroscience perspective on adolescent risk-taking. *Developmental Review*. 2008; 28(1): 78-106.
27. Somerville LH. The teenage brain: Sensitivity to social evaluation. *Current Directions in Psychological Science*. 2013; 22(2): 121-127.
28. Sherman LE, Payton AA, Hernandez LM, Greenfield PM, Dapretto M. The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science*. 2016; 27(7): 1027-1035.

29. van der Meulen M, Veldhuis J, Braams BR, Peters S, Konijn EA, Crone EA. Brain activation upon ideal-body media exposure and peer feedback in late adolescent girls. *Cognitive, Affective & Behavioral Neuroscience*. 2017; 17(4): 712-723.
30. Casey BJ, Jones RM, Hare TA. The adolescent brain. *Annals of the New York Academy of Sciences*. 2008; 1124: 111-126.
31. Galván A. The teenage brain: Sensitivity to rewards. *Current Directions in Psychological Science*. 2013; 22(2): 88-93.
32. Sherman LE, Hernandez LM, Greenfield PM, Dapretto M. What the brain 'likes': Neural correlates of providing feedback on social media. *Social Cognitive and Affective Neuroscience*. 2018; 13(7): 699-707.
33. Park B, Hyun Han D, Roh S. Neurobiological findings related to Internet use disorders. *Psychiatry and Clinical Neurosciences*. 2017; 71(7): 467-478.
34. Ethridge P, Kujawa A, Dirks MA, Arfer KB, Kessel EM, Klein DN, Weinberg A. Neural responses to social and monetary reward in early adolescence and emerging adulthood. *Psychophysiology*. 2017; 54(12), 1786-1799.
35. Wells G, Horwitz J, Seetharaman, D. Facebook knows Instagram is toxic for teen girls, company documents show. The Wall Street Journal. Sep. 14, 2021. Accessed Mar. 2, 2024. <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739>
36. Bejar A. Written Testimony of Arturo Bejar before the Subcommittee on Privacy, Technology, and the Law. Nov. 7, 2023. Accessed Mar. 2, 2024. https://www.judiciary.senate.gov/imo/media/doc/2023-11-07_-testimony_-bejar.pdf
37. Orben A, Przybylski AK, Blakemore SJ, Kievit RA. Windows of developmental sensitivity to social media. *Nature Communications*. 2022; 13(1): 1649.
38. Auxier B, Anderson M. *Social media use in 2021*. Pew Research Center. Apr. 7, 2021. Accessed Mar. 2, 2024. <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021>
39. Costello C, McNeil D, Binder R. Adolescents and social media: Privacy, brain development, and the law. *The Journal of the American Academy of Psychiatry and the Law*. 2016; 44(3): 313-321.
40. Amanda Raffoul, Zachary J. Ward, Monique Santoso, Jill R. Kavanaugh & S. Bryn Austin. Social media platforms generate billions of dollars in revenue from U.S. youth: Findings from a simulated revenue model. *PLOS ONE*. 2023; 18(12): e0295337.
41. Horowitz J. Meta designed products to capitalize on teen vulnerabilities, states allege. Wall Street Journal. Nov. 25, 2023. Accessed Mar. 2, 2024. <https://www.wsj.com/business/media/meta-designed-products-to-capitalize-on-teen-vulnerabilities-states-allege-6791dad5>
42. Unfair Trade Practices, Subchapter 25: Automated Employment Decision Tools, N.Y. LAW § 20-871. New York City Code. Accessed Mar. 2, 2024. <https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCadmin/0-0-0-135843>
43. Costello N, Sutton R, Jones M, Almassian M, Raffoul A, Ojumu O, Salvia M, Santoso M, Kavanaugh JR, Austin SB. Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media. *American Journal of Law & Medicine*. 2023; 49(2-3): 135-172.

44. The United States Attorney's Office for the Southern District of New York. United States attorney resolves groundbreaking suit against Meta Platforms, Inc., formerly known as Facebook, to address discriminatory advertising for housing. U.S. Department of Justice. Jun. 21, 2022. Accessed Mar. 2, 2024. <https://www.justice.gov/usao-sdny/pr/united-states-attorney-resolves-groundbreaking-suit-against-meta-platforms-inc-formerly>

45. Roth E. Meta's new ad system addresses allegations that it enabled housing discrimination. The Verge. Jan. 9, 2023. Accessed Mar. 2, 2024. <https://www.theverge.com/2023/1/9/23547191/meta-equitable-ads-system-settlement>

46. United States District Court. United States v. Meta Platforms, Inc. Settlement Agreement at 6. No. 1:22-cv-05187 (S.D.N.Y. June 21, 2022).

47. Sukunesan S, Huynh M, Sharp G. Examining the pro-eating disorders community on Twitter via the hashtag #proana: Statistical modeling approach. *JMIR Mental Health*. 2021; 8(7): e24340.

48. Harriger JA, Evans JA, Thompson JK, Tylka T. The dangers of the rabbit hole: Reflections on social media as a portal into a distorted world of edited bodies and eating disorder risk and the role of algorithms. *Body Image*. 2022; 41: 292-297.

49. Maheshwari S. *Young TikTok users quickly encounter problematic posts, researchers say*. New York Times. Dec. 14, 2022. Accessed Mar. 2, 2024. <https://www.nytimes.com/2022/12/14/business/tiktok-safety-teens-eating-disorders-self-harm.html>

50. WSJ Staff. Inside TikTok's algorithm: A WSJ video investigation. Wall Street Journal. Jul. 21, 2021. Accessed Mar. 2, 2024. <https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477>

51. Zauderer v. Office of Disciplinary Counsel, 471 U.S. 626 (1985)