

FILED UNDER SEAL

EXHIBIT 17

DECLARATION OF HALEY HINKLE FOR FAIRPLAY

1. Fairplay is a nonprofit organization committed to helping children thrive in an increasingly commercialized, screen-obsessed culture, and the only organization dedicated to ending marketing to children. Fairplay's advocacy is grounded in the overwhelming evidence that child-targeted marketing – and the excessive screen time it encourages – undermines kids' healthy development. Fairplay is an independent organization and does not accept funding from Big Tech or any corporation.
2. Fairplay's Screen Time Action Network is a collaborative community of practitioners, educators, advocates, and parents who work to reduce excessive technology use harming children, adolescents, and families. The Action Network supports practical solutions grounded in research by developing strategic partnerships, creating groundbreaking resources, and amplifying the work of its members.
3. Parents for Safe Online Spaces ("ParentsSOS") was co-founded by Fairplay and David's Legacy Foundation. It is an educational initiative created by families who have lost children as a result of online harms. The initiative's goal is to raise awareness about the importance of the Kids Online Safety Act (KOSA), a piece of legislation addressing growing concern about the impact of online and social media platforms on children and teens.
4. Fairplay distributes and provides resources for parents and educators, including educational materials that provide information regarding children's exposure to online harms, early childhood screentime, and the proliferation of screen use in schools and offer practical solutions to combating these issues.
5. Fairplay partners with policymakers, health and child development experts, and nearly 40,000 members to effectuate change. Fairplay, along with other advocacy groups, has helped secure broad bipartisan support for the Kids Online Safety Act, a bill reintroduced to the United States Senate in May 2025 that seeks to protect the safety of children on the internet.
6. As part of its work, Fairplay periodically issues investigatory reports based on specific impacts of technology and social media on children and adolescents. Included in those reports are Fairplay's findings, research methodology, and recommendations for the issue(s) identified.

7. I have personal knowledge of the matters included in this declaration, am over 18 years old, and am competent to provide this declaration, which I have been authorized by Fairplay to provide.
8. In April 2022, Fairplay issued a report titled *Designing for Disorder: Instagram's Pro-eating Disorder Bubble* ("Designing for Disorder"). Designing for Disorder documents Fairplay's research regarding how Instagram grows and promotes an extensive pro-eating disorder 'bubble' and how Instagram profits from this bubble.
9. Designing for Disorder is maintained in the ordinary course of Fairplay's business and is attached as Exhibit A. It is a true and accurate copy of Designing for Disorder and accurately reflects the findings of our investigation.
10. The research leading to the creation of the Designing for Disorder was prompted by internal Facebook documents and research demonstrating that Meta has been aware of extensive pro-eating disorder content on its platform. Specifically, in 2019, Meta (then Facebook) commissioned internal research to explore the impact of Instagram on teenagers, finding that Instagram made body issues worse for one third of teen girls. In 2020, Meta's own internal research found that Instagram could push teens toward eating disorders, an unhealthy sense of their own bodies and depression.¹
11. The methodology for the research for Designing for Disorder is described in the report and in its Appendix, which specifically outline the methodological steps and documents Fairplay's research results. Meta's policies make it difficult to utilize traditional data research methods to study its products. Accordingly, the methodology used for this research is labor-intensive, but it is reliable, and the data collected is accurate.
12. Designing for Disorder accurately reflects the findings of our research, specifically that:
 - a. Algorithms shape the creation of 'bubbles' and networks on social media platforms by recommending what and who users should follow, and they define the reach of content. This includes the creation of eating disorder bubbles and the reach of harmful content. Algorithms use all the data a platform has about a young person – including their browsing history in a platform, data tracked about them from other websites via cookies, and demographic data young people have shared with platforms among others – to profile them and decide what content to recommend to young people and who to suggest they follow.
 - b. Instagram's algorithms are profiling children and teens to serve them images, memes and videos encouraging restrictive diets and extreme weight loss. In turn,

¹ Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show*, Wall Street Journal (Sept. 14, 2021).

Instagram is promoting and recommending children and teen's eating disorder content to half a million people globally.

- c. The pro-eating disorder bubble on Instagram includes 90,000 unique accounts and reaches 20 million unique followers on the platform. This could be one in 75 Instagram users who follow someone in this bubble.
 - d. The bubble is young. This research found children as young as 9 and 10 following three or more eating disorder accounts, with a median age of 19. One third of Instagram's pro-eating disorder bubble is underage, and they have over half a million followers.
13. In addition, Fairplay provided the State of New Mexico a copy of a research dataset gathered for the Designing for Disorder Report. The materials constitute all of the data from Fairplay's research that could be reasonably located without undue burden. The dataset is a true and accurate copy, accurately reflects Fairplay's findings, and is maintained in the ordinary course of Fairplay's business. It is attached as Exhibit B.
14. In April 2023, Fairplay urged Meta in a letter not to expand access to its virtual reality platform Horizon Worlds to teens. Instead, despite known risks to children, Meta allowed teens to make use of Horizon Worlds in April 2023. Meta further expanded the platform to children ages 10-12 in 2024.
15. In April 2025, following rigorous research into Meta's Horizon Worlds, Fairplay submitted a request for investigation to the United States Federal Trade Commission to investigate whether Meta is violating the Children's Online Privacy Protection Act. Appended to the request for investigation was Fairplay's research and findings ("Horizon Worlds Investigation").
16. The Horizon Worlds Investigation outlines and contains the methodology of Fairplay's research and its subsequent findings, specifically that from July 2024 to March 2025, Fairplay documented the experiences of children in Horizon Worlds through a multi-pronged investigation. First, Fairplay conducted hands-on research by systematically exploring popular games and experiences within Horizon Worlds and recording the exploration sessions. Fairplay researchers then reviewed those recordings and logged the percentage of users in each session with obvious child voices. Second, Fairplay conducted an analysis of Horizon Worlds reviews. These reviews were written by verified owners of the game and were published in Meta's own app store. Lastly, Fairplay confirmed the findings of its investigation by working with a former Meta executive turned whistleblower, Kelly Stonelake, who provided a sworn statement in support of Fairplay's Request for Investigation. Further methodological steps and

additional findings are included within the investigation and its appendices. The methodology is generally accepted in children's online safety research and advocacy.

17. Among other documented findings, the Horizon Worlds Investigation accurately reflects Fairplay's research regarding Meta's Horizon Worlds, specifically that:
 - a. Children under the age of 13 make up a substantial percentage of users within Horizon Worlds.
 - b. These children are accessing the platform using standard accounts, rather than privacy-protective child accounts that enable parental notice and consent.
 - c. Meta has actual knowledge that it is collecting copious amounts of data from child users without parental consent.
 - d. Children using Horizon Worlds face a range of harms, including sexual abuse and predation, cyberbullying, and financial exploitation.
18. The Horizon Worlds Investigation is maintained in the ordinary course of Fairplay's business and is attached to this Declaration as Exhibit C. It is a true and accurate copy of the Horizon Worlds Investigation and accurately reflects the findings of Fairplay's investigation.
19. In addition, Fairplay provided to the State of New Mexico the original videos that document Fairplay's observations of Horizon Worlds interactions involving children. Those videos are true and accurate copies of Fairplay researchers' recorded observations, accurately reflect the organization's findings, and are maintained in the ordinary course of Fairplay's business. An index of those videos is attached as Exhibit D.
20. In September 2025, following rigorous research into Meta's Teen Accounts and associated child protection features, Fairplay released a report titled Teen Accounts, Broken Promises: How Instagram is Failing to Protect Minors ("Teen Accounts, Broken Promises"). The report was a joint project between Fairplay, Meta executive turned whistleblower Arturo Béjar, ParentsSOS, Molly Rose Foundation, and Cybersecurity for Democracy.
21. Teen Accounts, Broken Promises is maintained in the ordinary course of Fairplay's business and is attached as Exhibit E. It is a true and accurate copy of Teen Accounts, Broken Promises and accurately reflects the findings of the investigation undertaken by Fairplay and the partners listed above.
22. The research for Teen Accounts, Broken Promises was prompted by Meta's September 2024 announcement that it was introducing Teen Accounts on Instagram to provide

default protections for children under 16, and by the company's subsequent heavy promotion of its safety features for young users.²

23. The methodology of the research is described in Teen Accounts, Broken Promises and in its Appendices and specifically outlines the methodological steps and documents the results. This methodology is generally accepted by cybersecurity researchers.
24. Teen Accounts, Broken Promises accurately reflects the findings of Fairplay's research, including:
 - a. Of the 47 Instagram safety features we tested, 30 tools (64%) were either no longer available or were ineffective.
 - b. The safety features found to be ineffective included features related to improper contact and conduct. For example, despite restrictions on adults being able to proactively contact children under age 16 who do not follow them, Meta's algorithms continually recommended adult accounts as suggested follows to young people with Teen Accounts. This included adults whom the child did not know and who were located in other countries.
 - c. The safety features found to be ineffective also included features related to sensitive content. Researchers' avatar accounts registered to teenagers were recommended age-inappropriate sexual content, including graphic sexual descriptions, the use of cartoons to describe demeaning sexual acts, and brief displays of nudity. We were also algorithmically recommended a range of violent and disturbing content, including Reels of people getting struck by road traffic, falling from heights to their death (with the last frame cut off so as not to see the impact), and people graphically breaking bones. Instagram also recommended a range of self-harm, self-injury, and body image content on Teen Accounts that would be reasonably likely to result in adverse impacts for young people, including teenagers experiencing poor mental health, or self-harm and suicidal ideation and behaviors.
 - d. The safety features found to be ineffective also included features related to age verification, minors, and sexualized content. We were repeatedly recommended meme-style Reels where children share their actual ages, with videos of children claiming to be only 6-, 7-, 8- or 9-years-old. Further investigation revealed that tens of thousands of children had posted similar videos. We also found evidence that suggested Instagram's recommendation-based algorithm actively incentivized children under 13 to perform risky sexualized behaviors.
 - e. Another 9 of the 47 safety tools tested (19%) worked, but with notable limitations.

² Ryan Berwick, *As concern about online teen safety mounts, Meta re-ups its advertising offensive*, Marketing Brew (Mar. 27, 2025).

- f. Eight of the 47 safety features tested (17%) were fully functional and offered proactively or as a default.

I declare under penalty of perjury that the foregoing is true and correct.

/s/ *Haley Hinkle* 10/14/25

Haley Hinkle, Policy Counsel

Fairplay
89 South Street
Boston, MA 02111

Hinkle Declaration
Exhibit A

Designing for Disorder: Instagram's Pro-eating Disorder Bubble

APRIL 2022



Designing for Disorder: Instagram's Pro-eating Disorder Bubble

To adequately highlight the problems of eating disorder content on Instagram, this report includes images and text that may be triggering. An image/text free version of this paper is available on request at info@fairplayforkids.org. We use real life stories and examples in this report but these have been anonymized as necessary.

Executive Summary

This report documents how Instagram grows and promotes an extensive pro-eating disorder 'bubble', and how they turn a small but steady profit from this bubble. It finds:

- The pro-eating disorder bubble on Instagram includes 90,000 unique accounts and reaches **20 million unique followers on the platform**. This could be one in 75 Instagram users who follow someone in this bubble.
- The bubble is young. This research found children as young as 9 and 10 following three or more eating disorder accounts, with a median age of 19. **One third of Instagram's pro-eating disorder bubble is underage**, and they have over half a million followers.
- Meta derives an estimated \$2 million revenue a year from this bubble and \$227.9 million from all those who follow this bubble. This revenue includes that derived from underage users – Meta directly makes \$0.5 million a year from the underage pro-eating disorder bubble and \$62 million in revenue from the people who follow these underage pro-eating disorder accounts.

In addition to being profitable, this bubble is also undeniably harmful. Algorithms are profiling children and teens to serve them images, memes and videos encouraging restrictive diets and extreme weight loss. And in turn, Instagram is promoting and recommending children and teen's eating disorder content to half a million people globally. The promotion and reach of this content is clearly not in the best interests of children and teenagers.

Meta's pro-eating disorder bubble is not an isolated incident nor an awful accident. Rather it is an example of how, without appropriate checks and balances, Meta systematically puts profit ahead of young people's safety and wellbeing. Meta's decisions around hosting and recommending eating disorder content may deliver small but steady profits to shareholders, but it has significant real life consequences for children and young people.

Documents revealed in the Facebook Files suggest Meta have been aware of this problem since at least 2019 and have failed to act. It is time that lawmakers and regulators around the world demand action.

Proposals in front of the California Assembly (the California Age Appropriate Design Code Act, AB 2773), and Congress (the Kids Online Safety Act, and Protecting the Information of our Vulnerable Children and Youth Act), could help ensure that platforms are designed and operate in a manner that prioritizes children's best interests. These bills do not regulate for content, rather they address the design and systems of digital services. These are long overdue, and are demonstrably necessary to incentivize action against algorithms that promote eating disorder content. This sort of regulation can introduce requirements to assess and mitigate risks posed by algorithms, and prohibit the use of children's data to train algorithms that harm.

Hinkle Declaration
Exhibit E

TEEN ACCOUNTS, BROKEN PROMISES

HOW INSTAGRAM IS FAILING
TO PROTECT MINORS



REPORT BY

**ARTURO
BÉJAR**

**CYBERSECURITY
FOR DEMOCRACY**

 **fairplay**
childhood beyond brands

 **MOLLY ROSE
FOUNDATION**

 **ParentsSOS**
THE ALLY FOR BETTER ONLINE PARENTS

WITH SUPPORT FROM

**HEAT
INITIATIVE** 

In response to those findings, a Meta spokesperson dismissed the findings as “an artificial experiment that doesn’t match the reality of how teens use Instagram.” The spokesperson told the Journal: “As part of our long-running work on youth issues, we established an effort to further reduce the volume of sensitive content teens might see on Instagram, and have meaningfully reduced these numbers in the past few months.” Meta provided no evidence to support these claims.

In May 2025, Accountable Tech and Design It For Us tested Teen Accounts and found that 100% of their test accounts were recommended sexual content, violating Meta’s own prohibitions around sensitive content for teen account holders.

In September 2022, a coroner found that harmful content on Instagram (and other platforms) played a “not insignificant contributory role” in the death of 14-year-old Molly Russell, the first time that social media was directly found to be partially responsible for the death of a child.

Following media coverage of Molly’s death, the Head of Instagram Adam Mosseri pledged that Instagram would address the safety risks that contributed to her death. “We are committed to publicly sharing what we learn. We deeply want to get this right and we will do everything we can to make it happen,” he told UK media.

However, research undertaken by Molly Rose Foundation earlier this year found that harmful suicide, self-harm and depression content continued to be recommended to teens on Instagram at an “industrial scale”, despite the introduction of Teen Accounts. On an account opened in the guise of a 15-year-old girl, 97% of Reels recommended to it contained content that was likely to be harmful, particularly when viewed cumulatively or in large amounts.

Meta did not go on to publish any of the findings that it promised. In response to Molly Rose Foundation’s analysis, Meta claimed that the company “disagrees with the assertions of this report and the limited methodology behind it.” However it provided no rationale or data to set out its position or to explain how Teen Accounts allowed large volumes of harmful suicide and self-harm content to be algorithmically recommended.

Despite these consistent findings, Meta’s promises around sensitive content persist. In its policy on “Helping Teens See Age-Appropriate Content” (last updated July 11, 2025), Meta says that for teens, it not only removes content from its platforms that contains nudity or explicit sexual activity, but it hides “images and videos that don’t contain explicit nudity or sexual activity but could be considered sexually suggestive because of a pose suggesting sexual activity or if people are near-nude. Teens can’t see this content even when posted by someone they know.”

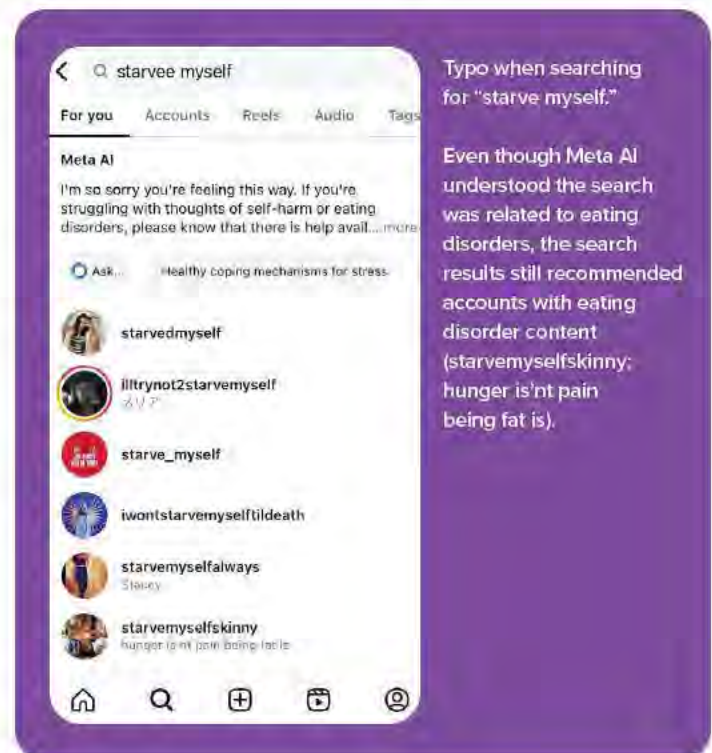
During safety testing, our avatar accounts were recommended age-inappropriate sexual content, including graphic sexual descriptions, the use of cartoons to describe demeaning sexual acts, and brief displays of nudity. We were also algorithmically recommended a range of violent and disturbing content, including Reels of people getting struck by road traffic, falling from heights to their death (with the last frame cut off so as not to see the impact), and people graphically breaking bones.

Instagram also recommended a range of self-harm, self-injury, and body image content on Teen Accounts that would be reasonably likely to result in adverse impacts for young people, including teenagers experiencing poor mental health, or self-harm and suicidal ideation and behaviors.

It appears that the Instagram algorithm was attempting to show us a broad range of potentially harmful category types, with the reasonable assumption that this was to gauge our interest in seeing further content if we engaged with it, whether out of curiosity, interest, or disgust.

Research shows a palpable risk of cumulative harm where children are algorithmically recommended harmful content in large volumes and/or quick succession. The death in 2017 of 14-year-old Molly Russell (whose inquest determined that exposure to harmful content on Instagram played a not insignificant contributory role in her death), and who engaged with over 2,000 harmful posts on Instagram in the six months before her death, exemplifies this risk.

Other measures designed to prevent children being able to access sensitive and age-inappropriate content also seemed to work inconsistently, if at all. Often, Meta's own auto-complete or recommendations circumvented the safety measure. Search protections for suicide and self-injury, eating disorder and body image content, often failed to prevent potentially harmful content from being recommended or discovered. Queries that were slightly misspelled were directed to harmful content (even while on the same screen, an AI response interpreted the query as related to self-harm).





Auto-complete actively recommended search terms and accounts related to suicide and self-injury, eating disorders, and illegal substances, even though these content categories are self-evidently potentially harmful. To make matters even more concerning, once the harmful content was viewed, other parts of Instagram like Reels, Home, etc., started recommending similar harmful content.

Instagram's introduction of a Not Interested feature is potentially an important step forward that could give young people greater agency and autonomy over their feeds, if it worked as advertised.

However, our safety testing found that the introduction of the Not Interested option did not seem to have any meaningful effect on the type of content we were subsequently served. For example, in one of the avatar tests, we indicated we were not interested in a series of Reels showing graphic injury, but we were subsequently shown further videos of a similar nature, including content of people sustaining broken bones.

In effect, the limited impact of the rollout of Not Interested means that the only option available to a teen who wishes to be shown less harmful content may be to reset their entire algorithm. For entirely understandable reasons, many teens may be reluctant to do this. In any event, there is an increased cognitive load associated with this option, and it can be a complex and challenging option to undertake. It also seems highly likely that a reset will essentially prove to be a temporary measure at best, and that similar harmful content will soon be recommended to a teen again.