List of Social Media Harms

Social media platforms, apps, and freemium games are designed to exploit attention and extract individualized data for the sake of advertising -- it is central to their business models. From January 1 to Sept. 30, 2021, Facebook made $82.3 billion from advertising.\(^1\) During the same period, YouTube made $20.2 billion from advertising.\(^2\) Roblox made $1.35 billion, including advertising and in-game purchases, and users, the majority of whom are under 16, spent 30.6 billion hours playing.\(^3\) Due to lack of regulation, profit derived from attention is the foremost motivator of design. How does this impact our kids and teens?

## Manipulative Design Features

The impact of these features is not universally negative, but that does not negate their inherently manipulative purpose, which is the promotion of user engagement.

<table>
<thead>
<tr>
<th>Design Feature</th>
<th>Outcomes</th>
<th>Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likes, emoji reactions, comments, views</td>
<td>Addiction; social comparison; body image issues</td>
<td>Facebook; Instagram; TikTok; YouTube; Discord; Twitter; Twitch</td>
</tr>
<tr>
<td>Autoplay and infinite scroll</td>
<td>Addiction; fear of missing out (FOMO)</td>
<td>Facebook; Instagram; TikTok; YouTube; Twitch</td>
</tr>
<tr>
<td>Push notifications</td>
<td>Addiction; social pressure; distraction; FOMO</td>
<td>Facebook; Instagram; TikTok; Twitter; Roblox; Snapchat; Discord; Twitch; messaging apps</td>
</tr>
<tr>
<td>&quot;Beautifying&quot; filters</td>
<td>Social comparison; body image issues</td>
<td>Snapchat; Instagram; TikTok</td>
</tr>
<tr>
<td>Gamification of social interactions</td>
<td>Addiction; social comparison; social pressure</td>
<td>Snapchat; Reddit; Twitch; TikTok</td>
</tr>
<tr>
<td>Ephemeral content</td>
<td>Addiction; FOMO</td>
<td>Snapchat; Instagram; Facebook; TikTok (pilot of stories)</td>
</tr>
<tr>
<td>Marking messages as read or unread</td>
<td>Social pressure; distraction</td>
<td>Snapchat; Facebook; Instagram; messaging apps</td>
</tr>
<tr>
<td>Virtual currencies and in-app purchases</td>
<td>Financial exploitation; social comparison; addiction</td>
<td>TikTok; Roblox; Twitch; Instagram</td>
</tr>
</tbody>
</table>

---

Opaque Recommendation Algorithms

Recommendation algorithms are difficult to study externally due to lack of transparency. Below, we discuss a selection of studies and cases that illustrate the harms caused by algorithms. Leaked internal Facebook research features prominently due to the unique research insights provided.⁴

**Prioritization of divisive, hateful, sensationalist content**
- Facebook's newsfeed ranking algorithm, which was altered in 2018 to ostensibly boost "meaningful interactions," actually heavily prioritized emoji reactions, long comments, and reshares of posts, resulting in the promotion of controversial and emotion-inducing content.⁵ News organizations and political parties noticed this change and felt pressured to "rely more on negativity and sensationalism." Facebook researchers found increased promotion of "civic misinfo, civic toxicity, health misinfo, and health antivax content."⁶ Internal researchers proposed solutions that were rejected by Mark Zuckerberg due to concerns that user engagement would decrease.⁷

**Promotion of negative social comparison**
- According to a 2020 Facebook study of 50,590 people in 10 countries, 33% of people compared their appearance to others on Instagram and 26% always or often saw content that made them feel worse about their appearance.⁸ It's worse for teen girls: 48% compared appearance, 37% felt worse about themselves, and 34% felt “a lot” or “extreme” pressure to look perfect. In another 2020 Facebook study, 40% of teen boys experienced negative social comparison due to Instagram, but body image contributed less than other factors like economic status.⁹

---

A 2019 study, focusing on young women, found frequency of Instagram use to be correlated with depressive symptoms, decreased self-esteem, general and physical appearance anxiety, and body dissatisfaction. Additionally, exposure to beauty and fitness images, in particular, significantly decreased self-rated attractiveness.

A 2020 study found that just seven minutes on Instagram, compared to Facebook or playing a game, led to decreased body satisfaction and increased negative emotions among college-age women.

Content moderation guidelines for TikTok, which were leaked in 2020, instructed moderators to suppress content featuring fatness, wrinkles, "abnormal body shape," "ugly facial looks," and poverty, among other categories.

**Promotion of disordered body image content**

- Communities that encourage disordered eating use coded language to evade platform filters and to share memes, "thinspiration" photos, goal weights, and tips on Twitter, TikTok, Instagram, Facebook, YouTube, Discord, Tumblr, and Snapchat. Instagram, in particular, was shown to recommend search terms like "appetite suppressant" to people with eating disorders.

- Several investigations, including by Sen. Blumenthal’s office and CNN, have created fake youth accounts on Instagram that like a few dieting posts and then, within a day or two, are shown majority pro-anorexia content.

- A 2019 study of 12- to 13-year-olds in Australia found that girls with Snapchat, Instagram, and Tumblr accounts and boys with Snapchat, Facebook, and Instagram were significantly more likely to report disordered eating than youths not using social media. Greater daily time spent using Instagram and Snapchat was associated with significantly higher disordered eating behaviors among girls.

---


- YouTube, Instagram, and Facebook are central to the encouragement and sale of steroids, particularly to boys, pushed through algorithms recommending content, ads, groups to join, and accounts to follow.\(^{17}\)

**Promotion of self-harm and suicide content**
- Internal Facebook research revealed that among teens who reported suicidal thoughts, 13% of British users and 6% of American users traced the desire to kill themselves to Instagram.\(^{18}\)
- A 2018 study found that Instagram posts that mentioned suicide ideation elicited higher engagement than posts that did not.\(^{19}\)
- A 2019 study found that 43% of young adults surveyed have seen self-harm content on Instagram and, of those, 80% did not seek it out (recommended or accidental).\(^{20}\) The content led 39% to think about how it would feel to harm themselves. Surprisingly, 32.5% indicated that they have performed the same self-harming behavior due to seeing self-harm content on Instagram. This correlated significantly with suicide ideation and suicide risk.

**Promotion of extremist content**
- A 2020 study of YouTube found that 9% of users watched extremist or white supremacist videos and 22% viewed "alternative" fringe content.\(^{21}\) During the two months studied, the mean numbers of videos watched were 11.5 (extremist) and 64.2 (alternative). Despite YouTube's recent algorithmic changes, 37.6% of recommendations on alternative videos and 29.3% of recommendations on videos from extremist channels were to other videos of the same type.
- A 2021 study of TikTok found that 30% of videos promoted white supremacy, 24% supported terrorists, 14.9% were antisemitic, 13.5% anti-Black, 8.7% anti-LGBTQ, and 7.9% anti-Muslim.\(^{22}\) This follows 2020 findings that 14 QAnon-related hashtags had over 488 million combined views on TikTok.\(^{23}\)

---


Inadequate Content Moderation

Hate speech
- By internal measures, Facebook’s hate speech algorithm prevents only 3–5% of hate speech views and 0.6% of all content that violates policies against violence and incitement.24 Publicly, Mark Zuckerberg had claimed 94% removal.25
- Facebook knew that its hate speech algorithm disproportionately protected White people and men, and inadequately protected those who are Black, Jewish, Muslim, and LGBTQ, but didn't prioritize changing the algorithm, resulting in greater harms and increased prevalence of racist language.26
- Facebook spends 84% of its misinformation budget on the U.S.27 and does not have translators for the majority of languages,28 so hate content is worse abroad.

Sexual abuse online
- In a 2020 study, 25% of U.S. 9- to 17-year-olds reported having a sexual encounter online with someone who they believed to be an adult.29 Of youths 9–12 years old, 19% had interacted with an adult sexually online. 55% of youths were recontacted by individuals they had blocked and/or reported (45% on the same platform under a new account, 43% on a different platform). 83% of kids will not tell a trusted adult about abuse online. 94% will not tell a trusted adult about unsolicited nudes they have received from adults online.
- There was a 77% increase in "self-generated" child sexual content from 2019 to 2020.30 In 80% of the cases, victims were 11- to 13-year-old girls.
- Most companies use tools to detect child sexual abuse material (87% use image "hash-matching"); only 37% currently use tools to detect the online grooming of children.

---