Synthetic Images; Real Feelings

Generative AI and Teen Well-Being

NYU | CENTER ON TECH POLICY

Scott Babwah Brennen

About the Center on Technology Policy

NYU's Center on Technology Policy seeks to craft public policy for a better internet. Utilizing an interdisciplinary academic framework, CTP works to identify knowledge gaps and develop actionable policy frameworks that will enable us to realize the potential benefits of technology while minimizing its harms. By working closely with students and expanding the University's offerings in technology policy analysis, we seek to cultivate and train the field's future practitioners.

About the Author

Scott Babwah Brennen is the director of the Center on Technology Policy at NYU.

This report was designed by Adrienne Testa.

Acknowledgements

I would like to thank Zeve Sanderson, John Perino, Matt Perault, Aliya Bhatia, Amanda Reid, and Bonnie Brennen for their advice and feedback on the report.

Executive Summary

Last November, a woman in Orlando filed a lawsuit against the AI chatbot company, Character.AI, claiming that one of the company's models had encouraged her son to take his own life. Amid the resulting media firestorm from this and a similar lawsuit filed in Texas, regulators in several states have introduced bills that would impose new requirements for chatbots that interact with children, including, as one New York state bill stipulates, that any "AI companion" must announce in at least 16pt font and capital letters at the beginning of every interaction with either a minor or adult:

"THE AI COMPANION [OR NAME OF THE AI COMPANION] IS A COMPUTER PROGRAM AND NOT A HUMAN BEING. IT IS UNABLE TO FEEL HUMAN EMOTION"

As parents, policymakers, and platforms all work to balance the benefits and risks that digital life poses to adolescents, online child safety on social media platforms has become one of the most prominent technology-related issues for state and federal policymakers. Yet, even as generative AI disrupts many sectors, there has been less attention to how the growing adoption of generative AI on social media may Impact — for better or worse — the psychological well-being of adolescents.

This report examines the impact that the incorporation of generative AI tools on social media may have on the psychological well-being of adolescents. In addition to reviewing the nascent literature addressing this issue, the report analyzes what we know about the mechanisms that connect social media use and psychological well-being in general, offering a series of predictions for how generative AI may improve or harm adolescent well-being.

Specifically, the report concludes:

1. Despite growing public concern, existing empirical literature suggests the increasing prominence of generative AI on social media platforms presents both real benefits as well as real risks to adolescent's psychological well-being.
2. Generative AI may improve adolescent well-being by providing or connecting users to others who can provide social support, while allowing users to create and communicate rich and complex identities and self-conceptions. But it may worsen adolescent well-being by exacerbating upward social comparison, providing vulnerable users poor emotional support, displacing sleep and other healthy behaviors, and increasing the likelihood that teens engage in both risky posting behavior and risky offline behavior.
3. This dichotomy offers an opportunity for policymakers and platforms to minimize the harm and maximize the benefits of generative AI on social media platforms.

Recommendations

Given these findings, we offer ten recommendations to help policymakers and platforms strengthen positive impacts of generative AI on social media on adolescent well-being, lessen negative impacts, and add to our knowledge base.

To maximize the benefits of generative AI on social media:

1. Rather than a one-size-fits-all approach, platforms should personalize AI product recommendations for different users. For example, they could promote life-like bots for those who have expressed an interested or who already have strong on-network social ties while promoting human connection for those who do not.
2. Platforms should design chatbots that bridge or connect human users to help grow users' human networks.
3. Social media platforms should ensure that integrated generative AI tools are trained on diverse datasets and offer sophisticated tools users can use to craft complex and distinctive identities. They may consider delaying wide-scale integration of specific tools until they are able to deploy more advanced models that can serve a wide-range of needs.

To minimize the harms of generative AI on social media:

4. Federal lawmakers should enact comprehensive consumer data privacy protections that include strong protections for and restrictions on the collection, sale, or use of adolescents' sensitive data. In the absence of federal legislation, states should continue to pass and extend comprehensive consumer data protection laws.
5. Platforms should integrate user-focused generative AI tools directly into platforms. While this may undermine competition with third-party developers, integrated tools will help platforms address some of the most serious risks posed by widespread generative AI use.
6. When platforms know or detect that an image of a person has been substantially modified — by generative AI or other image or video editing tools — in ways likely to exacerbate misleading upward social comparison, they should disable viewership metrics, append a tested label or disclaimer, and algorithmically deprioritize or derank that image in users' feeds.
7. Platforms should continue to iterate self- and parental controls, providing teens and/or guardians features including limiting total platform use or use after certain times and supplying less engaging or calmer material later in the day.

8. State policymakers should require and fund programming in schools on the harms of both consuming and producing NCII and CSAM.

9. Policymakers, platforms, and civil society should focus additional effort addressing victims of NCII and CSAM. While new laws may punish perpetrators, they will do little to ameliorate the existing harm done to victims. For example, policymakers could earmark additional resources for school psychological counseling, training programs for school psychologists on helping victims of NCII, or other public resources for victims. Platforms could fund and host informational campaigns or connect users to psychological services.

To build our knowledge base:

10. Platforms should expand research partnerships with independent researchers and provide data and access needed to better address specific questions related to generative AI integration and psychological well-being, including: How does generative AI impact the amount and quality of time spent on social media platforms? Does it increase overall time spent? Does it impact *how* users engage on platforms?

Generative AI on Social Media

Although there is a great deal of experimentation with generative AI on social media, we recognize three common ways that these AI tools are being integrated into user experience.

User-Generated Content

Simply put, generative AI tools make it easier for users to produce or edit content. The ultimate promise of these tools is to permit users to better express themselves faster, cheaper, and with less need for editing software or high-quality cameras. Platforms are slowly rolling out new integrated features. For example, Snapchat offers paying subscribes an image generator tool and all users a tool that generates new images based on submitted selfies.

Some social media platforms also offer stand-alone tools for creators that use AI tools to create or edit content. The Tiktok Effect House permits creators to create AI-generated augmented reality content. Meta Movie Gen is a text to video AI generator–and while it is not directly integrated into Instagram, Meta is touting the ways that Instagram influencers are using the tool to create video content.

Some commentators have suggested Meta's entire approach to open weighting its Llama models not only recognizes the value of community-led iterative development, but also understands the widespread adoption of LLMs to create more (and more engaging) social media content. That is to say, Meta can essentially give away Llama because it will ultimately benefit from users spending more time on its platforms.

Advertising

While social media platforms have been slow integrating generative AI tools into user content editing or creation, some have been quicker to adopting generative AI tools for advertising. Google Ads has tools that let advertisers generate some ad assets. Meta has a suite of AI tools for both Instagram and Facebook that allow advertisers to develop creative content for ads, along with creating variations of ads.

While platforms envision that generative AI will help advertisers create high-quality ads quickly and cheaply, in permitting advertisers to create variations of ads and then test those versions in the field, these tools allow advertisers to better personalize ad content to different audience segments. Personalization is, of course, the core of algorithmic-driven feeds on social media, where platforms are constantly refining AI-based algorithms to serve users content that is personalized to them–or at least, is most likely to drive engagement.

U.S. Policy Approaches to Generative AI and Kids on Social Media

The regulation of artificial intelligence and of children's safety online have, arguably, been the two most prominent technology policy priorities for US policymakers over the last several years. And while there are increasing efforts by regulators to address the risks posed by AI to kids, these efforts remain nascent.

Rather than reviewing the landscape of state and federal AI and child online safety regulation broadly — which others have done — this section looks narrowly at how regulators are attempting to regulate the intersection of AI and child online safety.

Broadly speaking, we recognize five main ways that regulators are looking to regulate child AI safety: restricting AI-generated abusive content, limiting algorithmic curation, imposing data collection and use restrictions, regulating chatbots, and requiring risk or impact assessments.

Sexual Content and Abuse

Recently, dozens of states have introduced and passed bills that expand existing CSAM laws to include content created or modified by artificial intelligence. For example, the law passed in New Hampshire last year revises the existing statute to include "something which is portrayed to be a person under the age of 18 and therefore indistinguishable from a child." The law passed last year in South Dakota revises the previous law to include that which depicts: "An individual indistinguishable from an actual minor created by the use of artificial intelligence or other computer technology capable of processing and interpreting specific data inputs to create a visual depiction."

While not specifically referencing children, a handful of other states have outlawed the creation and dissemination of non-consensual AI-created intimate imagery (NCII). Some laws frame this as "revenge porn," with others taking a slightly broader frame. In April, federal legislators passed the TAKE IT DOWN Act, which will impose criminal penalties for the intentional distribution of non-consensual sharing of intimate visual imagery — both AI generated and not. While the law applies to both minors and adults, the penalties are worse if minors are harmed. Notably, the law also requires platforms to remove such content when notified that it exists. Violations will be prosecuted as unfair and deceptive practices under Section 5 of the FTC Act. The DEFIANCE Act of 2024, introduced by Sen. Richard Durbin (D-IL) and passed by the U.S. Senate in July 2024, would amend existing federal law to permit victims of non-consensual deep fake sexual imagery to bring civil actions against creators of that content.

In addition, last year, the Child Exploitation and Artificial Intelligence Expert Commission Act, sponsored by a bipartisan group of 22 House members, would have created a new commission to "investigate and assess how artificial intelligence may be used in the commission of a child exploitation crime." The bill, however, did not pass.

Algorithmic Curation

In 2024, legislators enacted laws in California, New York, and Utah that restrict platforms from serving algorithmically curated feeds to minors

(parents can opt-in to such systems for their children under the laws in New York and Utah). Legislators in several states have introduced similar bills into 2025 legislative sessions, but as of publication none has passed.

Similarly, one of the most prominent federal bills on online child safety, the Kids off Social Media Act (KOSMA), would also prohibit platforms from using recommendation algorithms to serve content to kids under 17. More broadly, the law would prohibit social media platforms from permitting children under 13 to create or maintain social media accounts.

Data collection limitations

In 2024 and 2025, several states have proposed amending existing data privacy laws to increase the protections for minors to better cover AI-related data use. For example, a bill in California would restrict companies from collecting minor's data for AI training without parental consent. Similarly, a bill introduced in New York this year, prohibits the creation of a digital likeness of a minor without explicit consent.

Chatbots

In response to a lawsuit filed last year against Character.AI, claiming one of the company's models was complicit in a teenager's suicide, legislators in California and Minnestoa have also introduced bills this year specifically addressing chatbots and minors. The Minnesota bill basically prohibits minors from using chatbots for "recreational purposes" (a term that is not defined), and requires covered platforms to institute age verification. The California bill attempts to prevent chatbots from employing addicting techniques toward children, such as "providing rewards to a minor user at unpredictable intervals or after an inconsistent number of actions or from encouraging increased engagement, usage, or response rates." The bill would also require annual third-party audits and platform reports including data about how many times the operator "detected exhibitions of suicidal ideation by minor users," how many times those minors "died by suicide," and how many times a "chatbot brought up suicidal ideation or action with the user."

Impact or Risk Assessments

Finally, taking a page from the Age Appropriate Design Code Laws passed in California and Maryland, the Leading Ethical Development of AI (LEAD) Act, introduced this year in California, would require pre and post-deployment assessments specifically focused on harm to children. Another bill introduced in Massachusetts would require third-party risk algorithm audits specifically considering harm to children.

Adolescent Well-being and Social Media

While interest in regulatory approaches to limiting child AI risk has grown, we still lack a detailed understanding of the potential risks that generative AI poses to children and the full range of potential benefits — on social media or off it.

Addressing this need, this report synthesizes existing literature to better understand the specific risks and benefits that AI on social media presents concerning the psychological well-being of minors. Social media is only one of the ways minors are using generative AI – perhaps not even the most common — and there are other forms of well-being besides psychological. However, there is a need for a better understanding of how this one use of generative AI mediates one form of well-being.

Below we examine how AI may impact the specific mechanisms that connect social media use and psychological well-being. However, first we offer a brief account of the state of research on social media and psychological well-being in general.

Recently, there have been several excellent research syntheses of this topic — most notably by the National Academies. Rather than duplicate that effort, we draw heavily on these existing syntheses. Ultimately, we conclude that while at the population level there is little connection between social media use and adolescent well-being, for some individuals, certain types of use is correlated with negative psychological well-being. We break this into three top-line conclusions below.

Across populations, there is no or only very small correlation between social media use and psychological well-being

Reviewing a large corpus of empirical analysis, the National Academies recently concluded that their review "did not support the conclusion that social media causes changes in adolescent health at the population level." Similarly, applying new statistical techniques to analyze three existing social data sets, Orben & Przybylski conclude that, at the population level, there is a very small negative relationship between social media and digital technology use and well-being. However, they note that in the data they review, "the association of well-being with regularly eating potatoes was nearly as negative as the association with technology use and wearing glasses was more negatively associated with well-being." That is to say, even though they saw a statistically significant correlation between using social media and well-being, the size of the relationship was very small — as small as that which can be ascribed to other things that we generally consider to be insignificant to well-being. An umbrella review — a meta review of meta reviews — concluded that "SNS use is weakly associated with higher levels of ill-being but also with higher levels of well-being."

Even when research suggests a correlation between social media use and worse well-being, determining the causal relationship can be extremely difficult. The vast majority of studies are cross-sectional, essentially taking a snapshot at a specific moment in time, rather than longitudinal, which would look across time. Cross-sectional studies can show that two things are related, but not how

they are related. For example, for outcomes such as anxiety and depression, there is some indication that either rather than or instead of social media use causing depression or anxiety, those with depression and anxiety may be more likely to use social media. A third option is that some other factor is causing both anxiety and social media use.

Speaking broadly, qualitative research paints a complex picture of teens' experience on social media. Many scholars now recognize that digital and social media define both the media environments and lived experience of teens. Separating out causal effects of specific media content within a complex digital media ecology is extraordinarily difficult, if not impossible. Yet, qualitative research has demonstrated the ambivalence of social media on adolescent well-being. While some discuss the psychological harms resulting from body comparison, cyberbullying, or fear of judgment, others discuss the benefits of increased social connection and community support.

There are important differences across individuals and groups in the benefits of and risks to well-being of social media use

Although research suggests there may not be a large or even significant correlational or causal relationship between social media use and well-being at the population level, there is indication that there are notable differences in outcomes across individuals or certain groups and that certain groups may be more at risk of negative psychological consequences using social media.

First, for adolescents with existing mental health issues, including anxiety, depression, or body dysmorphia, social media use is associated with greater negative outcomes. Similarly, those with mental health issues and those with poor family support are more likely to engage in extreme or problematic use of social media — in terms of time spent or intensity of use — which itself is linked to negative well-being.

Additionally, negative outcomes of social media use may be worse for younger teens, as compared to older teenagers. Specifically, younger teens are more likely to engage in problematic social comparison (see below) that may lead to negative mental health outcomes.

For marginalized groups, there appears to be both greater potential for harm and for benefits. Research has suggested that teens in marginalized groups often look to social media for community and social support. For racial minorities, use of social media was correlated to higher risk of bullying and exposure to discrimination and bias, which "can result in increased depressive symptoms and even posttraumatic stress disorder symptoms." And while researchers have also seen that LGBTQ+ teens may face more discrimination and bias online, "social media use was associated with fewer depressive symptoms but more risk of discrimination."

How kids use social media matters

In addition to differences across individuals or groups of adolescents, there is evidence that how kids use social media may matter to the potential psychological risks or benefits. There is strong evidence supporting that those adolescents who engage in problematic social media use are more likely to have negative mental health outcomes. Scholars use "problematic use" to characterize "individuals who experience addiction-like symptoms as a result of their social media use. Problematic social

media use reflects a non-substance related disorder by which detrimental effects occur as a result of preoccupation and compulsion to excessively engage in social media platforms despite negative consequences." One meta review observes a significant correlation between problematic use and depression, anxiety, and stress. Of course, much of this research is also cross-sectional, and therefore can not speak to the causal relationship, leaving it unclear if problematic social media use is causing mental health issues, or vice versa.

Additionally, there has been ongoing debate regarding whether there may be a difference in health outcomes between active and passive social media use. Active use is defined as when a user creates content, communicates with others, or actively builds networks. Passive use is defined as only viewing or reading content. While some early studies suggested that passive use of social media was more associated with depression and other negative health outcomes, more recent research has questioned that relationship. In a metaview focused on the topic, Valkenburg et al. conclude there is not sufficient evidence of a difference between active and passive social media use, which they argue may be a result of how each is operationalized and measured. Similarly, Hancock et al. saw no significant overall difference between active and passive use, but did see that active social media use was associated both with higher well-being and with depression.

Social Media, Generative AI, and Psychological Well-Being

Researchers have identified the specific mechanisms that connect social media use with psychological well-being. Hancock et al. and Orban et al. both offer useful syntheses of the mechanisms identified across the literature, grouping them into general categories. Hancock et al. describe three categories of mechanisms: social structure, psychological processes, and behavioral processes. Looking closely at how generative AI may moderate each mechanism offers one way of estimating or predicting the impact that the technology may have on adolescent psychological well-being.

As things stand, there remains a great deal we do not know about how generative AI on social media may impact these mechanisms and therefore psychological well-being. However, we identify those mechanisms where the existing literature on generative AI and social media may be instructive. Specifically, we consider mechanisms in which AI is likely to improve adolescent well-being, those likely to impair well-being, and those that are ambivalent or require specific additional research. In each section, we offer a set of specific recommendations for both platforms and policymakers meant to increase the probability that generative AI improves psychological well-being. These recommendations are highlighted below; please see the Executive Summary for the complete list of recommendations.

Size and Quality of Social Network and Support

Hancock et al. observe that the size and type of connections in one's social networks directly impact psychological well-being. The limited available evidence implies that, while there are significant risks, for certain users in certain cases, generative AI can positively impact perceptions of social support and with it, adolescent psychological well-being.

Hancock et al. note that the "perceptions of tangible and intangible assistance from one's social network" can decrease depression and loneliness. For social media in general, greater feelings of assistance from one's social network are associated with better psychological well-being.

But can human users receive social support from chatbots or AI models, and if so, how does it impact psychological well-being? Even while social science research on genAI is nascent, there are a surprising number of studies exploring the connection between AI companions and feelings of emotional support, specifically around loneliness. It is important to note, however, that much of this work remains grounded in short-term experimental studies, which may have limited external validity — that is, limited applicability to real world settings.

Broadly, there is reason to suspect that, as one mixed-methods study of ChatGPT finds, the impact of interactions with chatbots on emotional well-being is "highly nuanced," and, similar to the impact of social media on well-being, is highly dependent on both "personal factors" and the duration of usage. That study found that most users did not look to ChatGPT for emotional support — a finding that echoes results of research on earlier models that many users did not form strong emotional connections with models because they lacked qualities one looks for in friends. Yet, a "small sub-population" of users may seek out and form stronger emotional ties to models.

Yet, there remains much public concern that adolescents will look to chatbots for emotional support, even as those chatbots lack the emotional intelligence to offer real help. Some have observed an "empathy gap" in models, "whilst adept at simulating empathy, their lack of genuine emotional understanding can occasionally result in user interactions going awry." Worse, there are documented instances of models "encouraging suicide, eating disorders, self-harm, or violence," to emotionally vulnerable users looking for mental health support. For vulnerable users — especially those in crisis — flawed responses from chatbots, whether irrelevant or actively harmful, may pose significant risks.

Once users become emotionally dependent on a chatbot persona, they may also experience hardship if they lose access to that persona. At the same time, chatbots may compel vulnerable users to share highly sensitive data or information which can be exploited by model developers. Users — especially adolescents — may not understand company data policies and may not understand how companies may be using their data in the future. **While we generally support federal comprehensive consumer privacy legislation, in its absence, state lawmakers should continue to enact strict restrictions and protection surrounding minors' sensitive data.**

However, recent experimental studies have shown that both physical robots and AI chatbots can provide meaningful social and emotional support for some users, especially older users, and ultimately positively impact wellbeing. In particular, AI companions may provide safe, non-judgemental, and readily available emotional support for those who do not receive social support elsewhere or who face social ostracism. For example, Laestadius et al. observed that for one model, "users struggling with loneliness generally portrayed themselves as seeking out Replika because they were already lonely, rather than becoming lonely because of Replika." Indeed, there is a growing, if still nascent, literature on the use of chatbots for mental health treatment.

In these instances, the social support depends on perceptions of "realness," "sentience," or "mind perception." The benefits derived from engaging with chatbots are greater for users who "perceive chatbots as having human-like qualities." Yet, users may have mixed feelings about communicating with chatbots for emotional support, and may continue to prefer human connection and the support offered by humans — especially on certain sensitive topics.

Given the ambiguity in the existing research and that individual differences may mean users have very different perceptions of social support, **rather than a one-size-fits-all approach, platforms should personalize genAI product recommendations**. For example, they could give user the option to select the types of connections they are looking to build. Platforms could promote life-like bots for those who have strong on-network social ties, and focus on promoting human connection for those that do not. Alternatively, for some who lack strong support networks, AI companions could be an important source of social support. Some users may see real benefit in platforms promoting more life-like bots. For others, who already have strong networks, there is less need – and indeed it may be better to deprioritize life-like bots and promote existing human relationships.

Beyond the perception of support, the size of one's network, measured in terms of the number of engagements with content and number of online contacts, can increase self esteem. As described above, one of the key ways that generative AI is being integrated into platforms is through creating bots with which users can interact. While researchers have seen an increase in well-being related to an increase in likes and follows, that doesn't mean that bots should simply serve as social filler, boosting everyone's likes and follows. There is likely an inflationary effect, where bot engagement increasingly raises the bar on what is considered success. More importantly, there is indication that not all members of one's network are equal: for example, one study observed that users who had stronger connections in their networks indicated lower loneliness compared to those with more weak ties. This implies that the solution is not simply to have bots like every post and follow every user. **Rather, platforms should design chatbots that bridge or connect human users to help grow users' human networks. If bots are seen less as social media filler, and more as an opportunity to link real people together, there is greater chance of positively impacting well-being**

Self-Concept Development

Orban et al. define "self-concept" as "a person's beliefs and evaluations about their own quality and traits," noting that "a negative and unstable self-concept has been associated with negative mental health outcomes." Specifically, affordances like personalization of social media

can impact self-concept development. In their "algorithmic crystal framework," Lee et al. show that users assess the quality of algorithmic personalization "by its ability to aggregate and recommend content that reflected the many and multiple facets of participants' self-concept," and their ability to shape the algorithm through both organic interactions and "strategic attempts to influence the algorithmically-curated content."

This suggests that more sophisticated personalization methods rooted in generative AI may improve both user experience and adolescent well-being. Generative AI-facilitated personalization may better reflect users' "multifaceted and dynamic self concept." In contrast, more simplistic algorithmic representation "may be particularly harmful for certain communities," such as those from marginalized identities, when personalization is inaccurate or based on stereotypes.

These findings suggest that platforms and developers should ensure that generative AI tools on platforms are trained on data from diverse populations. It is essential that AI tools can indeed reflect users' "multifaceted and dynamic" self concept. Models that are trained on smaller, more homogeneous data sets may not be able to give users the broad palette that would permit them to creatively reflect complex identities.

Therefore, delaying wide scale integration of models/tools/bots until it is possible to deploy more advanced models may be better than starting with more simplistic prototypes.

How Generative AI May Decrease Adolescent Psychological Well-Being

Social Comparison

There is strong evidence that users — especially teens — frequently engage in social comparison on social media, which is when users compare themselves to others. Researchers have determined that the type of social comparison matters. Downward social comparison — where a user evaluates another as worse off — has been associated with an increase in well-being. In contrast, upward comparison — where a user evaluates another as better off than them — is often associated with a decrease in well-being and depression.

There is strong support that social media use can facilitate upward comparison for teens — especially as it concerns body image. Orban et al. note that one review suggests "that social media affordances create a 'perfect storm' for body image concerns that can contribute to both socio-emotional and eating disorders." Specifically, they argue that the "quantifiability" of social media, where one's success is quantified in metrics including "likes" or "follows," facilitates social comparison, and editability of posts contributes to a positivity bias, worsening upward comparison.

Generative AI offers at least two affordances relevant to social comparison. First, AI tools make it easier for users to create and edit visuals of idealized situations and idealized bodies. While photo editing on social media is by no means new, the promise of generative AI — especially integrated directly into social media platforms — is to make photo (and eventually, video) editing simple and fast. Widespread visual editing may worsen the positive bias already seen across social media, and may likely increase upward social comparison. In a recent

study, Tufail et al. saw evidence of this, where viewing AI-generated content was correlated with lower self-esteem and body image satisfaction, relationships that were mediated by social comparison.

While this suggests generative AI may worsen upward social comparison, and with it negative well-being, there are a series of design changes that platforms could make that may mitigate some of these negative impacts.

First, we recommend that platforms integrate generative AI editing tools directly into their platforms. There are important trade-offs to consider here. On one hand, integrating generative AI tools directly into platforms will likely not only improve user experience, it may improve a platform's capacity to identify and track content manipulation. Despite many ongoing efforts, we still lack tools that can reliably detect content that has been modified by generative AI. Integrated tools mean that platforms can have greater certainty when user-generated content has been edited. That being said, there are downsides to direct integration. A robust ecosystem of third-party tools may foster competition and innovation — including in safety measures, as different developers have to compete for user adoption. Ultimately, however, integrated tools can better permit platform interventions to attempt to ameliorate some of the most pernicious risks that generative AI presents to adolescent psychological well-being.

Second, when platforms know or detect that a user has made substantial edits to an image of a body that are likely to worsen upward social comparison, they should action those images. For example, platforms could disable viewership metrics to attempt to undercut the reward mechanism underlying social comparison. Alternatively, they could deprioritize or derank manipulated images of bodies, especially from influencers with large followings. Finally, they could apply labels to images of bodies that have been manipulated. Platforms could also create interstitial labels that users must click through to share a manipulated image and that requires users to confirm that they understand an image has been manipulated.

This recommendation poses at least two significant complications. First, deciding what counts as "substantial" will be difficult, for example: should applying a filter count? Does altering skin color or slimming a waist size count? Platforms will have to make difficult decisions on when and how to act on content at scale in ways that identify the most problematic forms of body manipulation.

Second, labels are unlikely to have large effects. However, the goal of labels isn't to be a silver bullet solution, but rather to help users better understand the content with which they engage.

Displacement

One of the best supported mechanisms connecting social media use and well-being is displacement — specifically of sleep, but also of a range of other activities, such as exercise, studying, in-person interaction, and other media activities. When users — especially teens — sleep less, they report worse outcomes across the board, including in mental and physical well-being. On one hand, this would be similar if an adolescent missed sleep to watch TV, play video games, or listen to the radio. What is perhaps more important is whether there are features or affordances on social media that drive continued use worsening displacement. There is evidence that teens may be especially susceptible to design elements on social media that drive continued and excessive use, including personalization or "dark patterns,"

which the Electronic Frontier Foundation (EFF) has defined as "design tactics used in websites and apps to manipulate you into doing things you probably would not do otherwise."

As noted above, the platforms themselves seem to be counting on generative AI tools increasing user engagement. So far, they have released little data confirming this, however, available evidence suggests displacement remains a significant contributor to the negative psychological impacts of social media use.

Over the last several years, a number of scholars have suggested design mitigations to help mitigate the displacement effect of social media. Many of these mitigations likely apply to generative AI on social media as well. For example, a recent report from the Knight Georgetown Institute Expert Working Group on Recommender Systems, argues that "Platforms must provide easily accessible ways for users to set their preferences about types of items to be recommended and to be blocked."

Most importantly, we recommend platforms continue to iterate better self and parent controls, such as limiting total amount of platform use, or restricting use after certain times. Tools could also incorporate a wider range of design choices, such as altering recommendations in the evening to surface less engaging, less exciting content.

Most of the major social media platforms offer tools with some of these functions; however, fewer offer tools that make it easy for individuals to restrict *their own* usage. As our understanding of the benefits and risks of generative AI increase, it is essential that platforms continue to update these tools to reflect our changing understanding.

Risky Posting Behaviors

Orban et al. identify risky posting behavior as one additional behavioral mechanism connecting social media use and adolescent well-being. This can mean posting potentially embarrassing or revealing content, connecting with dangerous users, or filming engaging in risky physical activities. On one hand, adolescents often engage in more risky behavior in general than other age groups, in part because they underestimate the risks of that behavior. But Orban et al identify a series of affordances of social media that help explain this: quantifiability (of approval) and visibility can drive riskier behavior in pursuit of social approval online.

While we lack data to understand how AI-mediated personalized content may increase or decrease risky behavior in general, we do have extensive analysis of one source of risk AI poses online: non-consensual (generated) intimate imagery (NCII). New AI tools make it extraordinarily easy to generate fake intimate or sexual images. There has been a great deal of attention about this problem both within schools and regarding prominent celebrities. The growing threat of NCII means that posting *any* images online exposes adolescents to the risk of NCII being created and circulated.

Luckily, this is one of the few areas where regulators and platforms have moved quickly. Congress just passed the Take It Down Act with bipartisan support. If signed into law, the bill would criminalize the production and distribution of both generated NCII and CSAM.

Yet, legal restrictions are likely insufficient to address the harm this content poses — especially to adolescents. Given this, we recommend two further approaches. **First, state policymakers should require and fund**

programming in schools on the harm of both consuming and producing AI-generated NCII and CSAM. More and more schools and school districts are integrating digital media literacy programs, and it is essential these programs include discussion of the harm of producing generated NCII and CSAM.

Second, policymakers, platforms, and civil society should all focus more effort helping victims of NCII and CSAM. While new laws may punish perpetrators, they will do little to ameliorate the existing harm done to victims. For example, policymakers could earmark additional resources for school psychological counseling, training programs for school psychologists on helping victims of NCII, or other public resources for victims. Platforms could fund and host informational campaigns and connect users to psychological services. This is an opportunity for policymakers, platforms, and civil society to work together to supply better resources for victims of this content.

Where Generative AI May Have Ambivalent Impacts

In this final section, we look at mechanisms where social media may have ambivalent impact, largely to identify specific areas where we need additional research.

Observing thousands of studies, Orban et al. identify fear-of-missing-out (FOMO) as a key mechanism between social media use and psychological well-being. However, the research points to two opposite relationships between FOMO and well-being. Roberts & David found that the relationship between FOMO and well-being is mediated by both (feelings of) social connection and social media intensity. As far as FOMO drives social connection, it can ultimately *increase* well-being by motivating users to be more connected. But, insofar as it worsens the intensity of use of social media, it may ultimately decrease well-being.

Social media intensity describes the ways that people use social media, and one common scale assesses the amount of time spent on social media and the number of friends one has, but also a more subjective sense of whether a platform "is part of my everyday routine," and how one would feel if it shut down.

The general consensus is that higher social media intensity is correlated with worse psychological well-being. Similarly, "problematic use" of social media, which is often assessed by measuring "addiction" is also correlated with worse psychological well-being. Importantly, there remains much debate about the direction and existing of causality in this relationship.

As far as how this relates to generative AI on social platforms, we see three useful data points that indicate specific future research questions. First, despite a growing concern among activists and policymakers that algorithmic targeting and personalization can increase addiction or problematic use, there is mixed evidence for this effect. More directly, we currently have little data on how generative AI may impact problematic use of social media.

Second, as described above, platforms expect that better integration of generative AI on social platforms will improve user experience and increase usage. Yet, there remains much debate about the relationship between time spent on

social media and well-being. While a number of cross-sectional studies have seen correlation between time spent and worse well-being, some longitudinal studies do not show a causal relationship. That is, instead of more time spent on a platform causing worse psychological well-being, it could be that those with worse well-being are spending more time on the platform. Even if generative AI motivates greater time spent on the platform, it is unclear the impact that may have on well-being.

Third, there is some indication that whether social media usage is "active"–creating posts, actively engaging with contacts, or "passive" — simply viewing content, it is correlated with well-being. While the operating consensus has been that passive use is associated with worse well-being, and active with better, one recent meta analysis, throws that simplistic account into question. Valkenburg et al. observe that the majority of studies do not support that account, but instead point to a far more complex picture. Again, however, even if passive use is correlated with worse well-being, it remains unclear the direction of the relationship: it is both plausible that poor well-being drives social media use, and that social media use drives poor well-being.

In addition to continuing to analyze the effects of active vs passive social media use, we need a better understanding on how generative AI impacts the type of use. On one hand, AI tools may make it easier for users to create content, increasing active use. On the other hand, influencers and for-profit companies are already adopting these tools more rapidly than many regular users, flooding platforms with generated content. Depending on how platforms treat such content, this may ultimately lead to more passive social media use.

Ultimately, we need significant further study on how generative AI mediates the underlying mechanisms that link social media and adolescent psychological well-being. **Specifically, we recognize a need for research addressing: How does generative AI impact the amount and quality of time spent by users on social media platforms? That is, does it increase time spent or problematic use? Does it change when, why, or how users engage on social media platforms?**

Conclusion

Nearly two and a half years after ChatGPT dropped, generative AI still excites near unbridled hype. Though the pace of advancement is occasionally shocking, there remains a significant gap between the extraordinary claims made about generative AI and its current capacity. We have not yet entered a world where most white-collar jobs are outsourced to models, where we are awash in deceptive visuals, or where human relationships have lost their value and meaning.

Similarly, as the techlash matures, there remains a great deal of concern over the harm that social media poses to adolescent well-being. Yet, years of empirical analysis indicates that there is a gap between the fears and extraordinary claims made about the harm of social media on kids, and our understanding of the impact social media actually has on adolescents.

Generative AI and child online safety are not only two of the most prominent issues in technology policy, but two issues where public debate remains permeated with hype, baseless prognostication, and fear. As policymakers consider the risks and benefits of generative AI and the need for regulation, some are already matching unrestrained hype for generative AI with groundless fears about the dangers to adolescents. In contrast, we need careful, empirically based analysis that leverages the existing research base to understand the specific ways that generative AI on social media may help or hurt kids. It is only when we have a real understanding of the specific ways that generative AI may amplify or nullify certain risks, that we can design better systems to maximize the benefits and minimize the risks that generative AI poses to children.