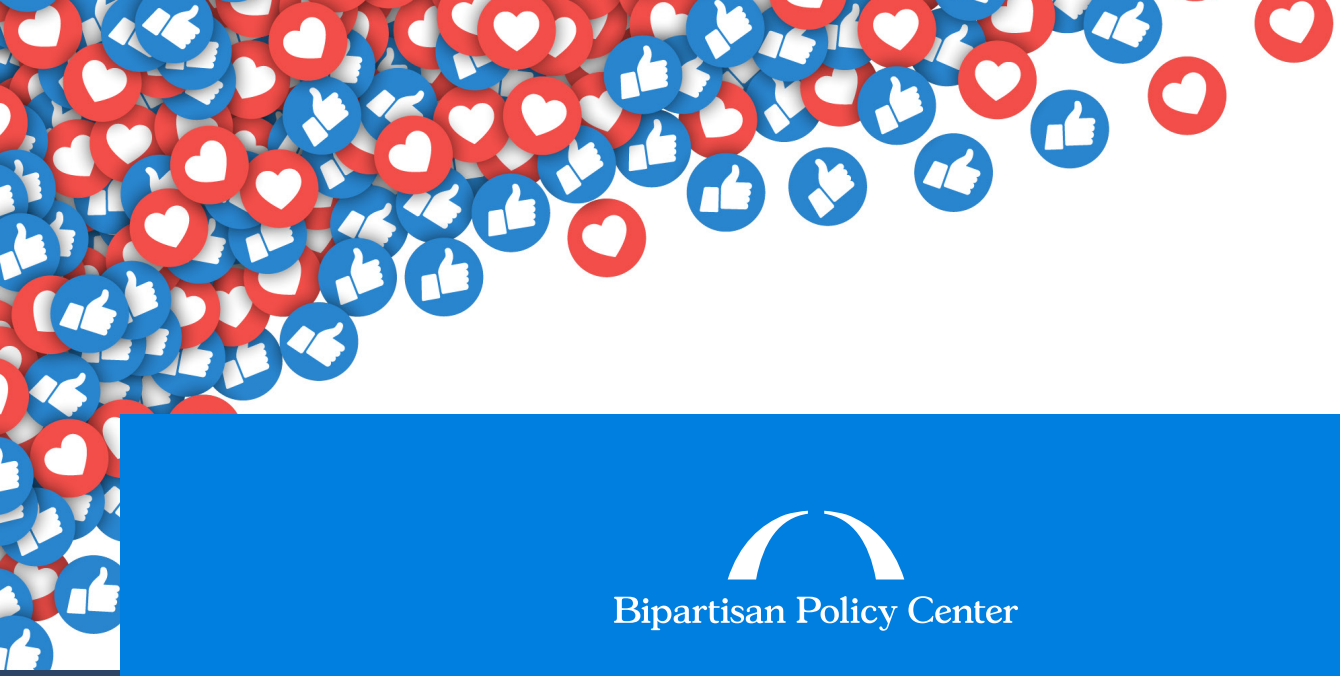

Bipartisan Policy Center

The Pros and Cons of Social Media Algorithms

AUTHORS

Danielle Draper
Project Manager,
Technology Project,
Bipartisan Policy Center

Sabine Neschke
Policy Analyst, Technology Project,
Bipartisan Policy Center

Point-of-view: you are scrolling your social media timeline. Have you ever seen an ad for the exact pair of shoes you were considering buying online earlier that week? Or did you recently adopt a dog and now your social media feed is flooded with posts targeted toward new dog owners? It's no coincidence. Today's AI-powered social media platforms use advanced algorithms to curate content just for you.

Social media algorithms are built into the platform design process for various use cases, such as sorting data and ranking content at a speed and scale that would be impossible for humans. Notably, algorithmic content moderation forms the backbone of social media security by restricting illegal, harmful, or inappropriate content. Algorithms ensure the safety and effectiveness of digital platforms.

While social media algorithms have powerful technological capabilities, maintaining a safe digital environment for billions of people among an abundance of content poses many challenges. As Congress looks to address the offline ramifications of social media, there are bipartisan concerns algorithms may influence these real-world harms.[1] If policymakers intend to regulate how digital platforms use algorithms, they must understand the implications and tradeoffs of this rapidly advancing technology. This explainer provides a brief high-level overview of the numerous benefits of social media algorithms while contrasting their limitations.

Impacts to Children's Online Safety

Benefits of algorithms:	Limitations of algorithms:
Algorithmic recommendation systems can help youth find valuable resources encouraging self-discovery and connections with like-minded peers, especially for marginalized youth, such as LGBTQ+ youth and youth who identify as racial minorities. This creates influential social communities that are important to a child's upbringing. Content moderation algorithms ensure children's online safety by detecting and removing various forms of dangerous content, including content linked with self-harm. This can help foster a safe online environment for kids by mitigating the spread of content that could incite substance abuse, sexual harassment, or other harmful conduct.	Social media companies design algorithms to keep users engaged for extended periods of time by feeding users customized content. The addictive nature of algorithms can impact a person's quality of sleep, which is linked to mental health concerns in youth (e.g., anxiety and depression).[2] Algorithmic recommendations entice users to remain on the platform longer, and studies found that increased time spent online is correlated with increased exposure to age-inappropriate content, unrealistic standards of beauty, or cyberbullying.[3] For example, a child with low self-esteem will likely be algorithmically recommended pictures that may exacerbate social comparison or negative feelings. Overexposure to harmful content poses serious mental health risks to children and adolescents, particularly for vulnerable groups. Lastly, algorithmic recommendations use personal data which, despite the intent, can track and exploit children's online behavior by serving them specific ads. Many teens report feeling they have little control over the personal information social media companies collect about them.[4]

Impacts to Online Speech

Benefits of moderation algorithms:	Limitations of moderation algorithms:
Like most businesses, many social media platforms seek to safeguard their companies from liability and reputational damage. Digital platforms that deploy proactive community guidelines, employ trust and safety teams, and use algorithmic content moderation have successfully reduced the risk of harmful content online.[5] Prohibiting social media companies from using algorithms to moderate speech could have far-reaching implications, such as an influx of dangerous and graphic content. With the rise of false information, hate speech, and violent threats online, it is vital that platforms continue to invest in the research and development of advanced content moderation technologies.	While social media companies use algorithms to police their sites for harmful content, there are growing concerns this allows for too much editorial power. Should artificial intelligence (AI) have the ultimate authority or expertise in determining what is and is not true? Algorithmic downranking and shadow banning suppress certain forms of content while promoting others. Screening algorithms that inaccurately classify legitimate content as harmful can trigger the removal of false positives (i.e., over-blocking content). Unfair censorship can negatively impact content creators and lead to information gaps. To facilitate communication and free expression—social media's primary functions— many worry the increased use of algorithms could unintentionally silence marginalized communities.

Improving Political Literacy	
Benefits of recommendation algorithms:	**Limitations of recommendation algorithms:**
Many Americans get their political news primarily through social media.[6] Given social media's role in civic engagement and social activism, algorithmic recommendations help connect users with like-minded people who share aligned values. This can help political participants form supportive networks, build collective action, and increase awareness about societal issues.[7] Algorithmic recommendations can also facilitate the free exchange of information and ideas—a hallmark of a functioning democracy—across the digital environment. Social media connects users with factual, authoritative sources of information. For example, some platforms attach accurate voting information to posts about elections and use algorithms to amplify posts from election offices.[8]	Unfortunately, human psychology tells us that bad news is more likely to get our attention.[9] Designing algorithms based on human behavior means content that is more headline-grabbing and polarizing tends to go viral and rank higher in users' feeds—often called algorithmic amplification. This phenomenon can inadvertently steer more users towards hyper-partisan news, both on the left and the right.[10] Furthermore, algorithms attempt to personalize each user's experience by primarily showing political news based on one's political beliefs. This feedback loop may limit users' exposure to different viewpoints and push users into filter bubbles or echo chambers. This ideological segregation could lead to confirmation bias and polarize public opinion in ways that are not deliberate.[11] Lastly, the rise of AI-generated content can lead to the proliferation of election-related disinformation (e.g., such as deepfakes impersonating election officials) which social media algorithms will then recycle. Flooding social media with AI-generated photos, videos, and audio makes it difficult to differentiate fact from fiction, which can erode the public's confidence in democracy.

Transparency and Explainability around Algorithmic Decision-Making	
Benefits:	**Limitations:**
Clear information-sharing, audits, impact assessments, disclosure requirements, or transparency reports around automated content moderation systems could increase research, accountability, and people's trust in platforms. The use of AI algorithms is often considered a challenge to transparency suggesting that knowledge about this complex technology is not yet broadly accessible. The inability to understand how algorithms reach their conclusions is referred to as the "black box problem." For example, users frequently express confusion over why they receive certain social media posts or why their content was deleted.[12] Algorithms determine what billions of people watch, work, buy, and think on the internet, but information about these internal processes is not always easy to find or understand or is cost-prohibitive.[13]	To promote transparency, researchers need access to thorough data from the AI systems used by social media companies. Algorithms are confidential and well-kept mathematical formulas. The question of how to perform effective algorithm audits without compromising the integrity of platforms' trade secrets and intellectual property rights should be considered. Fully disclosing explanations about algorithms could encourage malicious actors to abuse AI systems to their advantage, such as conducting foreign disinformation campaigns. Maintaining American leadership in AI innovation is especially important given the global competitiveness of these technologies.

Mitigating AI System Bias	
Benefits of AI innovation:	**Limitations of AI innovation:**
The degree to which algorithms operate accurately and effectively relies on careful design and continuous evaluation by the computer scientists who built them. Social media software engineers continually test algorithms to provide the best user experience and improve platform design. AI algorithms are trained and retrained to account for accuracy, discrimination, or other ethical concerns. Algorithms are designed to provide equal visibility to content shared by people of different ethnicities, genders, sexual orientations, and backgrounds. For example, in online advertising, social media algorithms can be trained to recognize and filter out content that may reinforce harmful stereotypes or offensive materials.[14]	Algorithms are designed and trained by humans, and all humans are marred by unconscious bias. This means social media algorithms may have built-in biases that can exacerbate societal challenges or disproportionally affect marginalized groups. If algorithms are trained on unrepresented, incomplete, or skewed data, it can lead to automation bias against certain groups regarding their ethnicity, political affiliation, sexual preference, gender, or race. For example, algorithms might recommend or amplify divisive content that reinforces racial stereotypes, ultimately perpetuating historical inequities. Conversely, algorithms may disproportionately screen and suppress content that challenges cultural norms, which may reinforce prejudiced viewpoints, limiting opportunities for minority groups.[15]

Collecting User Data on Social Media	
Benefits of algorithms:	**Limitations of algorithms:**
Algorithms enhance user experiences online by harnessing data to provide tailored recommendations. By analyzing users' behaviors, algorithms create a personalized experience online by delivering curated content that resonates with their interests.[16] Algorithms can also combat cyber-attacks or data breaches by monitoring for unwanted or unrecognized behaviors. They can quickly detect and safeguard against a security breach, securing individuals' private information, including account details.	Algorithms rely heavily on data to run effectively, which raises numerous privacy implications. Some users are concerned with the nature of data being collected and inferred about them. For example, social media algorithms may recommend content based on a user's demographics, geographic location, or search history. People's comfort in disclosing information and concern about their privacy online ranges according to many factors and is often referred to as the Privacy Paradox.[17] Another concern is around how algorithms store and transfer data, especially as social media companies might share it with third parties without the individual's consent or knowledge. The monetization of personal data through targeted advertising may exacerbate these problems, prompting some to worry about users' ability to opt-out of such practices and the sale of their information.

Language Complexities in Content Moderation	
Benefits of NLP algorithms:	**Limitations of NLP algorithms:**
Natural Language Processing (NLP) algorithms and Large Language Models enable computer programs to comprehend human language as it is recorded or typed. These algorithms allow for real-time sentiment analysis and speech translation, which play important roles in harmful speech detection. By deploying NLP, digital platforms can analyze millions of online conversations, searching for hateful, harassment, swearing, and other inappropriate language in posts. As the volume of online content increases, NLP algorithms continue to train on new language trends, such as slang, slurs, or ambiguous euphemisms, thereby enhancing its content moderation capabilities.	There are more than 7,000 languages spoken across the world today. The inherent complexity of language can negatively impact the capabilities of NLP. For example, new research demonstrates that algorithmic detection is flawed for many languages other than English.[18] Most AI models are trained predominantly on content in English, a bias that leaves a significant amount of online content vulnerable to inconsistent moderation.[19] This shortcoming creates the risk of under-or over-moderating harmful speech globally. For example, misinformation in other languages is more likely to spread undetected on social media than in English.[20]

Social Media Algorithms and Section 230 Immunity	
Benefits of Section 230 immunity:	**Limitations of Section 230 immunity:**
Section 230 of the Communications Decency Act is arguably one of the most important laws in tech policy because it shields social media companies from liability for user-generated content.[21] Recently, there has been controversy over whether Section 230 liability protections should protect decisions made by algorithms. Historically, courts have broadly interpreted the use of algorithms as a traditional editorial function that passively organizes content (within Section 230 scope) versus actively creating content (outside Section 230 scope). If digital platforms were held accountable for the actions of their algorithms, it could lead to an influx of lawsuits. This liability could disproportionately hurt startups or smaller platforms that do not have sophisticated content moderation systems or cannot afford liability charges.	Currently, algorithms do not have legally enforceable safety and efficacy standards. Some argue that it's time to rethink and revise Section 230 protections for platform design actions. For example, if an algorithm flags dangerous content but it is not removed, should the digital platform be penalized for knowingly hosting or profiting from material that violates its Terms of Service? Another growing concern is whether Section 230 protects the emerging use of generative AI on social media given that the machine learning algorithm, as the concept suggests, "generates" new content. As generative AI expands its capabilities, these deployments come with potentially significant legal risks and a rapidly changing policy debate around Section 230. Currently, the courts and Congress have yet to answer the Section 230 vs. generative AI question.[22]

Role of Human Moderators in Content Moderation

Benefits of increased use of algorithms:	Limitations of increased use of algorithms:
In the early days of social media, content moderation was performed by small groups of employees who made split-second decisions on removing content. The sheer volume of content on today's digital platforms requires the use of AI. When platforms adopt automated content analysis systems, the algorithms behind the tools can eliminate large volumes of inappropriate content—before it reaches a human moderator for further review. This shields employees from constant exposure to disturbing violence, egregious conspiracy theories, and graphic imagery which can lead to considerable mental health risks. The advanced capabilities of algorithms decrease the reliance on human moderators and increase the speed and scale of effective content moderation.	As content moderation becomes more automated, it's important to note that AI systems are not perfect and can make incorrect conclusions. Algorithms can have trouble parsing the intent or context of social media posts, which is why a human-in-the-loop approach is still needed to ensure the accuracy of content removal. When algorithms flag potentially harmful content, human moderators should have the ability to review or override the decision of the machine. It is critical human oversight plays an active role in ensuring AI systems are used responsibly. Unfortunately, recent layoffs in the tech sector reduced the size of many trust and safety teams, which may indicate companies are divesting resources dedicated to detecting harmful content.[23]

Detecting Extremist Networks

Benefits of moderation algorithms:	Limitations of recommendation algorithms:
When digital platforms are equipped with algorithmic detection systems, they can more proactively suspect illegal activity and intervene before offline attacks occur. For example, machine learning algorithms can predict people's trajectories toward violent extremism by studying their online behavior.[24] In 2017, Facebook, Twitter, Google, Microsoft, and other major tech companies established the Global Internet Forum to Counter Terrorism, which invests in the innovation and distribution of leading technological tools used to identify terrorist propaganda.	Concerned critics question the role algorithms play in the radicalization and recruitment of extremist networks. Algorithms may help connect susceptible bad actors, specifically young adults, who may search, consume, and spread harmful content with like-minded people online. In 2022, the Supreme Court heard *Gonzalez v. Google*— a case about an international terrorist attack in which plaintiffs alleged YouTube's algorithmic recommendations aided and abetted ISIS. The case was not the first-time social media platforms came under scrutiny for their correlation with offline violence. For example, many domestic terrorists are particularly active on social media leading up to mass shootings.[25] While many mainstream platforms develop advanced AI tools to mitigate offline violence, defining violent extremism is a contested concept that makes effective content moderation even more challenging. Many threat actors are simply moving their strategies to smaller platforms that don't moderate content.

Detecting Child Sexual Abuse Material	
Benefits of moderation algorithms:	**Limitations of moderation algorithms:**
U.S. federal law requires tech companies to report known child sexual abuse material (CSAM) to the National Center for Missing & Exploited Children's (NCMEC), which sends those reports to law enforcement. Every day, mainstream social media platforms flag thousands of images containing child abuse using algorithmic detection methods. From the smallest startups to the largest tech companies in the world, the Tech Coalition is an alliance of global corporations that work together to proactively advance AI technology to combat online child sexual exploitation. For example, Tech Coalition members are given access to advanced image classification systems and hash-matching technology. This technology leverages algorithms to identify and match social media posts against the NCMEC's database of known child abuse imagery.	Whereas a social media algorithm that recommends a low-quality shirt may result in a bad purchase, algorithms that promote pedophile networks have much more serious allegations and consequences. Shocking investigations have revealed that recommendation algorithms help connect a vast network of social media accounts that propagates CSAM.[26] If CSAM proliferates on public feeds via major tech platforms, one can only predict the extent of the problem on smaller platforms or direct messaging apps. Perpetrators continuously evolve their methods to avoid online detection, which leaves platforms attempting to play catch-up and update their algorithms. For example, researchers at the Stanford Internet Observatory are concerned about the rise of AI-generated CSAM, in which machine learning algorithms scrape images of real children as source material.[27]

1 Megan McCluskey. "How Addictive Social Media Algorithms Could Finally Face a Reckoning in 2022." TIME. January 4, 2022. Available at: https://time.com/6127981/addictive-algorithms-2022-facebook-instagram/.

2 Rea Alonzo, Junayd Hussain, Saverio Stranges, Kelly K. Anderson. "Interplay between social media use, sleep quality, and mental health in youth: A systematic review." *Sleep Medicine Reviews*. 56. April 2021. Available at: https://www.sciencedirect.com/science/article/abs/pii/S108707922030157X?via%3Dihub.

3 Kira Riehm, Kenneth Feder, Kayla Tormohlen, Rosa Crum, Andrea Young, Kerry Green, Lauren Pacek, Lareina La Flair, Ramin Mojtabai. "Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth." *JAMA Psychiatry*. 76(12): 1266-1273. December 2019. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6739732/#yoi190054r28.

4 Emily Vogels and Risa Gelles-Watnick. "Teens and social media: Key findings from Pew Research Center surveys." Pew Research Center. April 24, 2023. Available at: https://www.pewresearch.org/short-reads/2023/04/24/teens-and-social-media-key-findings-from-pew-research-center-surveys/.

5 Niall McCarthy. "Facebook Removes Record Number Of Hate Speech Posts [Infographic]." Forbes. May 13, 2020. Available at: https://www.forbes.com/sites/niallmccarthy/2020/05/13/facebook-removes-record-number-of-hate-speech-posts-infographic/?sh=78fe4a130356.

6 "Social Media and News Fact Sheet." Pew Research Center. September 20, 2022. Available at: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

7 Stefania Milan. "When Algorithms Shape Collective Action: Social Media and the Dynamics of Cloud Protesting." *Social Media + Society*. 1(2). December 2015. Available at: https://journals.sagepub.com/doi/full/10.1177/2056305115622481.

8 Naomi Gleit. "Launching The Largest Voting Information Effort in US History." Meta. June 16, 2020. Available at: https://about.fb.com/news/2020/06/voting-information-center/.

9 Kent Campbell. "Why do people click on bad news? Negativity bias." Reputation X. July 29, 2023. Available at: https://blog.reputationx.com/what-makes-us-drawn-to-negative-online-content.

10 Paul Barrett, Justin Hendrix, Grant Sim. "How tech platforms fuel U.S. political polarization and what government can do about it." Brookings. September 27, 2021. Available at: https://www.brookings.edu/articles/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/.

11 Fernando P. Santos, Yphtach Lelkes, Simon A. Levin. "Link recommendation algorithms and dynamics of polarization in online social networks." *Proc Natl Acad Sci.* 118 (50). December 2021. Available at: https://www.pnas.org/doi/full/10.1073/pnas.2102141118.

12 Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, Jillian York. "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation." *International Journal of Communication*. 13(18). 2019. Available at: https://ijoc.org/index.php/ijoc/article/view/9736.

13 Mehil Mohan. "The Rising Costs of Social Media APIs: The Twitter-Reddit Scenario." Codedamn. June 11, 2023. Available at: https://codedamn.com/news/api/cost-of-apis-twitter-reddit.

14 Lokke Moerel. "Algorithms can reduce discrimination, but only with proper data." Iapp. November 16, 2018. Available at: https://iapp.org/news/a/algorithms-can-reduce-discrimination-but-only-with-proper-data/.

15 Theodora (Theo) Lau, Uday Akkaraju. "When Algorithms Decide Whose Voices Will Be Heard." Harvard Business Review. November 12, 2019. Available at: https://hbr.org/2019/11/when-algorithms-decide-whose-voice-will-be-heard.

16 David Doty. "A Reality Check On Advertising Relevancy And Personalization." Forbes. August 13, 2019. Available at: https://www.forbes.com/sites/daviddoty/2019/08/13/a-reality-check-on-advertising-relevancy-and-personalization/?sh=68c5eb297690.

17 Patricia A. Norberg, Daniel R. Horne, David A. Horne. "The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors." *Journal of Consumer Affairs*. 41(1), 100–126. March 2007. Available at: https://doi.org/10.1111/j.1745-6606.2006.00070.

18 Gabriel Nicholas, Aliya Bhatia. "The Dire Defect of 'Multilingual' AI Content Moderation." Wired. May 23, 2023. Available at: https://www.wired.com/story/content-moderation-language-artificial-intelligence/.

19 "Social Media Algorithms: Content Recommendation, Moderation, and Congressional Considerations." Congressional Research Service. July 27, 2023. Available at: https://crsreports.congress.gov/product/pdf/IF/IF12462.

20 Stephanie Valencia. "Misinformation online is bad in English. But it's far worse in Spanish." The Washington Post. October 28, 2021. Available at: https://www.washingtonpost.com/outlook/2021/10/28/misinformation-spanish-facebook-social-media/.

21 47 U.S. Code § 230

22 Tony Phillips, Jaria Martin. "Will Generative AI Create a Break in the Impenetrable Wall That Is Section 230?" Pillsbury Law. June 16, 2023. Available at: https://www.pillsburylaw.com/en/news-and-insights/generative-ai-section-230.

23 Hayden Field, Jonathan Vanian. "Tech layoffs ravage the teams that fight online misinformation and hate speech." CNBC. May 6, 2023. Available at: https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html.

24 "Research Rooted in Machine Learning Challenges Conventional Thinking About the Pathways to Violent Extremism." National Institute of Justice. July 24, 2023. Available at: https://nij.ojp.gov/topics/articles/research-rooted-machine-learning-challenges-conventional-thinking-about-pathways.

25 Jillian Peterson, James Densley, Jamie Spaulding, Stasia Higgins. "How Mass Public Shooters Use Social Media: Exploring Themes and Future Directions." February 26, 2023. Available at: https://journals.sagepub.com/doi/10.1177/20563051231155101.

26 Jeff Horwitz, Katherine Blunt. "Instagram Connects Vast Pedophile Network. The Wall Street Journal. June 7, 2023. Available at: https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189?mod=hp_lead_pos7.

27 David Thiel, Melissa Stroebel, Rebecca Portnoff. "Generative ML and CSAM: Implications and Mitigations." Stanford Digital Repository. June 23, 2023. Available at https://purl.stanford.edu/jv206yg3793.