Position: Generative AI Regulation Can Learn from Social Media Regulation

Ruth E. Appel [1]

Abstract

There is strong agreement that generative AI should be regulated, but strong disagreement on how to approach regulation. In this position paper, we argue that the debates on generative AI regulation can be informed by evidence on social media regulation. For example, AI companies have faced allegations of political bias similar to the allegations social media companies have faced. We discuss four specific policy recommendations based on the evolution of social media and their regulation: (1) counter bias and perceptions thereof (e.g., via transparency, oversight boards, researcher access, democratic input), (2) address specific regulatory concerns (e.g., youth wellbeing, election integrity) and invest in trust and safety, (3) promote computational social science research, and (4) take on a more global perspective. Applying lessons learnt from social media regulation to generative AI regulation can save effort and time, and prevent avoidable mistakes.

1. Introduction

When Google's generative AI model Gemini produced images of racially diverse Nazis in early 2024, it led to a public outcry and allegations of anti-conservative bias (Robertson, 2024). Almost a decade earlier, the first allegations of anti-conservative bias were made against social media platforms like Facebook (Barrett & Sims, 2021), and they have persisted (Romm, 2019; Barrett & Sims, 2021). This shows that the content moderation challenges that emerging technologies face are not entirely new. Media scholars have called attention to the fact that new technologies often elicit similar questions and concerns as their predecessors (Wartella & Reeves, 1985). Generative AI is the latest, but not the first technology to garner widespread attention and raise societal and regulatory concerns.

In this paper, we argue that **generative AI regulation can learn from social media regulation**, which has evolved over the past two decades. While there is strong agreement that generative AI should be regulated—evidenced by the large number of recent regulatory efforts across countries and stakeholders (Zaidan & Ibrahim, 2024)—, there is strong disagreement on how to approach regulation. Some argue that AI regulation should mostly rely on extensions of existing laws (Huttenlocher et al., 2023), while others argue that entirely new laws and regulations are needed and have proposed laws and regulations such as the EU AI Act (European Parliament and Council, 2024) or California's vetoed AI Safety Bill SB 1047 (Wiener et al., 2024). Analyzing the evolution of social media regulation can provide insights into which approaches to regulation are promising when it comes to generative AI, which in turn can save effort and time, and prevent avoidable mistakes.

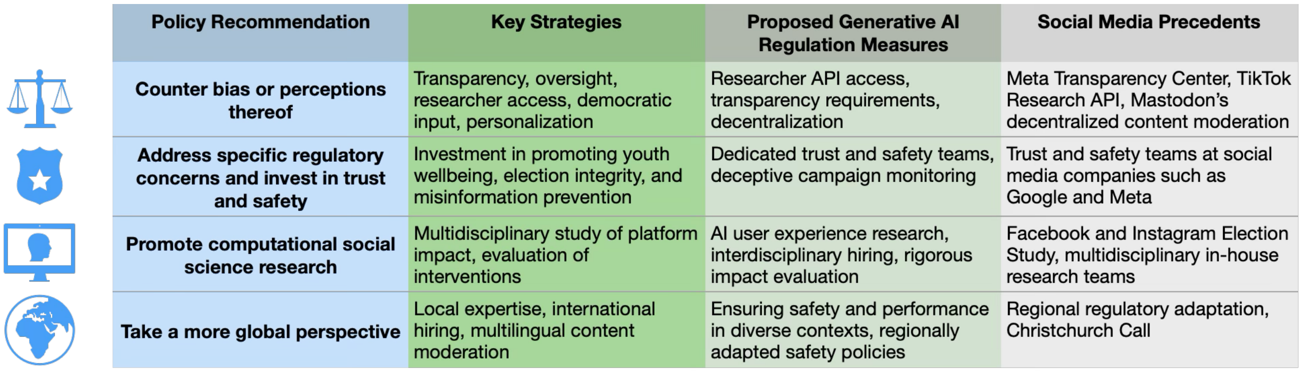
The focus of this paper is on content moderation, i.e. how to regulate the content generated by generative AI models or shown on social media platforms. Further, the paper focuses on regulation in a broad sense, which can include self-regulation of industry players and formal laws such as the EU AI Act (European Parliament and Council, 2024). We discuss specific policy recommendations based on the evolution of social media and their regulation.

2. Learnings from social media regulation for generative AI regulation

Generative AI and social media share important features (see Appendix A for an analysis), including the use of AI and content moderation. Although generative AI and social media differ on some dimensions, these differences are mostly differences in degree, not in kind, when it comes to regulation. Thus, lessons learnt from social media regulation are relevant to generative AI regulation. This paper provides four policy recommendations for generative AI regulation based on social media regulation: (1) counter bias and perceptions thereof (e.g., via transparency, oversight boards, researcher access, democratic input), (2) address specific regulatory concerns (e.g., youth wellbeing, election integrity) and invest in trust and safety, (3) promote computational social science research, and (4) take on a more global perspective (see Figure 1 for an overview).

[1] Stanford University, Stanford, CA, United States. Correspondence to: Ruth E. Appel <rappel@cs.stanford.edu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).



Policy Recommendation	Key Strategies	Proposed Generative AI Regulation Measures	Social Media Precedents
Counter bias or perceptions thereof	Transparency, oversight, researcher access, democratic input, personalization	Researcher API access, transparency requirements, decentralization	Meta Transparency Center, TikTok Research API, Mastodon's decentralized content moderation
Address specific regulatory concerns and invest in trust and safety	Investment in promoting youth wellbeing, election integrity, and misinformation prevention	Dedicated trust and safety teams, deceptive campaign monitoring	Trust and safety teams at social media companies such as Google and Meta
Promote computational social science research	Multidisciplinary study of platform impact, evaluation of interventions	AI user experience research, interdisciplinary hiring, rigorous impact evaluation	Facebook and Instagram Election Study, multidisciplinary in-house research teams
Take a more global perspective	Local expertise, international hiring, multilingual content moderation	Ensuring safety and performance in diverse contexts, regionally adapted safety policies	Regional regulatory adaptation, Christchurch Call

Figure 1. **Policy recommendation overview.** Overview of the lessons generative AI regulation can learn from social media regulation.

2.1. Counter bias and perceptions thereof

Given that both generative AI and social media share the key features—use of content moderation, use of AI, black-box nature, abstraction of the complexity of algorithmic decision-making such that much of the decision-making is intransparent—, it is no surprise that both generative AI companies and social media companies have faced allegations of bias, including allegations of anti-conservative political bias (Robertson, 2024; Barrett & Sims, 2021). While there is no evidence of anti-conservative bias for social media (Barrett & Sims, 2021), multiple studies have shown political bias in generative AI (Rozado, 2023; Röttger et al., 2024).

Generative AI models have also been shown to exhibit other forms of bias, such as anti-Muslim bias (Abid et al., 2021) and stereotypical depictions of race, gender, age, nationality, and socioeconomic status (Nangia et al., 2020).

Addressing such biases is as important as it is challenging. It is important to address biases because biases can harm and manipulate users. For example, political bias in generative AI models can influence users' opinions (Bai et al., 2023; Matz et al., 2024; Potter et al., 2024) and behavior (Fisher et al., 2024). Biases may also lead to lower-quality output, entrench historical biases and stereotypes, and undermine trust. It is challenging to address biases because they are challenging to measure accurately, e.g. due to sensitivity to the prompt design (Röttger et al., 2024) and order effects (Dominguez-Olmedo et al., 2024) and can arise at different points in the development and deployment of generative AI (Suresh & Guttag, 2021; Ferrara, 2023).

Social media companies have taken different approaches to address biases or perceptions thereof that mainly focus on transparency about algorithms and decision-making, gathering input from users and learning from case studies, and increasing user choice.

2.1.1. Increase transparency and researcher access

The shared features content moderation, use of AI, black-box and abstraction give rise to transparency challenges for social media and generative AI. Generative AI transparency is lacking (Bommasani et al., 2023; 2024). Social media companies have pursued multiple different approaches to increase transparency and generative AI can learn from this playbook. For example, Facebook's parent company Meta introduced features such as "Why am I seeing this ad?" that allowed users to understand why they were served certain ad content (Thulasi, 2019) and established an independent oversight board of experts for contentious content moderation decisions (Meta, 2024). These initiatives do not come without problems. In response to the launch of Facebook's oversight board, "The Real Facebook Oversight Board" was created, which brought experts together to argue for more independence, transparency and regulation (The Real Facebook Oversight Board, 2022). Company policies are also not guaranteed to be permanent. In January 2025, Meta starkly shifted its content moderation policy, limiting its efforts to reduce misinformation and harmful speech (Isaac & Schleifer, 2025; Iyer, 2025).

An important aspect of transparency is allowing for third-party evaluations. Efforts to create research platforms or APIs accessible to researchers, such as the Meta Researcher Platform (Li et al., 2022) and the TikTok Research API (TikTok, 2025), or to design academic-industry collaboration such as the Facebook and Instagram Election Study (Clegg & Nayak, 2020) are helpful, but imperfect (Wagner, 2023). The Coalition for Independent Technology Research was founded after researchers at different institutions faced difficulty maintaining or gaining access to social media data for research purposes (Coalition for Independent Technology Research, 2022). Importantly, we can learn from these shortcomings. Researcher access programs to evaluate technology should be characterized by sufficient resources (including staffing, infrastructure, and funding), incentives that are compatible with academic research (e.g., data retention policies, persistent API access and publication permission for researchers), sound knowledge sharing processes between internal and external researchers to help understand data availability and analysis feasibility, helpful documentation, privacy preserving measures and timeliness in terms of data access, publication review and addressing issues that researchers discovered. To protect researchers involved, researchers have called for legal protection for researchers pursuing legitimate research purposes, initially for social media (Abdo et al., 2022) and more recently for generative AI (Longpre et al., 2024).

Regulations like the Digital Services Act prescribe transparency by requiring audits of social media companies (European Commission, 2023), and similar auditing efforts are imaginable for generative AI. In fact, some scholars suggest to extend and adapt DSA rules for social media platforms to generative AI (Hacker et al., 2023).

While the specific implementation of these transparency efforts may be contentious and requires nuance, there is a broader lesson: Generative AI regulation can incentivize measures for increasing transparency, such as short and accessible explanations of the technology, independent oversight mechanisms, researcher access and mandatory audits.

2.1.2. Gather democratic input to inform technology

Generative AI and social media share features that make them complex, including that the content they feature can pertain to a variety of domains, that there is potential for personalization, and that content could be moderated in various different ways. One approach to determine what a good content moderation system may look like is to gather input directly from users to inform design choices. Different initiatives have been launched over the past few years to gather input from users and enable democratic decisions about the nature of regulation and content moderation for social media and AI (Wetherall-Grujić, 2023). Social media also offers case studies of networks where content moderation seems to be broadly accepted and deliver productive results, such as in the case of the deliberation platform vTaiwan (Miller, 2019). Finally, social media researchers have studied how to embed important societal values into social media AI (Bernstein et al., 2023), which could inform how such values can be embedded into generative AI.

2.1.3. Promote user choice

Another option to empower users to make choices in the face of features such as content moderation and the varied nature of content is to enable users to set up rules for a subset of the system. The social media platform Mastodon is a prominent example in terms of increasing user choice in such a way. Mastodon is built on the idea that different communities can create their own servers and set and enforce their own content moderation rules (Mastodon, 2024). This highlights that the feature of personalization may be a potential route for resolving content moderation dilemmas. Content moderation questions with regards to generative AI and social media are similar and it is not clear what opinion representation should be the default, but increased personalization of models may be an answer (Redpoint, 2020).

2.2. Address specific regulatory concerns and invest in trust and safety

The feature of content moderation that generative AI and social media share comes with challenges such as preventing the spread of harmful misinformation and protecting user wellbeing. Social media companies have invested in teams that address these specific regulatory concerns. Examples include teams at companies like Google and Microsoft working on youth wellbeing and mental health, election integrity, preventing spam, preventing the spread of child sexual abuse material, preventing harmful misinformation, detecting deceptive campaigns, and ensuring trust in the platform and safety of its users.

Generative AI chatbot performance has already been rated with regards to certain principles that apply just as much to social media (see e.g. Common Sense Media, 2024).

Yet, generative AI companies do not have teams at the same scale as social media companies to address these issues. Generative AI companies are much smaller and younger than some of the social media giants, thus it is not surprising that they do not have as much dedicated staff to work on these issues. Going forward, however, adding diverse staff beyond engineers that can bring in expertise to address issues such as user mental health or combating misinformation is important to address the variety of risks and harms that generative AI models pose (for taxonomies of risks and harms related to generative AI, see Weidinger et al., 2021; Marchal et al., 2024). Investment in trust and safety teams seems particularly crucial, and it is encouraging to see that companies like OpenAI and Anthropic are investing in this area, with OpenAI publishing the first-ever report on the activity of deceptive campaigns on generative AI platforms in May 2024 (Nimmo, 2024).

The policies social media companies have put in place to decide how and when to moderate individual users, and the best practices they have developed to uncover abuse such as deceptive campaigns that try to interfere with elections or spam users, could inform the approaches generative AI companies take. This includes developing a repertoire of content moderation approaches, which could include bans, warnings and strikes for misbehavior, more guardrails and throttling usage for users that try to abuse generative AI models. Social media companies also gained experience in involving the user community in content moderation decisions (e.g., in the case of BirdWatch (Wojcik et al., 2022)) and how to collaborate across platforms, and generative AI companies could consider how these approaches could be adapted to their platforms.

2.3. Promote computational social science research

Both generative AI and social media allow users to express themselves and allow for a connection, be it to other users or to an AI with a vast pool of knowledge. How these media interact with users is a key part of what makes them so influential. They are neither purely technical, nor purely social systems. This suggests that multidisciplinary study—computational social science—is needed to understand, evaluate and shape these systems (Gillespie et al., 2024).

Social media companies have hired researchers from many disciplines, including computer science, psychology, political science, communication, law and others, to better understand how their platforms impact society, and how certain interventions influence society and their revenue.

Rigorous computational social science evaluations, whether conducted in-house or via external researchers with platform access, are key to ensuring that technologies such as generative AI and social media meet their goal of being helpful and not harmful to society. Further investment in research is needed because generative AI has features that differ from previous technologies, so its impact and user preferences are not clear. Even the impact of previous technologies such as social media has not yet been comprehensively evaluated and needs further investment. Rigorous research can inform platform and public policy when it comes to regulation, and it can enhance user trust.

This implies the need to invest in diverse research teams that understand the interaction of humans and technology and can evaluate the societal implications of that technology. While AI company recruiting often focuses heavily on engineers, and some companies are more concerned with extreme risks in the more distant future, social media companies have shown the value of multidisciplinary teams to address current risks such as biases. Multidisciplinary teams allow companies to test different product features and interventions effectively, e.g. to reduce misinformation spread. Guidance on building effective red teams for generative AI models also highlights the importance of diverse teams (Ofcom, 2024; Metcalf & Singh, 2024; Ahmad et al., 2024; Oremus, 2023). Computational social scientists from any background, data scientists and user experience researchers would be especially helpful to address questions at the intersection of humans and technology, such as which emotional bonds may be formed between humans and AI.

While content moderation on social media is far from a resolved issue, there is a large and growing body of academic literature addressing user preferences and content moderation approaches (e.g., Persily & Tucker, 2020; Appel et al., 2023; Kozyreva et al., 2024), which could inform content moderation for generative AI.

2.4. Take on a more global perspective

Both generative AI and social media can be used in a variety of contexts. Generative AI companies have grown rapidly and are serving users around the world, similar to social media companies. However, compared to social media companies, many generative AI companies are more heavily focused on the US. To address problems like biases, it is crucial that companies take on a global perspective and embrace local expertise in multiple countries. The reasoning mirrors that for the benefits of diversity in AI red teaming (Ofcom, 2024; Metcalf & Singh, 2024; Oremus, 2023), i.e. that broader representation allows for a better understanding of user preferences and the harms that a technology may pose. The stakes are high. If companies fail to invest in taking user preferences and risk factors outside of the US seriously, the technology may serve large numbers of users worse, contain undiscovered harms (Metcalf & Singh, 2024; Oremus, 2023), and could even promote violence in conflict regions (Amnesty International, 2022). Given the increasing amount of national and local regulations on generative AI, global expertise is also important to keep up with local laws.

3. Conclusion

There are strong disagreements about the approach that should be taken to regulate generative AI. This paper argued that the regulation of generative AI can be informed by the evolution of the regulation of social media. While social media is by far not the only analogy proposed for generative AI (Maas, 2023), generative AI and social media share key features that make a comparison of the two worthwhile. This paper outlined recommendations regarding transparency, researcher access, gathering democratic input, promoting user choice, addressing specific regulatory concerns, increasing investments into computational social science research, and taking on a more global perspective. Analyzing social media regulation may inform and accelerate the process of developing generative AI regulation. Regulation takes time and effort, so where possible, resources should be saved and mistakes avoided by looking at the lessons that social media regulation and research hold for generative AI regulation.

Acknowledgements

The author is grateful to Jennifer Pan for feedback on an earlier draft.

References

Abdo, A., Krishnan, R., Krent, S., and Woods, A. K. A safe harbor for platform research, 2022. URL https://knightcolumbia.org/content/a-safe-harbor-for-platform-research.

Abid, A., Farooqi, M., and Zou, J. Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6):461–463, 2021. ISSN 25225839. doi: 10.1038/s42256-021-00359-2. URL https://doi.org/10.1038/s42256-021-00359-2.

Ahmad, L., Agarwal, S., Lampe, M., and Mishkin, P. OpenAI's approach to external red teaming for AI models and systems. Technical report, OpenAI, 1 2024. URL https://cdn.openai.com/papers/openais-approach-to-external-red-teaming.pdf.

Amnesty International. The social atrocity: Meta and the right to remedy for the Rohingya. Technical report, Amnesty International, 2022. URL https://www.amnesty.org/en/documents/asa16/5933/2022/en/.

Appel, R. E., Pan, J., and Roberts, M. E. Partisan conflict over content moderation is more than disagreement about facts. *Science Advances*, 9(44):1–10, 2023. URL https://doi.org/10.1126/sciadv.adg6799.

Bai, H., Voelkel, J. G., Eichstaedt, J. C., and Willer, R. Artificial intelligence can persuade humans on political issues, Feb 2023. URL osf.io/stakv_v1.

Barrett, P. M. and Sims, J. G. False accusation: The unfounded claim that social media companies censor conservatives. Technical Report February, NYU Stern Center for Business and Human Rights, 2021. URL https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/6011e68dec2c7013d3caf3cb/1611785871154/NYU+False+Accusation+report_FINAL.pdf.

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

Bernstein, M. S., Christin, A., Hancock, J. T., Hashimoto, T., Jia, C., Lam, M., Persily, N., Piccardi, T., Saveski, M., Tsai, J. L., Ugander, J., and Xu, C. Embedding societal values into social media algorithms. *Journal of Online Trust and Safety*, 2(1):1–13, 2023. URL https://doi.org/10.54501/jots.v2i1.148.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The Foundation Model Transparency Index, 2023. URL http://arxiv.org/abs/2310.12941.

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., and Liang, P. The Foundation Model Transparency Index v1.1, 2024. URL https://doi.org/10.48550/arXiv.2407.12929.

Bostrom, N. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013. doi: https://doi.org/10.1111/1758-5899.12002. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12002.

Cattell, S., Ghosh, A., and Kaffee, L.-A. Coordinated disclosure for AI: Beyond security vulnerabilities, 2024. URL http://arxiv.org/abs/2402.07039.

Center for Humane Technology. The attention economy: Why do tech companies fight for our attention? Technical report, Center for Humane Technology, 2021. URL https://cdn.prod.website-files.com/5f0e1294f002b15080e1f2ff/612f8e3fa20df8374659a774_1-TheAttentionEconomyIssueGuide.pdf.

Clark, H. H. *Using Language*. Cambridge University Press, 1996. URL https://doi.org/10.1017/CBO9780511620539.

Clegg, N. and Nayak, C. New Facebook and Instagram research initiative to look at US 2020 presidential election, August 2020. URL https://about.fb.com/news/2020/08/research-impact-of-facebook-and-instagram-on-us-election/.

Coalition for Independent Technology Research. Coalition for Independent Technology Research Founding Document, 2022. URL https://independenttechresearch.org/coalition-for-independent-technology-research-founding-document/.

Common Sense Media. AI Initiative, 2024. URL https://www.commonsensemedia.org/ai.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the survey responses of large language models, 2024. URL https://arxiv.org/abs/2306.07951.

European Commission. Shaping Europe's digital future Commission adopts rules on independent audits under the Digital Services Act, 2023. URL https://digital-strategy.ec.europa.eu/en/news/commission-adopts-rules-independent-audits-under-digital-services-act.

European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689.

Ferrara, E. Should ChatGPT be biased? challenges and risks of bias in large language models, 2023. URL http://arxiv.org/abs/2304.03738.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., and Reinecke, K. Biased AI can influence political decision-making, 2024. URL https://arxiv.org/abs/2410.06415.

Gillespie, T., Shaw, R., Gray, M. L., and Suh, J. AI red-teaming is a sociotechnical system. now what?, 2024. URL https://arxiv.org/abs/2412.09751.

Hacker, P., Engel, A., and Mauer, M. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1112–1123, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594067. URL https://doi.org/10.1145/3593013.3594067.

Huttenlocher, D., Ozdaglar, A., and Goldston, D. A framework for U.S. AI governance: Creating a safe and thriving AI sector. Technical report, MIT Schwarzman College of Computing, 2023. URL https://computing.mit.edu/wp-content/uploads/2023/11/AIPolicyBrief.pdf.

Isaac, M. and Schleifer, T. Meta says it will end its fact-checking program on social media posts. *The New York Times*, January 2025. URL https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking. Updated January 15, 2025.

Iyer, R. To evaluate meta's shift, focus on the product changes, not the moderation. *Tech Policy Press*, January 2025. URL https://www.techpolicy.press/to-evaluate-metas-shift-focus-on-the-product-changes-not-the-moderation/.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., Panizza, F., Pennycook, G., Rand, D. G., Rathje, S., Reifler, J., Schmid, P., Smith, M., Swire-Thompson, B., Szewach, P., van der Linden, S., and Wineburg, S. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, pp. 1–9, 2024. ISSN 23973374. URL https://doi.org/10.1038/s41562-024-01881-0.

Kumar, V. Making "freemium" work. *Harvard Business Review*, May 2014. URL https://hbr.org/2014/05/making-freemium-work.

Li, D., Pyke, R., Jiang, R., and Jagadeesh, K. Introducing the Researcher Platform: Empowering independent research analyzing large-scale data from Meta, January 2022. URL https://research.facebook.com/blog/2022/1/introducing-the-researcher-platform-empowering-independent-research-analyzing-large-scale-data-from-meta/.

Longpre, S., Kapoor, S., Klyman, K., Ramaswami, A., Bommasani, R., Blili-Hamelin, B., Huang, Y., Skowron, A., Yong, Z. X., Kotha, S., Zeng, Y., Shi, W., Yang, X., Southen, R., Robey, A., Chao, P., Yang, D., Jia, R., Kang, D., Pentland, A., Narayanan, A., Liang, P., and Henderson, P. Position: A safe harbor for AI evaluation and red teaming. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32691–32710. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/longpre24a.html.

Lukoff, K., Lyngs, U., Zade, H., Liao, J. V., Choi, J., Fan, K., Munson, S. A., and Hiniker, A. How the design of YouTube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445467. URL https://doi.org/10.1145/3411764.3445467.

Maas, M. M. AI is like… a literature review of AI metaphors and why they matter for policy. *SSRN Electronic Journal*, October 2023. doi: 10.2139/ssrn.4612468. URL https://doi.org/10.2139/ssrn.4612468.

Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., and Isaac, W. Generative AI misuse: A taxonomy of tactics and insights from real-world data, 2024. URL https://arxiv.org/abs/2406.13843.

Mastodon. Mastodon, 2024. URL https://joinmastodon.org/.

Matz, S. C., Teeny, J., Vaid, S., Harari, G., and Cerf, M. The potential of generative AI for personalized persuasion at scale. *Nature Scientific Reports*, 14(4692): 1–16, 2024. URL https://doi.org/10.1038/s41598-024-53755-0.

McGrenere, J. and Ho, W. Affordances: Clarifying and evolving a concept. *Graphics Interface*, pp. 1–8, 2000. URL https://graphicsinterface.org/wp-content/uploads/gi2000-24.pdf.

Meta. Oversight Board, 2024. URL https://transparency.fb.com/en-gb/oversight/oversight-board-recommendations/.

Metcalf, J. and Singh, R. Scaling up mischief: Red-teaming AI and distributing governance. *Harvard Data Science Review*, May 2024. URL https://hdsr.mitpress.mit.edu/pub/ded4vcwl.

Miller, C. Crossing Divides: How a social network could save democracy from deadlock, 2019. URL https://www.bbc.com/news/technology-50127713.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models, 2020. URL https://aclanthology.org/2020.emnlp-main.154/.

Nimmo, B. AI and covert influence operations: Latest trends. Technical Report May, OpenAI, 2024. URL https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf.

Ofcom. Red teaming for GenAI harms: Revealing the risks and rewards for online safety. Discussion paper, Office of Communications, 7 2024. URL https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/red-teaming/red-teaming-for-gen-ai-harms.pdf.

Oremus, W. Meet the hackers who are trying to make AI go rogue, 2023. URL https://www.washingtonpost.com/technology/2023/08/08/ai-red-team-defcon/.

Persily, N. and Tucker, J. A. (eds.). *Social Media and Democracy*. Cambridge University Press, 2020. ISBN 9789162893439. URL https://www.cambridge.org/core/books/social-media-and-democracy/E79E2BBF03C18C3A56A5CC393698F117.

Potter, Y., Lai, S., Kim, J., Evans, J., and Song, D. Hidden persuaders: LLMs' political leaning and their influence on voters. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4244–4275, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.244/.

Rafaeli, S. and Sudweeks, F. Networked interactivity. *Journal of Computer-Mediated Communication*, 2(4): JCMC243, March 1997. URL https://doi.org/10.1111/j.1083-6101.1997.tb00201.x.

Redpoint. Stanford professor Tatsu Hashimoto on AI biases and improving LLM performance, 2020. URL https://www.youtube.com/watch?v=pceYeZdT1D0.

Robertson, A. Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis. *The Verge*, 2024. URL https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

Romm, T. Senate Republicans renew their claims that Facebook, Google and Twitter censor conservatives, apr 2019. URL https://www.washingtonpost.com/technology/2019/04/10/facebook-google-twitter-under-fire-senate-republicans-censoring-conservatives-online/.

Ronzhyn, A., Cardenal, A. S., and Rubio, A. B. Defining affordances in social media research: A literature review. *New Media & Society*, 25(11):3165–3188, 2023. doi: 10.1177/14614448221135187. URL https://doi.org/10.1177/14614448221135187.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., and Hovy, D. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models, 2024. URL http://arxiv.org/abs/2402.16786.

Rozado, D. The political biases of ChatGPT. *Social Sciences*, 12(3), 2023. URL https://doi.org/10.3390/socsci12030148.

Sharma, S. and Murano, P. A usability evaluation of web user interface scrolling types. *First Monday*, 25(3), February 2020. URL https://doi.org/10.5210/fm.v25i3.10309.

Stern, J. AI is like ... nuclear weapons? *The Atlantic*, March 2023. URL https://www.theatlantic.com/technology/archive/2023/03/ai-gpt4-technology-analogy/673509/.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pp. 1–9. ACM, October 2021. doi: 10.1145/3465416.3483305. URL https://doi.org/10.1145/3465416.3483305.

The Real Facebook Oversight Board. The Real Facebook Oversight Board, 2022. URL https://the-citizens.com/real-facebook-oversight/.

Thulasi, S. Understand why you're seeing certain ads and how you can adjust your ad experience, 2019. URL https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/.

TikTok. Research tools - TikTok for developers, 2025. URL https://developers.tiktok.com/products/research-api/.

Wagner, M. W. Independence by permission. *Science*, 381 (6656):388–391, 2023. URL https://doi.org/10.1126/science.adi2430.

Wartella, E. and Reeves, B. Historical trends in research on children and the media: 1900-1960. *Journal of Communication*, 35(2):118–133, 1985. URL https://doi.org/10.1111/j.1460-2466.1985.tb02238.x.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models, 2021. URL https://arxiv.org/abs/2112.04359.

Wetherall-Grujić, G. The race to democratise AI, 2023. URL https://democracy-technologies.org/participation/the-race-to-democratise-ai/.

Wiener, Roth, Rubio, and Stern. Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, 2024. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., and Baxter, J. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, 2022. URL https://arxiv.org/abs/2210.15723.

Zaidan, E. and Ibrahim, I. A. AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, 11(1):1–18, 2024. URL https://doi.org/10.1057/s41599-024-03560-x.

A. Affordances of generative AI and social media

To shed light on the similarities and differences between specific media, we can analyze their affordances. For the purposes of this paper, we define affordances as the features that characterize a medium in its relationship to its users (for a detailed discussion of different definitions and the evolution of the term affordances, see McGrenere & Ho, 2000; Ronzhyn et al., 2023). Both generative AI, e.g. in the form of a chatbot like OpenAI's ChatGPT or Anthropic's Claude, and social media, e.g. in the form of Meta's Facebook or X (formerly Twitter), can be considered media that allow to create and distribute content and are shaped by specific features. The features discussed here pertain to a medium in general, but may not apply to every instance, that is, a specific generative AI application or social media platform may differ from the norm in terms of its affordances.

Based on an analysis of commonly used generative AI applications (e.g., ChatGPT and Claude) as well as social media applications (e.g., Facebook and X), we identified key features that generative AI and social media share or that differentiate them. The analysis of features is grounded in work by Clark (1996), who discusses several features of media that fall into three categories: medium, control, and immediacy. Since Clark (1996)'s features focus on the affordances of face-to-face communication,[1] we added new features and removed features that are less relevant to the comparison of generative AI and social media. We also added the feature of interactivity discussed by Rafaeli & Sudweeks (1997). We will point out each feature that is adapted from Clark (1996) or Rafaeli & Sudweeks (1997).

We will first address why social media is comparable to generative AI in key aspects that have implications for technology regulation. Then, we will engage with differences in the affordances of social media and generative AI to show that the analogy is useful, but imperfect.

Table 1. **Comparison of affordances of generative AI and social media**

Feature	Definition	Generative AI	Social Media
Medium			
Spatial separation	Content is generated in different locations	**Yes**	**Yes**
Direct connection	Medium is conversation partner	Yes	No
User connections	Medium connects user to other users	No	Yes
Interactivity	Medium responds interactively to user input	**Yes**	**Yes**
Dialogue-by-default	Actions occur in a dialogue	Yes	No
Recording	User actions are recorded	**Yes**	**Yes**
Personalization	User context and preferences are learnt over time	**Yes**	**Yes**
Single output	Medium presents usually just a single output	Yes	No
Infinite content	Content is served infinitely	No	Yes
General content	Content can pertain to any domain	**Yes**	**Yes**
General purpose	Medium serves many functions	Yes	No
Use of AI	Medium learns patterns from data	**Yes**	**Yes**
Abstraction	Medium hides its complexity	**Yes**	**Yes**
Black-box	How algorithmic decisions are made is intransparent	**Yes**	**Yes**
Control			
Content moderation	Content is moderated at all	**Yes**	**Yes**
Invisible content moderation	Most content moderation is not visible to the user	Yes	No
Content moderation pre-generation	Content is moderated before it is received by the user	Yes	No
Self-determination	User can decide themselves how to act	**Yes**	**Yes**
Self-expression	User can express themselves	**Yes**	**Yes**
Simultaneity	User can receive and produce content concurrently	No	Yes
Immediacy			
Instantaneity	Actions are perceived almost immediately	**Yes**	**Yes**
Evanescence	Medium quickly recedes to the background	**Yes**	**Yes**

Note: The features spatial separation, recording, self-determination, self-expression, simultaneity, instantaneity, and evanescence, as well as the categories medium, control and immediacy are based on Clark (1996). The feature interactivity is based on Rafaeli & Sudweeks (1997). Instances where features of generative AI are similar to features of social media are highlighted in bold.

[1] There are contextual differences between face-to-face communication on the one hand and generative AI and social media on the other, such as where and why they may be used. This paper focuses on the comparison of generative AI and social media, and therefore focuses on features in Clark (1996)'s model that are pertinent to generative AI and social media, but not the comparison to other media.

A.1. Generative AI and social media are comparable in key aspects

The analogy between generative AI and social media is valuable because both media share key features. Importantly, the shared affordances of generative AI and social media imply that both of these media necessarily moderate content and thus face complex content moderation challenges and public scrutiny.

Table 1 shows key similarities between generative AI and social media when it comes to the features of each ***medium***. Both generative AI and social media allow for *spatial separation*, that is, the conversation partners usually generate content in different physical spaces—e.g., in a home office and at a data center for generative AI—and are not copresent (copresence is one of the features described in Clark (1996)). Both media feature **interactivity** and respond interactively to user input, which makes them engaging (Rafaeli & Sudweeks, 1997) (interactivity is defined and discussed in Rafaeli & Sudweeks (1997)). Both generative AI and social media are *recording* user data (the recording feature is adapted from Clark (1996)'s recordlessness feature). Both media can learn about a user's context and their preferences over time for output *personalization*, e.g. by updating the chatbot's memory or personalizing a recommendation algorithm. Further, both generative AI and social media can feature *general content*, i.e. content on all kinds of domains (e.g., hobbies, jobs, politics). Both are powered by *artificial intelligence* (AI), that is, they rely on learning patterns from data to perform well on tasks such as generating text or recommending content, although generative AI relies on more recent deep learning models while social media tends to rely on traditional machine learning approaches such as recommender systems. Both media also feature *abstraction*, that is, they hide the complex technical implementation details from the user behind a simple user interface. Further, generative AI and social media algorithms tend to be *black-box*, that is, algorithmic decisions are intransparent—almost always for users, but often also for experts because mechanistic interpretability (Bereska & Gavves, 2024) that can explain why a deep learning model made a certain decision is in its infancy.

With regards to ***control*** features (Clark, 1996), both generative AI and social media feature *content moderation*, that is, the medium shapes what content is allowed to appear. Both media also meet Clark (1996)'s criteria for *self-determination*, i.e. a user's ability to decide themselves how to act, and *self-expression*, i.e. a user's ability to express themselves on a medium.

With regards to ***immediacy*** (Clark, 1996), both generative AI and social media share *instantaneity* (Clark, 1996), i.e. that actions are perceived almost immediately, and *evanescence* (Clark, 1996), i.e. that the medium recedes to the background quickly once it is not actively used anymore.

Beyond features, the evolution of generative AI is similar to the evolution of social media in that both are characterized by limited, lagging regulation and large inflows of funding for technology entrepreneurship in this space (Stern, 2023).

A.2. Generative AI and social media are not perfectly comparable

While the shared affordances highlight the value of comparing generative AI to social media, we acknowledge that the analogy is imperfect. By definition, an analogy is not a perfect match. As Jacob Stern put it: "[T]his is just the nature of analogies: They are illuminating but incomplete" (Stern, 2023).

Table 1 reveals differences in affordances between generative AI and social media. With regards to features of the ***medium***, generative AI and social media show some variation. While generative AI such as ChatGPT constitutes a conversation partner that is in *direct connection* with the user, social media foster *user connections*—connections between users. Whereas generative AI interacts in a *dialogue-by-default* manner with the user, social media is merely mediating between the user and their human conversation partners (e.g., when a social media algorithm displays one user's post on another user's feed) and tend to involve a sequence of one-off actions. While generative AI tends to respond to prompts, usually with a *single output* instead of multiple outputs, and does not continue to serve content unless the user requests it, social media often feature *infinite content* via mechanisms such as infinite scroll (Sharma & Murano, 2020) or auto-play (Lukoff et al., 2021), which serve content as long as the user is on the platform. The purpose of social media tends to be focused on social communication, while generative AI is considered a *general purpose* technology that could serve various functions, including as a text writer or reviewer, a calculator, a programmer and much more.

With regards to ***control*** features, a feature Clark (1996) proposed is *simultaneity*, which is the user's ability to receive and produce content concurrently. Simultaneity is given for social media—e.g., one user might send a message at the same time as another user is sending them a message—, but not for generative AI, which operates in a sequential dialogue of user input and model output. Important differences between generative AI and social media are related to content moderation: Even though both generative AI and social media feature content moderation, content moderation in generative AI tends to use *invisible content moderation* more than social media. Social media platforms may occasionally take hardly visible actions such as downranking posts, but many social media content moderation actions such as removal of a post or user are clearly visible. Generative AI models, on the other hand, are built and fine-tuned to moderate content in a certain way (e.g., to avoid providing dangerous information), without the user necessarily becoming aware of the moderation. Generative AI content moderation may be invisible to the user because the model will usually respond, and not necessarily provide a reason if it refuses to respond to a prompt directly, which makes moderation less obvious than a missing response or a refused response citing the reason for refusal. Relatedly, generative AI models tend to moderate *before* the content is shown to the user, e.g. by refusing to reply to a prompt, while social media content moderation tends to occur only *after* content made it onto a platform, e.g. when a post was reported as harmful misinformation.

Beyond specific features of generative AI and social media, there are differences in their context and potential consequences. In terms of business model, most social media companies rely on revenue from advertisements (Center for Humane Technology, 2021), while prominent generative AI companies have so far leaned towards freemium (Kumar, 2014) subscription models. While the potential harm of social media to democracy and society has been an important focus of scholarly and public attention (Persily & Tucker, 2020), some argue that the destructive potential of AI may be at another level since it may present a larger threat (Bostrom, 2013) or stronger geopolitical advantage (Stern, 2023). Generative AI and social media differ also in the level of uncertainty they bring. For example, auditing and discovering vulnerabilities in systems that are probabilistic (Cattell et al., 2024), like generative AI models, implies new complexities that traditional, deterministic social media algorithms do not entail. Finally, generative AI and social media may differ in areas that have so far remained legally uncertain, such as questions of liability (e.g., for harms results from media use) and copyright. This means the learnings for generative AI regulation should be based on, and not go beyond key shared features.

B. Alternative views

In this paper, we argued that generative AI regulation can learn from social media regulation. However, there are valid counterarguments related to the imperfect analogy between generative AI and social media and the fact that social media regulation has not been a model example of technology regulation.

First, as discussed in detail in Appendix A.2, there are important differences in the affordances of social media and generative AI, including whether the medium acts as conversation partner and how visible content moderation is. While this concern is valid, the recommendations in this paper build on the affordances that generative AI and social media share and do not go beyond those.

Second, as described in earlier parts such as Section 2.1.1, social media regulation has not been a model example of technology regulation. Initiatives like coalitions to protect independent researchers (Coalition for Independent Technology Research, 2022) show that the research community and the broader public has not been satisfied with how social media regulation played out. However, we can learn lessons from both past failures and past successes. The encouragement to learn from social media regulation does not mean that we should always take similar regulatory approaches for generative AI. It means that we should carefully assess what worked well, and what needs to be improved.

C. Impact statement

There are increasingly heated debates about what appropriate and effective generative AI regulation should look like. This paper shows that we do not have to reinvent the wheel when it comes to questions such as how to ensure that generative AI is safe and moderated in alignment with users' preferences. Instead, we can learn lessons from social media regulation.

Concrete lessons we can learn include the importance of investing in trust and safety and taking a more diverse perspective, both in terms of geography and research disciplines involved.

Learning lessons from social media regulation can help prevent avoidable mistakes and utilize resources more effectively, which can ultimately improve AI policy and AI safety.