Generative AI Regulation Can Learn From Social Media Regulation

Ruth E. Appel
Department of Communication
Stanford University
Stanford, CA 94305
rappel@cs.stanford.edu

Abstract

There is strong agreement that generative AI should be regulated, but strong disagreement on how to approach regulation. While some argue that AI regulation should mostly rely on extensions of existing laws, others argue that entirely new laws and regulations are needed to ensure that generative AI benefits society. In this paper, I argue that the debates on generative AI regulation can be informed by the debates and evidence on social media regulation. For example, AI companies have faced allegations of political bias regarding the images and text their models produce, similar to the allegations social media companies have faced regarding content ranking on their platforms. First, I compare and contrast the affordances of generative AI and social media to highlight their similarities and differences. Then, I discuss specific policy recommendations based on the evolution of social media and their regulation. These recommendations include investments in: efforts to counter bias and perceptions thereof (e.g., via transparency, researcher access, oversight boards, democratic input, research studies), specific areas of regulatory concern (e.g., youth wellbeing, election integrity) and trust and safety, computational social science research, and 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 and Sims, 2021), and have continued to persist e.g. during Senate hearings (Romm, 2019) and when former President Trump was banned from Twitter (now X) and Facebook (Barrett and 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 and Reeves, 1985). Generative AI is the latest technology to garner widespread attention and raise societal and regulatory concerns, but so have social media and other technologies before it.

The aim of this paper is to show that what has been learnt with regards to social media regulation in the past two decades can inform the regulation of generative AI going forward. While there is strong agreement that generative AI should be regulated, there is strong disagreement on how to approach regulation. Some argue that AI regulation should mostly rely on extensions of existing laws (Huttenlocher, Ozdaglar and Goldston, 2023), while others argue that entirely new laws and regulations are needed and have proposed laws and regulations such as the EU AI Act, the White

2nd Workshop on Regulatable ML at NeurIPS 2024.

House Executive Order on AI, or California's AI Safety Bill SB 1047. 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 design and regulate the content that is generated by a generative AI model or shown on a social media platform. Further, the paper focuses on regulation in a broad sense, which can include self-regulation of industry players to ensure harmless output or avoiding bias, and formal laws such as the General Data Protection Regulation or the White House Executive Order on AI. The first section compares and contrasts the affordances of generative AI and social media to highlight their similarities and differences. The second section discusses specific policy recommendations based on the evolution of social media and their regulation.

2 Affordances of generative AI and social media

To shed light on the similarities and differences between specific media, we can analyze their affordances. Affordances are the features that characterize a medium. 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), I 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's features were originally meant to capture affordances of face-to-face communication,[1] I added new features and removed features that are less relevant to the comparison of generative AI and social media. I will point out each feature that is adapted from Clark.

I will first address why social media are an imperfect analogy to generative AI, and then outline why this analogy can still be very helpful to highlight key features that can inform regulation.

2.1 Generative AI and social media are not perfectly comparable

By definition, an analogy is not a perfect match, otherwise the objects of comparison would be the same thing. As Jacob Stern puts 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 for the user and interacts in a dialogue with the user, social media are merely mediating between the user and their human conversation partner (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. Relatedly, while social media foster direct connections between users, generative AI is usually used by a single person at a time. 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 scroll or auto-play that serve content as long as the user is on the platform. The purpose of social media tends to be limited to 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

[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's model that are pertinent to generative AI and social media, but not the comparison to other media.

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
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). Instances where features of generative AI are similar to features of social media are highlighted in bold.

for generative AI, which operates in a sequential dialogue of user input and model output. Important differences between generative AI and social media concern content moderation: Even though both generative AI and social media feature content moderation, content moderation in generative AI tends to be less visible than on 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, while prominent generative AI companies have so far leaned towards freemium subscription models. While the potential harm of social media to democracy and society has been an important focus of scholarly and public attention (Persily and Tucker, 2020), some argue that the destructive potential of AI may be at another level since it may present a larger threat 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, Ghosh and Kaffee, 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.

2.2 Generative AI and social media are comparable in key aspects

The analogy between generative AI and social media is valuable despite its imperfection because their key features are similar. 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 of face-to-face communication in Clark (1996)). Both generative AI and social media are recording user data (the recording feature is adapted from Clark's recordlessness feature). Both media can learn about a user's context and their preferences over time to personalize their output, e.g. by updating the chatbot's memory or personalizing a recommendation algorithm. Further, both generative AI and social media can feature 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 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'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 money (Stern, 2023).

3 Learnings from social media regulation for generative AI regulation

As the review of the affordances has shown, generative AI and social media share important features, including the use of AI and content moderation. Although generative AI and social media differ on some dimensions, these differences suggest, for the most part, differences in degree, and not differences in kind when it comes to regulation. Thus, lessons learnt from social media regulation may be relevant to generative AI regulation. This paper addresses four policy recommendations for generative AI regulation based on the evolution of social media regulation: (1) counter bias and perceptions thereof (e.g., via transparency, oversight boards, researcher access, democratic input, multidisciplinary research), (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.

3.1 Counter bias and perceptions thereof

Given that both generative AI and social media share the key features content moderation, use of AI, that they are black-box and abstract the complexity of algorithmic decision-making away 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 and Sims, 2021). While there is no evidence of anti-conservative bias for social media (Barrett and Sims, 2021), multiple studies have shown political bias in generative AI. For example, compared to representative opinion polls, large language models were found to output biased opinions (Durmus et al., 2023; Santurkar et al., 2023), and multiple studies showed left-leaning bias in generative AI such as ChatGPT (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, Farooqi and Zou, 2021), bias towards Western culture (Naous, Ryan and Xu, 2023), 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 users by leading to lower-quality output, they can entrench historical biases and stereotypes, and they can undermine trust in model developers, model deployers, and regulators of generative AI. It is challenging to address biases because they are challenging to measure accurately (e.g., they may be sensitive to the specific prompt design (Röttger et al., 2024)) and because it is not clear where exactly biases stem from. Biases can arise at different points in the development and deployment of generative AI, including training and data curation, fine-tuning, evaluation and feedback, real-time moderation, customization and control of models (Suresh and Guttag, 2019; 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.

3.1.1 Increase transparency and researcher access

The features content moderation, use of AI, black-box and abstraction also give rise to transparency challenges for social media and generative AI. Generative AI transparency has been poor as shown in the Foundation Model Transparency Index (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), created blog posts and a Transparency Center providing some information on the role of AI and other factors in content recommendation (Clegg, 2023; Meta, 2024*a*), and established an independent oversight board of experts that adjudicates particularly contentious content moderation decisions (Meta, 2024*b*). 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).

An important aspect of transparency is allowing for third-party evaluations. Efforts to create APIs accessible to researchers, such as the Facebook Open Research and Transparency Researcher API and the TikTok Research API, or to design academic-industry collaboration such as the Facebook and Instagram Election Study 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 feasibility, helpful documentation, privacy preserving measures (e.g., aggregation of user data) and timeliness in terms of publication or addressing of issues that researchers discovered. To protect researchers involved, researchers have called for "safe harbors," that is, 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). Additional proposals to facilitate external generative AI research include data donations (Sanderson, 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, Engel and Mauer, 2023).

While the specific implementation of these transparency efforts may be contentious and requires nuance, ideas such as short and accessible explanations of the technology, independent oversight mechanisms, researcher access and mandatory audits are viable options for increasing transparency via generative AI regulation.

3.1.2 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. Given the vast set of choices that developers face, one approach is to gather input directly from users to determine what a good system may look like. In terms of gathering input from users to enable democratic decisions about the nature of regulation and content moderation, different initiatives have been launched over the past few years to deliberate issues ranging from cyberbulling on social platforms to the rules and constitutions that inform generative AI models (Wetherall-Grujić, 2023), relying on the much older idea of deliberative democracy (Eagan, 2016). 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) or a neighborhood-focused social network (Oremus, 2024). Finally, social media researchers have studied how to embed important societal values into AI (Bernstein et al., 2023), which could also inform how such values can be embedded into generative AI.

3.1.3 Promote user choice

Another option to empower users to make choices in the face of features like 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 (Redpoint, 2020). This suggests that increased personalization of models may be an answer (Redpoint, 2020).

3.2 Address specific regulatory concerns and invest in trust and safety

The feature of content moderation that generative AI and social media share comes with especially thorny issues 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, Meta and Microsoft working on youth wellbeing and mental health in general, 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 overall.

Generative AI chatbots have already been rated on AI-related principles that apply just as much to social media. Common Sense Media published rankings of different generative AI models with regards to the following principles: put people first, prioritize fairness, be trustworthy, keep kids and teens safe, be effective, help people connect, use data responsibly, and be transparent and accountable (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 seems important. 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, but also more cautious interventions such as warnings and strikes for misbehavior, putting more guardrails in place or throttling usage for users that tried to abuse generative AI models in the past. 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.

3.3 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. These features suggest that both of these media are so important and powerful because of how they interact with users. 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.

In fact, the recommendations above, whether regarding measures to reduce bias or enhance user wellbeing, all require computational social science research to test their effectiveness. 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.

Whether research is conducted in-house or via access to the platforms for external researchers, rigorous evaluations are key to ensure that media like 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 does have features that differ from previous technologies, so its impact and user preferences (e.g., with regards to privacy, personalization or content moderation) are not clear. Even the impact of previous technologies like 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 investing in diverse research teams that understand the interaction of humans an a given medium and that can evaluate the societal implications of a product. 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 addressing current risks such as biases and of creating multidisciplinary teams to do so. This allows companies to test different product features and interventions effectively, e.g. to reduce spam or misinformation spread. Computational social scientists from any background, data scientists and user experience researchers would be especially helpful to address questions at the intersection of technology and humans, such as which emotional bonds may be formed between humans and AI, and what type of personalization could be implemented.

While content moderation on social media is far from a resolved issue, there is a large and growing body of academic literature that speaks to promising approaches and could inform content moderation for generative AI (e.g. Persily and Tucker, 2020; Kozyreva et al., 2022).

3.4 Take on a more global perspective

As the features spatial separation, general content, and use of AI imply, 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, generative AI companies, at least the startups among them, seem more heavily focused on the US due to their location (with exceptions like Google DeepMind in the UK and Mistral AI in France). To address problems like biases, it is crucial that even small companies take on a global perspective and become global companies with local expertise in multiple countries. This could take the form of local offices and a focus on hiring internationally. 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 (e.g., due to under-investment in non-English language content generation) and could even result in catastrophes such as promoting 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.

For effective regulation, local expertise needs to be integrated into a global perspective. For example, the former Prime Minister of New Zealand suggested that a model for governing AI could follow the Christchurch Call, which is a multinational, multi-stakeholder effort bringing together governments, tech companies and civil society to eliminate violent extremist and terrorist content online (Ardern, 2023).

4 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), and by no means a perfect analogy, generative AI and social media share key features that make a comparison of the two worthwhile. Taking a close look at social media regulation efforts — including self-regulation and laws — reveals interesting approaches and best practices. 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. In the case of social media, self-regulation did not always work, which has resulted in multiple new laws being proposed in the past few years. These laws, but also the forms of self-regulation that were put in place, including specific approaches to increasing transparency, enhancing user choice, and investing in research, can be valuable pointers for those looking to regulate generative AI. Analyzing social media regulation may speed up the process of developing generative AI regulation. The EU AI Act may not have been able to address general purpose models as fast as it did had it not already been concerned with other forms of machine learning much earlier. Regulation takes time and effort, so where possible, resources should be saved and mistakes avoided by looking at social media regulation and research.

Acknowledgments and Disclosure of Funding

The author has worked on two research projects in collaboration with Meta. The author attended a Meta event where food was paid for by the company. The author is grateful to Jennifer Pan for feedback on an earlier draft.

References

Abdo, Alex, Ramya Krishnan, Stephanie Krent and Andrew Keane Woods. 2022. "A Safe Harbor for Platform Research.".
URL: *https://knightcolumbia.org/content/a-safe-harbor-for-platform-research*

Abid, Abubakar, Maheen Farooqi and James Zou. 2021. "Large language models associate Muslims with violence." *Nature Machine Intelligence* 3(6):461–463.
URL: *http://dx.doi.org/10.1038/s42256-021-00359-2*

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

Ardern, Jacinda. 2023. "There' s a model for governing AI. Here it is.".
URL: *https://www.washingtonpost.com/opinions/2023/06/09/jacinda-ardern-ai-new-zealand-planning/*

Barrett, Paul M. and J. Grant Sims. 2021. False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives. Technical Report February NYU Stern Center for Business and Human Rights.
URL: *https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/6011e68dec2c7013d3caf3cb/1611785871154/NYU+False+Accusation+report_FINAL.pdf*

Bernstein, Michael S, Angèle Christin, Jeffrey T Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, Jeanne L Tsai, Johan Ugander and Chunchen Xu. 2023. "Embedding Societal Values into Social Media Algorithms." *Journal of Online Trust and Safety* 2(1):1–13.

Bommasani, Rishi, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej and Percy Liang. 2024. "The Foundation Model Transparency Index v1.1.".

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang and Percy Liang. 2023. "The Foundation Model Transparency Index.".
URL: *http://arxiv.org/abs/2310.12941*

Cattell, Sven, Avijit Ghosh and Lucie-Aimée Kaffee. 2024. "Coordinated Disclosure for AI: Beyond Security Vulnerabilities.".
URL: *http://arxiv.org/abs/2402.07039*

Clark, Herbert H. 1996. *Using Language*. Cambridge University Press.

Clegg, Nick. 2023. "How AI Influences What You See on Facebook and Instagram.".
URL: *https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/*

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

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

Durmus, Esin, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark and Deep Ganguli. 2023. "Towards Measuring the Representation of Subjective Global Opinions in Language Models.".
URL: *http://arxiv.org/abs/2306.16388*

Eagan, Jennifer L. 2016. "deliberative democracy.".
URL: *https://www.britannica.com/topic/deliberative-democracy*

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

Ferrara, Emilio. 2023. "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models.".
URL: *http://arxiv.org/abs/2304.03738*

Hacker, Philipp, Andreas Engel and Marco Mauer. 2023. Regulating ChatGPT and other Large Generative AI Models. In *ACM International Conference Proceeding Series*. pp. 1112–1123.

Huttenlocher, Dan, Asu Ozdaglar and David Goldston. 2023. A Framework for U.S. AI Governance: Creating a Safe and Thriving AI Sector. Technical report MIT Schwarzman College of Computing.
URL: *https://computing.mit.edu/wp-content/uploads/2023/11/AIPolicyBrief.pdf*

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan Herzog, Ullrich Ecker, Stephan Lewandowsky and Ralph Hertwig. 2022. "Toolbox of Interventions Against Online Misinformation and Manipulation." *PsyArXiv Preprints* pp. 1–24.
URL: *https://psyarxiv.com/x8ejt/*

Longpre, Shayne, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng Xin Yong, Suhas Kotha, Yi Zeng, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Alex Pentland, Arvind Narayanan, Percy Liang and Peter Henderson. 2024. Position: A Safe Harbor for AI Evaluation and Red Teaming. In *Proceedings of the 41st International Conference on Machine Learning*, ed. Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett and Felix Berkenkamp. Vol. 235 of *Proceedings of Machine Learning Research* PMLR pp. 32691–32710.
URL: *https://proceedings.mlr.press/v235/longpre24a.html*

Maas, Matthijs M. 2023. "AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy." *SSRN Electronic Journal* (October).

Mastodon. 2024. "Mastodon.".
URL: *https://joinmastodon.org/*

Meta. 2024*a*. "Our Approach to Facebook Feed Ranking How Feed Ranking Works for Connected Content.".
URL: *https://transparency.fb.com/features/ranking-and-content/*

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

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

Nangia, Nikita, Clara Vania, Rasika Bhalerao and Samuel R. Bowman. 2020. "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models." *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* pp. 1953–1967.

Naous, Tarek, Michael J. Ryan and Wei Xu. 2023. "Having Beer after Prayer? Measuring Cultural Bias in Large Language Models.".
URL: *https://arxiv.org/abs/2305.14456v1*

Nimmo, Ben. 2024. AI and Covert Influence Operations: Latest Trends. Technical Report May OpenAI.
URL: *https://downloads.ctfassets.net/kftzwdyauwt9/5IMxzTmUclSOAcWUXbkVrK/3cfab518e6b10789ab8843bcca18b633/Threat_Intel_Report.pdf*

Oremus, Will. 2024. "The friendliest social network you've never heard of.".
URL: *https://www.washingtonpost.com/technology/2024/08/10/front-porch-forum-vermont-research-new-public/*

Persily, Nathaniel and Joshua A. Tucker, eds. 2020. *Social Media and Democracy*. Cambridge University Press.
URL: *https://www.cambridge.org/core/books/social-media-and-democracy/E79E2BBF03C18C3A56A5CC393698F117*

Redpoint. 2020. "Stanford Professor Tatsu Hashimoto on AI Biases and Improving LLM Performance.".
URL: *https://www.youtube.com/watch?v=pceYeZdT1D0*

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

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

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze and Dirk Hovy. 2024. "Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models.".
URL: *http://arxiv.org/abs/2402.16786*

Rozado, David. 2023. "The Political Biases of ChatGPT." *Social Sciences* 12(3).

Sanderson, Zeve. 2024. "Beyond Competition: Designing Data Portability to Support Research on the Digital Information Environment." *SSRN Electronic Journal* .

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning (ICML, oral)*.
URL: *http://arxiv.org/abs/2303.17548*

Stern, Jacob. 2023. "AI Is Like . . . Nuclear Weapons ?" *The Atlantic* .
URL: *https://www.theatlantic.com/technology/archive/2023/03/ai-gpt4-technology-analogy/673509/*

Suresh, Harini and John V. Guttag. 2019. "A Framework for Understanding Unintended Consequences of Machine Learning.".
URL: *http://arxiv.org/abs/1901.10002*

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

Thulasi, S. 2019. "Understand Why You're Seeing Certain Ads and How You Can Adjust Your Ad Experience.".
URL: *https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/*

Wagner, Michael W. 2023. "Independence by permission." *Science* 381(6656):388–391.

Wartella, Ellen and Byron Reeves. 1985. "Historical Trends in Research on Children and the Media: 1900-1960." *Journal of Communication* 35(2):118–133.

Wetherall-Grujić, Graham. 2023. "The Race to Democratise AI.".
URL: *https://democracy-technologies.org/participation/the-race-to-democratise-ai/*

Wojcik, Stefan, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman and Jay Baxter. 2022. "Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation." *Proceedings of ACM Conference (Conference'17)* 1(1).
URL: *http://arxiv.org/abs/2210.15723*