SOCIAL SCIENCE IS NECESSARY FOR OPERATIONALIZING SOCIALLY RESPONSIBLE FOUNDATION MODELS

Adam Davies
Siebel School of Computing and Data Science
The Grainger College of Engineering
University of Illinois Urbana-Champaign
adavies4@illinois.edu

Elisa Nguyen
Tübingen AI Center
University of Tübingen

Michael Simeone
School of Complex Adaptive Systems
Arizona State University

Erik Johnston
School for the Future of Innovation in Society
Arizona State University

Martin Gubri
Parameter Lab

ABSTRACT

With the rise of foundation models, there is growing concern about their potential social impacts. Social science has a long history of studying the social impacts of transformative technologies in terms of pre-existing systems of power and how these systems are disrupted or reinforced by new technologies. In this position paper, we build on prior work studying the social impacts of earlier technologies to propose a conceptual framework studying foundation models as sociotechnical systems, incorporating social science expertise to better understand how these models affect systems of power, anticipate the impacts of deploying these models in various applications, and study the effectiveness of technical interventions intended to mitigate social harms. We advocate for an interdisciplinary and collaborative research paradigm between AI and social science across all stages of foundation model research and development to promote socially responsible research practices and use cases, and outline several strategies to facilitate such research.

1 INTRODUCTION

While the rapid recent development of generative foundation models is exciting for many potential applications (see, e.g., Touvron et al. 2023; Jiang et al. 2023; Aryabumi et al. 2024; Dubey et al. 2024, etc.), important social impacts come along with rapid adoption, including worker displacement (Ludec et al., 2023; Casilli, 2025; Capraro et al., 2024), use of copyrighted data for training models (Carlini et al., 2021; Somepalli et al., 2022; Samuelson, 2023; Grynbaum & Mac, 2023), energy requirements and associated climate impact (Tamburrini, 2022; Luccioni & Hernandez-Garcia, 2023), and data privacy (Carlini et al., 2021; Jo & Gebru, 2020b; Nasr et al., 2023; Kim et al., 2023). To develop socially responsible foundation models, we argue for *proactive* consideration of such concerns across the whole research and development (R&D) lifecycle from ideation to retirement of the technology. Anticipating social concerns can enable early discovery of unintended problems in the pipeline (e.g., biased data collection – see Sambasivan et al., 2021) and inform interventions to mitigate undesired impacts (Hardt et al., 2016; Bommasani et al., 2021; Sun et al., 2019).

Understanding and considering social impacts in the research and development of AI technology requires knowledge and experience studying complex social systems and interactions – i.e., expertise in social science. However, the domains of AI and social science research are largely siloed (Selbst et al., 2019; Dahlin, 2021; Sartori & Theodorou, 2022), manifesting in differences in vocabulary (Krafft et al., 2020), publishing venues, and publishing practices (e.g., the prestige of journals vs. conferences). Often, simplifying assumptions are made in AI research about social structures which may not hold in real life (Sartori & Theodorou, 2022) – e.g., crowdsourcing annotators to align

LLMs with so-called "human values" via reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ouyang et al., 2022), despite the fact that such "values" are unique to the individual and may vary widely across cultures. To facilitate more socially responsible research, we advocate for an interdisciplinary paradigm integrating expertise in AI and social science throughout the technology lifecycle to anticipate and study potential social impacts of foundation models. First, we explore several relevant notions from social science to better contextualize how new technologies can impact society, highlighting how past failures to anticipate sociotechnical impacts have led to real social harms. Building on these ideas, we propose a conceptual framework for integrating social science in foundation model research to understand social responsibilities, anticipate potential impacts, and develop technical innovations to create and deploy more socially responsible foundation models. Finally, we consider incentives for tech firms and individual researchers that encourage or inhibit research and development toward socially responsible foundation models, and indicate several actionable suggestions to promote interdisciplinary collaboration between AI and social sciences through incentives, education, and skill development.

2 BACKGROUND

Social Systems of Power In one relevant intellectual tradition of social science, *systems of power* – the structures and institutions that shape, maintain, and distribute power within a society – have been researched and described across a variety of theoretical paradigms and approaches including post-structuralism, socio-cultural theory, network analysis, and organizational theory (Linstead, 2003; Roberts, 2012; Martin, 2024). Scholars explore how institutions (such as governments, corporations, and social norms) distribute power and privilege, shaping social outcomes like prosperity, inequality, and marginalization. Many approaches starting toward the end of the 20th century emphasize the interconnectedness of race, class, gender, and other identities in understanding power dynamics (Crenshaw, 1991; Collins, 2000). These systems of power can be reproduced by new technologies such as social media platforms, recommendation algorithms, and search engines, which can amplify existing biases by encoding them into technological systems (Eubanks, 2018; Noble, 2018; Benjamin, 2019).

Technological Affordances In an oft-cited example of how racial systems of power can be codified in technology, Noble (2018) examines how Google Search in 2011-2013 reinforced longstanding harmful misrepresentations of Black women in, e.g., racist and sexist stereotypes that appeared in top autocompletions beginning with "why are black women so" versus "why are white women so", and the hyper-sexualization of Black women evidenced by the extreme prevalence of pornographic results when queried with "black girls". We may understand such instances through the lens of *technological affordances* – i.e., the technology-mediated actions that are enabled, encouraged, or constrained by a technology with respect to an environment (Jones, 2020) – in the specific context of information seeking (Zhao et al., 2020; Hirvonen et al., 2023), where the actions taken by web users (e.g., selecting an autocompletion or following a search result) are influenced by technologies that can implicitly reproduce existing systems of power (such as harmful stereotypes or sexual objectification), reciprocally shaping the digital information environment by driving search traffic and influencing users' beliefs to reinforce the social harms and inequities embedded in the technology (Vicente & Matute, 2023). In this work, we consider the technological affordances of foundation models, and the importance of social science for understanding how these affordances can reproduce or reshape existing systems of power.[1]

Social Media and Teen Mental Health Before discussing foundation models, we first consider a more established technology where failing to take findings from social science and psychology into account has led to serious real-world harms: social media use (SMU) among teens, and its impact on their mental health. Many studies have found a strong correlation between SMU and diagnoses of mood and body-image disorders (Barry et al., 2017; Gupta et al., 2022; Costello et al., 2023; Weigle & Shafi, 2024); and while it is difficult to establish a direct causal relationship, the limited causal evidence available suggests that SMU is indeed an important contributor to these negative impacts (Bozzola et al., 2022; Weigle & Shafi, 2024). One possible solution that has been proposed to help mitigate such harms is to redesign content recommendation feeds to de-prioritize engagement metrics, as there is clear evidence that recommender systems optimized for user engagement suggest harmful

[1]Note that, while a primary focus of our work is the relevance of social science research to such considerations, the traditional subject matter and methods of humanities research are similarly critical (Klein et al., 2025).

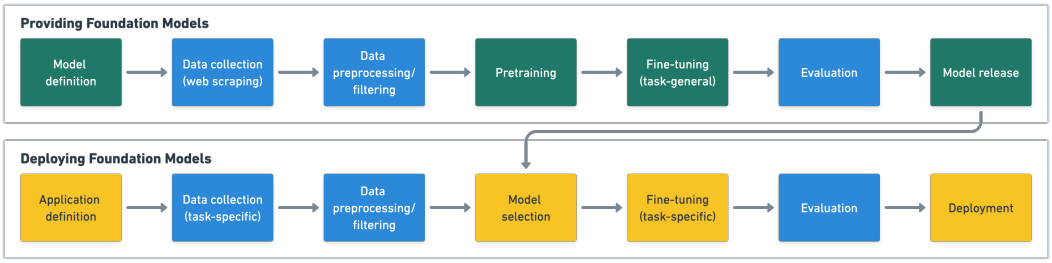

Figure 1: **Steps of the Foundation Model R&D pipeline.** The top pipeline illustrates the stages for training a foundation model (providers), while the bottom pipeline describes the stages of deploying foundation models (deployers).

content at a far higher rate than systems that do not (Banker & Khetani, 2019). For instace, despite early internal user studies conducted at Facebook and Instagram finding that simple adjustments to engagement-based algorithmic design choices could indeed mitigate negative impacts on teen mental health (Wells et al., 2021; Hao, 2021), the teams conducting this research were shuttered and the corresponding changes were never adopted at scale because they also led to lower advertising revenues (Hao, 2021; Mac & Kang, 2021; Costello et al., 2023; Protecting Kids Online, 2021).

The Foundation Model Pipeline Throughout this work, we will focus on a more recent, potentially socially-transformative technology, *foundation models* (i.e., self-supervised deep learning models trained on large-scale web data, such as LLMs). The process of creating or deploying foundation models can be visualized as a pipeline representing the different stages of the research and development (R&D) process,[2] as visualized in Figure 1, where decisions in each step of the pipeline can carry important consequences for later stages – for example, sub-optimal data collection and filtering choices can have serious implications for downstream model robustness and lead to preventable social harms (Sambasivan et al., 2021). Following the EU AI Act (European Parliament, 2023), we categorize model *providers* as those who develop a general-purpose AI model; and model *deployers* as those who develop a product or service leveraging such models for a specific use case where providers can also be deployers of their own models (e.g., OpenAI is the provider of ChatGPT, and a company that calls the ChatGPT API in a user-facing product would be a deployer).

3 Operationalizing Socially Responsible Foundation Models

To provide and deploy foundation models in a socially-responsible manner, we argue that it is necessary to involve social science expertise throughout the foundation model R&D process. In particular, we propose a conceptual framework to decompose this task into three key components:

1. **Understanding systems of power:** How do disruptive technologies like foundation models reproduce or reshape existing systems of power? What affordances would best promote desirable effects on these systems?
2. **Designing technical interventions:** How can the foundation model R&D pipeline be modified to align models with target affordances?
3. **Anticipating social impacts:** What social impacts may result from deploying a model with the target affordances in a specific context?

While AI researchers are well-positioned to study technical interventions (2), this is an entirely different question from understanding systems of power (1) or anticipating social impacts (3), which are better suited to social scientists. However, there is still a role in each component for AI research, as it is nonetheless important to provide robust, quantifiable, and computationally tractable definitions of desired foundation model affordances (e.g., it is necessary to specify affordances in terms that can be learned by models in encouraging socially representative model outputs, prohibiting the use of models for generating toxic content, etc.), as well as to carry out systematic empirical evaluations of corresponding model behaviors to predict alignment with the intended affordances, which are both

[2]Throughout this work, we use foundation model *research and development (R&D)* very broadly, where "research" is intended to cover all aspects of foundation model research – e.g., from basic research involving model architectures, loss functions, fine-tuning paradigms, etc., all the way to benchmarking existing models or developing applied techniques to improve models' performance for specific tasks.

tasks where AI expertise is essential. As such, interdisciplinary collaboration between AI and social science is required to address the challenges associated with each of these components. Specifically, we argue that it is critical to involve social science in foundation model research, development, and deployment in order to (1) *proactively consider interactions between foundation model affordances and sociotechnical systems of power,* and (2) *anticipate the impacts associated with deploying these models in a given context,* as explored below.

Responsible Model Providers Proactively Consider Systems of Power. Large scale, web-scraped data is an essential igredient for training all foundation models; and such data is shaped by sociotechnical systems of powers in subtle, complex, and systematic ways. For instance, Wikipedia, which has been heavily relied upon as a large and high-quality knowledge resource in many LLM training datasets (Touvron et al., 2023; Gao et al., 2021; Soldaini et al., 2024), underrepresents women and non-binary figures (Graells-Garrido et al., 2015; Hube, 2017; Falenska & Çetinoğlu, 2021; Tripodi, 2023; Ferran-Ferrer et al., 2023) – e.g., only 19% of biographies are about women (Tripodi, 2023). This *Wikipedia gender gap* is well studied in social science (see Ferran-Ferrer et al. 2023 for a comprehensive survey on the topic) as a complex systemic phenomenon. Using the conceptual framework of fields of visibility, Beytía & Wagner (2022) analyze content asymmetries on Wikipedia as a system composed of diverse agents affecting content in terms of representation, characterization, and structural placement. For model providers to avoid reinforcing systematic under- and mis-representation, it is important to be aware of such phenomena and act to mitigate resulting bias (e.g., by actively collecting under-represented data Jo & Gebru, 2020a; or implementing debiasing techniques Mehrabi et al., 2021; Parraga et al., 2025). However, such techniques are not a "silver bullet" solution (Anwar et al., 2024), given the wide variety of statistical notions of bias that can be contradictory and entail tradeoffs (Verma & Rubin, 2018; Carey & Wu, 2023) and the problematic simplifying assumptions required to mitigate biased representation by way of statistical methods (Bode, 2020). Thus, it is a key responsibility of model providers to study and transparently communicate learned biases to downstream model deployers, as understanding and documenting the sources and effects of potential biases can provide the necessary context for selecting the application areas of a model (Sherman et al., 2024; Klein et al., 2025). For example, Mitchell et al. (2019) argue that models should be distributed alongside *model cards* that report metrics at a disaggregated level for cultural, demographic, or phenotypic population groups,[3] and Klein et al. (2025) further advocate for detailed documentation of data collection and/or generation procedures.

To illustrate the importance of considering systems of power for all stages in the foundation model pipeline, consider a scenario where these systems are not taken into account by model providers. Here, whatever systemic inequities are present in the model's training corpus (e.g., under-representation of women, harmful stereotypes of racial or ethnic minorities, etc.) can easily be learned and reproduced by the model, naturally affording corresponding harmful use cases (Sambasivan et al., 2021; Weidinger et al., 2021). Despite the common counter-argument that web-scraped data simply reflects the reality of what content appears on the web, and that it is not the responsibility of model providers to mitigate any given notion of bias in one's pre-training corpus (as highlighted by Birhane et al. 2023), the alternative *laissez faire* approach, where systems of power are not taken into account whatsoever, can lead to an avoidable "race-to-the-bottom" collective action problem among model providers, deployers, and end-users. In this case, each deployer utilizing the provided model would need to decide whether and how to account for social risks or harms on their own, and those who make the greatest effort to mitigate them will incur a greater time and cost in doing so relative to less scrupulous competitors. That is, where many deployers might *prefer* that bias had been better mitigated by providers, it may not be possible for them to take on this task on their own while maintaining competitiveness; and ethics-minded employees may be disempowered to take collective action (Nedzhvetskaya & Tan, 2024) due to financial precarity, immigration status, workplace culture, or organizational incentives (Widder et al., 2023). Similarly, from a user's perspective, risks and harms might only be addressed (if at all) after some level of harm has already been done (given the competitive disadvantage associated with anticipating and proactively addressing possible harms); and providers will face a lack of trust in the safety of their models on the part of deployers and end users (Keymolen, 2024). In such cases, we argue that there should be a *duty of care* (Witting, 2005; Arbour, 2008; Welsh, 2012) to anticipate, transparently communicate, and act to mitigate the propagation of discriminatory (or otherwise harmful) systems of power to avoid the social dilemma

[3]E.g., model providers can report intended uses and potential limitations using Hugging Face's model and dataset card features, inspired by (Mitchell et al., 2019).

described above. We further consider pragmatic motivations for model providers to address such concerns in Section 4.

Responsible Model Deployers Address Application-Specific Social Impacts. In deploying existing foundation models for a specific application context, model deployers are best placed to consult social scientists in (a) anticipating potential social impacts associated with their specific intended application, and (b) designing and studying effective affordances, where various foundation model applications require expertise from different disciplines in social science. For example, consider an application leveraging foundation models to edit photos before they are posted to social media. Many popular social media platforms already afford users to edit selfies using "beauty filters" (Eshiet, 2020; Ryan-Mosley, 2021) that modify their appearance to align with a socially-constructed representation of conventional attractiveness or high social status (Javornik et al., 2022; Burnell et al., 2022). If we consider societies where such a notion includes being thin, then these filters are expected to reduce the apparent weight of users in photos (Eshiet, 2020; Ateq et al., 2024); or in the context of societies where this notion is associated with lighter skin tone, these filters have been shown lighten the apparent skin tone of users (Riccio et al., 2024; Trammel, 2023). In the former case, the filter affordance reinforces a culture of "fatphobia", stigmatizing heavier individuals and creating unrealistic body standards (Robinson et al., 1993); and in the latter case reinforces racial caste systems, such as White supremacy (Bonilla-Silva, 2001). Indeed, filter affordances predating the era of generative foundation models have already been implicated in teen body image disorders (Burnell et al., 2022; Tremblay et al., 2021), and it is reasonable to expect that more powerful generative models will potentially lead to further such harms. As such, just as in the case explored in Section 2 on the relationship between social media use and teen mental health, social psychologists should likewise be consulted in this case to anticipate the potential impacts of foundation model-enabled filters, and in studying the effectiveness of possible interventions to mitigate harmful affordances.

Note that, while we have focused here on social media platform affordances enabled by foundation models, analogous arguments can be made for many other aspects of society and require expertise from different disciplines in social science. For instance, in the workplace, foundation model-enabled affordances are predicted to carry widely varying net impacts on wages and labor markets depending on the speed and manner in which various workplace tasks are automated (Acemoglu et al., 2024; Acemoglu & Johnson, 2024); in education, they are expected to help democratize education worldwide while also leading to broader and more systemic bias in educational assessment and college admissions (Akgun & Greenhow, 2022; Baker & Hawn, 2022) or exacerbating the digital divide (Capraro et al., 2024; Mannekote et al., 2024); and so on. Each of these considerations requires consultation with relevant domain-area experts to anticipate potential impacts and design mitigation strategies.

4 Facilitating Socially Responsible Foundation Models

Despite the rationale and approach for researching and developing more socially-responsible foundation models articulated above, it is unrealistic to expect all stakeholders to opt for such an approach on ethical merit alone, as doing so may conflict with other incentives such as short-term profits. Below, we consider incentives for (1) tech firms providing and deploying foundation models, and (2) interdisciplinary AI + social science research, and suggest potential interventions that may aid in (re)structuring incentives in favor of interdisciplinary work toward more socially-responsible foundation models.

Incentives for Tech Firms Failure to anticipate and proactively address deleterious effects of social media use on teen mental health, as discussed in Section 2, has resulted in substantial brand harm and increased regulatory oversight for social media companies (Wells et al., 2021; Hao, 2021; Costello et al., 2023). In contrast, there is evidence to suggest that tech firms may benefit financially from prioritizing social responsibility in providing and deploying foundation models. As outlined by Gillan et al. (2021), there is a large and growing body of research in financial economics suggesting that more socially-responsible firms tend to see superior financial performance and stability in the long term – specifically, that the Environmental, Social, and Governance (ESG) and the Corporate Social Responsibility (CSR) profiles of firms are strongly related to lower risk, higher performance, and higher value. For example, Lins et al. (2016) show that high-CSR firms had better stock returns, profitability, growth, and sales per employee, compared to low-CSR firms during the 2008–2009 financial crisis, suggesting that investments in social capital can pay off in times of economic crisis.

Furthermore, Hong et al. (2019) estimate that, in the aggregate, high-ESG firms face 65% lower sanctions from prosecutors. Thus, we hypothesize that tech firms prioritizing social responsibility in providing and deploying foundation models may observe similar financial benefits.

Incentives for Interdisciplinary Collaboration Interdisciplinary collaboration between social science and AI research runs contrary to some key incentives for researchers' career advancement. For instance, interdisciplinary publication venues are unlikely to be among the top-tier venues in each respective discipline (Campbell, 2005); and while most top social science venues are journals, most top AI venues are conferences instead. As such, even the best interdisciplinary work is less likely to be adequately recognized, awarded, cited, and disseminated (Pellmar et al., 2000). Similar issues exist for other key factors in academic career advancement beyond publication venues, such as grant review (Bromham et al., 2016) and degree requirements for university students (Amelink et al., 2024). The following is a preliminary list of suggestions for attenuating the cost of interdisciplinary collaboration, though it is not intended to be exhaustive:

- As in the FAccT conference,[4] more AI conferences could offer the optional choice of non-archival paper submissions (in addition to the standard archival submission), allowing researchers from other fields to later submit their conference papers to discipline-specific journals.
- Research institutions could better consider interdisciplinary work in career advancement (Pellmar et al., 2000) and funding proposals assessment (Bromham et al., 2016), and offer specialized funding opportunities and sabbaticals, allowing researchers to explore new ideas and collaborations in a wider context (Ioppolo & Wooding, 2023).
- Existing practices for promoting socially-responsible research can be further promoted by publication venues (e.g., by making model and dataset cards Mitchell et al., 2019 a mandatory part of certain submission types) to expand their adoption as a standard in the research community.
- Promoting interdisciplinary education helps provide the next generation of researchers with the foundations to integrate methods from, and facilitate collaborations with, fields beyond their primary research area. For instance, awarding degree credit for courses from other disciplines encourages students to learn the essentials of these fields (Amelink et al., 2024), and embedding ethics education in technical courses can improve students' abilities to engage in relevant ethical discussions (Horton et al., 2022).
- Researchers considering a more interdisciplinary agenda could broaden their expertise with workshops, tutorials, or short courses provided by researchers from other fields. For example, we suggest that AI and social science conferences open tutorial calls to researchers outside their respective disciplines.

5 Conclusion

In this work, we have advocated for interdisciplinary research between AI and social science in the context of foundation models like LLMs, focusing on the importance of social science in understanding the affordances and social impacts of such transformative technologies. We outlined the importance of interdisciplinary expertise and collaboration throughout the foundation model R&D pipeline, highlighted the associated responsibilities and benefits for model providers and deployers, and provided actionable suggestions to promote collaboration between AI and social science. Finally, we discuss a few important considerations for future work in Appendix A.

Acknowledgements

We thank Alan Craig, Dave Buckley, and Linda Derhak for their help in facilitating this project and sharing valuable feedback. This work is supported in part by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award #2229612 (National AI Institute for Inclusive Intelligent Technologies for Education). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or the U.S. Department of Education. This material is based upon work supported in part by the National Science Foundation under Grant No. 2217706. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Elisa Nguyen. This work was supported in part by the Tübingen AI Center and the Parameter Lab company.

[4]See `https://facctconference.org/2024/cfp`.

REFERENCES

Daron Acemoglu and Simon Johnson. Learning from ricardo and thompson: Machinery and labor in the early industrial revolution and in the age of artificial intelligence. *Annual Review of Economics*, 16(1):597–621, 2024.

Daron Acemoglu, Fredric Kong, and Pascual Restrepo. Tasks at work: Comparative advantage, technology and labor demand. 2024.

Selin Akgun and Christine Greenhow. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI and Ethics*, 2(3):431–440, 2022.

Catherine T. Amelink, Dustin M. Grote, Matthew B. Norris, and Jacob R. Grohs. Transdisciplinary Learning Opportunities: Exploring Differences in Complex Thinking Skill Development Between STEM and Non-STEM Majors. *Innovative Higher Education*, 49(1):153–176, February 2024. ISSN 1573-1758. doi: 10.1007/s10755-023-09682-5. URL https://doi.org/10.1007/s10755-023-09682-5.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Louise Arbour. The responsibility to protect as a duty of care in international law and practice. *Review of International Studies*, 34(3):445–458, 2008.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

Khadijah Ateq, Mohammed Alhajji, and Noara Alhusseini. The association between use of social media and the development of body dysmorphic disorder and attitudes toward cosmetic surgeries: a national survey. *Frontiers in Public Health*, 12:1324092, 2024.

Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pp. 1–41, 2022.

Sachin Banker and Salil Khetani. Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, 38(4):500–515, 2019.

Christopher T Barry, Chloe L Sidoti, Shanelle M Briggs, Shari R Reiter, and Rebecca A Lindsey. Adolescent social media use and mental health from adolescent and parent perspectives. *Journal of adolescence*, 61:1–11, 2017.

Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019. ISBN 9781509526390. URL https://politybooks.com/bookdetail/?isbn=9781509526390.

Pablo Beytía and Claudia Wagner. Visibility layers: a framework for systematising the gender gap in Wikipedia content. *Internet Policy Review*, 11(1), March 2022. ISSN 2197-6775. URL https://policyreview.info/articles/analysis/visibility-layers-framework-systematising-gender-gap-wikipedia-content.

Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

Katherine Bode. Why you can't model away bias. *Modern Language Quarterly*, 81(1):95–124, 2020.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie,

Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Eduardo Bonilla-Silva. *White supremacy and racism in the post-civil rights era*. Lynne Rienner Publishers, 2001.

Elena Bozzola, Giulia Spina, Rino Agostiniani, Sarah Barni, Rocco Russo, Elena Scarpato, Antonio Di Mauro, Antonella Vita Di Stefano, Cinthia Caruso, Giovanni Corsello, et al. The use of social media in children and adolescents: Scoping review on the potential risks. *International journal of environmental research and public health*, 19(16):9960, 2022.

Lindell Bromham, Russell Dinnage, and Xia Hua. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684–687, 2016. doi: 10.1038/NATURE18315.

Kaitlyn Burnell, Allycen R Kurup, and Marion K Underwood. Snapchat lenses and body image concerns. *New Media & Society*, 24(9):2088–2106, 2022.

Lisa M Campbell. Overcoming obstacles to interdisciplinary research. *Conservation biology*, 19(2): 574–577, 2005.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, et al. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus*, 3(6), 2024.

Alycia N Carey and Xintao Wu. The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, 3(1):1–23, 2023.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. August 2021.

Antonio A. Casilli. *Waiting for Robots: The Hired Hands of Automation*. The France Chicago Collection. University of Chicago Press, Chicago, IL, January 2025. ISBN 978-0-226-82095-8. URL https://press.uchicago.edu/ucp/books/book/chicago/W/bo239039613.html.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Patricia Hill Collins. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York, 2 edition, 2000. ISBN 978-0-203-90005-5. doi: 10.4324/9780203900055.

Committee on Science and Public Policy and Committee on Facilitating Interdisciplinary Research. *Facilitating interdisciplinary research*. National Academies Press, 2005.

Nancy Costello, Rebecca Sutton, Madeline Jones, Mackenzie Almassian, Amanda Raffoul, Oluwadunni Ojumu, Meg Salvia, Monique Santoso, Jill R Kavanaugh, and S Bryn Austin. Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media. *American Journal of Law & Medicine*, 49(2-3):135–172, 2023.

Kimberle Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299, 1991. ISSN 0038-9765. doi: 10.2307/1229039. URL https://www.jstor.org/stable/1229039. Publisher: Stanford Law Review.

Emma Dahlin. Mind the gap! on the future of ai research. *Humanities and Social Sciences Communications*, 8(1):1–4, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Janella Eshiet. "real me versus social media me:" filters, snapchat dysmorphia, and beauty perceptions among young women. 2020.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, January 2018. ISBN 978-1-250-07431-7.

European Parliament. AI Act, 2023.

Agnieszka Falenska and Özlem Çetinoğlu. Assessing gender bias in Wikipedia: Inequalities in article titles. In Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster (eds.), *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pp. 75–85, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.9. URL https://aclanthology.org/2021.gebnlp-1.9.

Núria Ferran-Ferrer, Juan-José Boté-Vericad, and Julià Minguillón. Wikipedia gender gap: a scoping review. *El Profesional de la información*, 2023. URL https://api.semanticscholar.org/CorpusID:266344769.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

Stuart L. Gillan, Andrew Koch, and Laura T. Starks. Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal of Corporate Finance*, 66:101889, February 2021. ISSN 09291199. doi: 10.1016/j.jcorpfin.2021.101889. URL https://linkinghub.elsevier.com/retrieve/pii/S0929119921000092.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pp. 165–174, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333955. doi: 10.1145/2700171.2791036. URL https://doi.org/10.1145/2700171.2791036.

Michael M. Grynbaum and Ryan Mac. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*, December 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Chirag Gupta, Sangita Jogdand, and Mayank Kumar. Reviewing the impact of social media on the mental health of adolescents and young adults. *Cureus*, 14(10), 2022.

Karen Hao. The facebook whistleblower says its algorithms are dangerous. here's why. *MIT Technology Review*, 5(10):2021, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Noora Hirvonen, Ville Jylhä, Yucong Lao, and Stefan Larsson. Artificial intelligence in the information ecosystem: Affordances for everyday information seeking. *Journal of the Association for Information Science and Technology*, 2023.

Harrison G. Hong, Jeffrey D. Kubik, Inessa Liskovich, and José A. Scheinkman. Crime, Punishment and the Value of Corporate Social Responsibility, October 2019. URL https://papers.ssrn.com/abstract=2492202.

Diane Horton, Sheila A McIlraith, Nina Wang, Maryam Majedi, Emma McClure, and Benjamin Wald. Embedding ethics in computer science courses: Does it work? In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*, pp. 481–487, 2022.

Christoph Hube. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pp. 717–721. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349147. doi: 10.1145/3041021.3053375. URL https://doi.org/10.1145/3041021.3053375.

Becky Ioppolo and Steven Wooding. How academic sabbaticals are used and how they contribute to research – a small-scale study of the University of Cambridge using interviews and analysis of administrative data. *F1000Research*, 11:36, March 2023. ISSN 2046-1402. doi: 10.12688/f1000research.74211.2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10192944/.

Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, Graeme McLean, Yvonne Rogers, Paul Marshall, and Luk Warlop. 'what lies behind the filter?'uncovering the motivations for using augmented reality (ar) face filters on social media and their effect on well-being. *Computers in Human Behavior*, 128:107126, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Eun Seo Jo and Timnit Gebru. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 306–316, January 2020a. doi: 10.1145/3351095.3372829. URL http://arxiv.org/abs/1912.10389. arXiv:1912.10389 [cs].

Eun Seo Jo and Timnit Gebru. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 306–316, January 2020b. doi: 10.1145/3351095.3372829. URL http://arxiv.org/abs/1912.10389. arXiv:1912.10389 [cs].

Rodney Jones. Mediated discourse analysis and the digital humanities. In Svenja Adolphs and Dawn Knight (eds.), *The Routledge Handbook of English Language and Digital Humanities*, Routledge Handbooks in English Language Studies. Routledge, May 2020. URL https://centaur.reading.ac.uk/81524/.

Esther Keymolen. Trustworthy tech companies: talking the talk or walking the walk? *AI and Ethics*, 4(2):169–177, 2024.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: Probing Privacy Leakage in Large Language Models. In *NeurIPS 2023*, July 2023. URL http://arxiv.org/abs/2307.01881.

Lauren Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. Provocations from the humanities for generative ai research. *arXiv preprint arXiv:2502.19190*, 2025.

P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. Defining ai in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 72–78, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375835. URL https://doi.org/10.1145/3375627.3375835.

Karl V. Lins, Henri Servaes, and Ane Tamayo. Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis, October 2016. URL https://papers.ssrn.com/abstract=2555863.

Stephen Andrew Linstead. Organization theory and postmodern thought. 2003.

Alexandra Sasha Luccioni and Alex Hernandez-Garcia. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476*, 2023.

Clément Le Ludec, Maxime Cornet, and Antonio A Casilli. The problem with annotation. human labour and outsourcing between france and madagascar. *Big Data & Society*, 10(2): 20539517231188723, 2023. doi: 10.1177/20539517231188723.

Ryan Mac and Cecilia Kang. Whistle-Blower Says Facebook 'Chooses Profits Over Safety'. *The New York Times*, October 2021. ISSN 0362-4331. URL https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html.

Amogh Mannekote, Adam Davies, Juan D Pinto, Shan Zhang, Daniel Olds, Noah L Schroeder, Blair Lehman, Diego Zapata-Rivera, and ChengXiang Zhai. Large language models for whole-learner support: opportunities and challenges. *Frontiers in Artificial Intelligence*, 7:1460364, 2024.

R. Martin. *The Sociology of Power*. Routledge Revivals. Taylor & Francis, 2024. ISBN 9781003833826. URL https://books.google.com/books?id=N1YIEQAAQBAJ.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, January 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from (Production) Language Models, November 2023. URL http://arxiv.org/abs/2311.17035. arXiv:2311.17035 [cs].

Nataliya Nedzhvetskaya and J. S. Tan. The role of workers in ai ethics and governance. In *The Oxford Handbook of AI Governance*. Oxford University Press, 04 2024. ISBN 9780197579329. doi: 10.1093/oxfordhb/9780197579329.013.68. URL https://doi.org/10.1093/oxfordhb/9780197579329.013.68.

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN 978-1-4798-4994-9. doi: 10.2307/j.ctt1pwt9w5. URL https://www.jstor.org/stable/j.ctt1pwt9w5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 57(6):1–40, 2025.

Terry C Pellmar, Leon Eisenberg, et al. Barriers to interdisciplinary research and training. In *Bridging disciplines in the brain, behavioral, and clinical sciences*. National Academies Press (US), 2000.

Piera Riccio, Julien Colin, Shirley Ogolla, and Nuria Oliver. Mirror, mirror on the wall, who is the whitest of all? racial biases in social media beauty filters. *Social Media+ Society*, 10(2): 20563051241239295, 2024.

John Michael Roberts. Poststructuralism against poststructuralism: Actor-network theory, organizations and economic markets. *European Journal of Social Theory*, 15(1):35–53, 2012.

Beatrice "Bean" E Robinson, Lane C Bacon, and Julia O'reilly. Fat phobia: Measuring, understanding, and changing anti-fat attitudes. *International Journal of Eating Disorders*, 14(4):467–480, 1993.

Tate Ryan-Mosley. Beauty filters are changing the way young girls see themselves. *MIT Technology Review*, 2(2021):2021, 2021.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. 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: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023. doi: 10.1126/science.adi0656. URL https://www.science.org/doi/abs/10.1126/science.adi0656.

Laura Sartori and Andreas Theodorou. A sociotechnical perspective for the future of ai: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1):4, 2022.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Jihan Sherman, Romi Morrison, Lauren Klein, and Daniela Rosner. The power of absence: Thinking with archival theory in algorithmic design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 214–223, 2024.

Timothy F Slaper, Tanya J Hall, et al. The triple bottom line: What is it and how does it work. *Indiana business review*, 86(1):4–8, 2011.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.840.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, December 2022. URL http://arxiv.org/abs/2212.03860. arXiv:2212.03860 [cs].

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Guglielmo Tamburrini. The ai carbon footprint and responsibilities of ai scientists. *Philosophies*, 7 (1):4, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Juliana Maria Trammel. Artificial intelligence for social evil: Exploring how ai and beauty filters perpetuate colorism—lessons learned from a colorism giant, brazil. In *Black Communication in the Age of Disinformation: DeepFakes and Synthetic Media*, pp. 51–71. Springer, 2023.

Simon C Tremblay, Safae Essafi Tremblay, and Pierre Poirier. From filters to fillers: an active inference approach to body image distortion in the selfie era. *AI & society*, 36:33–48, 2021.

Francesca Tripodi. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, 25(7):1687–1707, 2023. doi: 10.1177/14614448211023772. URL https://doi.org/10.1177/14614448211023772.

US Senate Subcommittee on Consumer Protection, Product Safety, and Data Security. Protecting Kids Online: Testimony from a Facebook Whistleblower, October 2021. URL https://www.commerce.senate.gov/2021/10/protectingkidsonline:testimonyfromafacebookwhistleblower. Section: Hearings.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pp. 1–7, 2018.

Lucía Vicente and Helena Matute. Humans inherit artificial intelligence biases. *Scientific Reports*, 13 (1):15737, 2023.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Paul E Weigle and Reem MA Shafi. Social media and youth mental health. *Current psychiatry reports*, 26(1):1–8, 2024.

Georgia Wells, Jeff Horwitz, and Deepa Seetharaman. Facebook knows instagram is toxic for teen girls, company documents show. *The Wall Street Journal*, 14, 2021.

Jennifer M Welsh. Who should act?: Collective responsibility and the responsibility to protect. In *The Routledge Handbook of the Responsibility to Protect*, pp. 103–114. Routledge, 2012.

David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 467–479, 2023.

Christian Witting. Duty of care: an analytical approach. *Oxford Journal of Legal Studies*, 25(1): 33–63, 2005.

Yuxiang Chris Zhao, Yan Zhang, Jian Tang, and Shijie Song. Affordances for information practices: theorizing engagement among people, technology, and sociocultural environments. *Journal of Documentation*, 77(1):229–250, 2020.

A Future Work

An important consideration regarding the interventions suggested in Section 4 is that interdisciplinary research can be expensive and time-consuming (Pellmar et al., 2000), and bringing in diverse perspectives always carries the potential to dilute research focus with competing visions and priorities (Committee on Science and Public Policy and Committee on Facilitating Interdisciplinary Research, 2005). We suggest that future work could consider performing more comprehensive cost-benefit analyses along multiple dimensions (cf. Slaper et al., 2011) to assess the resources needed to achieve the benefits outlined above, making it possible to more effectively manage these research tradeoffs. More broadly, we recommend that AI experts and labs researching, developing, or deploying foundation models reflect on incorporating interdisciplinary collaboration within their team and their research topic more broadly, particularly in promoting socially responsible affordances and studying potential social impacts of their work. Neither AI nor social science holds all the answers regarding how to develop safe, beneficial, and socially responsible foundation models; and it is critical that both disciplines work more closely together toward this goal, rather than "siloing" research for such a potentially transformative technology.