Original Manuscript

Large AI Models Have a Prioritization Problem: Policy Implications and Solutions

Policy Insights from the Behavioral and Brain Sciences
1–9

DOI: 10.1177/23727322251408311
journals.sagepub.com/home/bbs
Sage

Joshua Conrad Jackson[1,2,†], Yuanze Liu[1,†], Zhao Wang[3], and William J. Brady[4]

Abstract

Artificial intelligence (AI) models trained on large corpora must prioritize some information over others. Distilling vast data into simple user-friendly representations can lead to a *prioritization problem*, in which large AI models neglect crucial information in their outputs. This prioritization problem manifests for newsfeed-ranking algorithms and generative AI such as large language models (LLMs). This paper discusses two flavors of this problem: (1) models trained to prioritize coherence or engagement may inadvertently prioritize misleading information over accurate information, and (2) models trained to prioritize prevalent information will underrepresent the heterogeneity of voices, reflecting cultural perspectives that dominate the training data. The prioritization problem can critically undermine knowledge, but it is not inevitable. Policymakers can address the prioritization problem by modifying the training material, training objectives, institutional safeguards, and messaging around large AI models.

Keywords

artificial intelligence, large language models, culture, social media, technology

Social Media

Large AI models simplify vast data by prioritizing some voices over others. Prioritization can pose a problem when models are trained on engagement or coherence rather than accuracy, and when prioritization homogenizes culture. New policies can address this problem.

Key Points

- Large AI models—like generative language models and newsfeed-ranking algorithms— are trained to distill vast high-dimensional data into simpler representations.
- Distillation requires prioritizing some information over others. While prioritization is inevitable, it can foster two knowledge problem.
- First, training models to prioritize coherence or engagement can inadvertently undermine accuracy.
- Second, training models to prioritize prevalent information homogenizes cultural perspectives and emphasize the voices of groups that dominate the training data.
- Academics and industry professionals can address the prioritization problem by creating culturally heterogeneous training materials, incentivizing variation during the training process, and making the model-building process more democratic and transparent.

Humans have long used institutions to compress complexity. Markets compress preferences, democracy compresses political opinions, and university degrees compress qualifications. In recent years, however, large artificial intelligence (AI) models have emerged as a novel form of information compression (Farrell et al., 2025). Large AI models are a new generation of AI built on deep neural network architecture. Their design gives them the capacity to reorganize and reconstruct large training datasets in transformative ways, thereby changing how people access and interpret information (Mitchell, 2020).

Large AI models are trained to compress their high-dimensional training data into simpler representations, which are lossy (incomplete, selective) but useful (Farrell, 2025; Farrell et al., 2025; Vaswani et al., 2017). For example, when someone asks a Large Language Model (LLM) "Why is the sky blue?" the model provides a concise explanation such as "air molecules scatter the Sun's shorter wavelength blue light more than its longer-wavelength red light." It does not enumerate every possible

[1]Booth School of Business, University of Chicago, Chicago, Illinois, USA
[2]Data Science Institute, University of Chicago, Chicago, Illinois, USA
[3]Computational Social Science Program, University of Chicago, Chicago, Illinois, USA
[4]Kellogg School of Management, Northwestern University, Evanston, Illinois, USA

[†]The authors contributed equally

Corresponding Author:
Joshua Conrad Jackson, Booth School of Business, University of Chicago, Chicago, Illinois, USA.
Email: joshua.jackson@chicago.booth.edu

scientific detail, nor does it list the full range of mythological or historical answers that humans have given. The same principle applies to social media feed-ranking algorithms, which must sift through millions of possible posts and comments and then present just a handful on a user's screen, reducing overwhelming informational complexity into a simple, digestible feed (Gillespie, 2018).

In providing a single or limited set of answers, AI systems engage in a process of simplification that necessarily involves *prioritization*. Models must decide which information is most relevant, which explanations to highlight, and which perspectives to omit (Vaswani et al., 2017). Although this prioritization process is core to the function of many commercial models, it can have negative consequences. First, when models prioritize, they can inadvertently prioritize compatible but inaccurate content. Second, when models prioritize prevalent information over rare information, they can emphasize the voices of groups that dominate their training materials, and homogenize cultural diversity.

These are two flavors of the "prioritization problem": defined here as the tension between trying to simplify complexity and trying to maximize accuracy and heterogeneity. The following sections describe these two manifestations of this problem in more depth, and discuss why they can undermine knowledge. The final section provides policy insights into how to mitigate these consequences and incentivize better large AI models. Figure 1 provides an overview of the paper.

Prioritizing Coherence and Engagement: Consequences for the Accuracy of Knowledge

AI prioritization becomes a problem for knowledge when the training objectives of AI models are misaligned with the criterion of factual truth. For instance, LLMs are trained primarily to speak fluently and convincingly, not to verify the correctness of the information they produce (Devlin et al., 2019; Radford et al., 2018). Similarly, recommender systems are optimized to maximize user engagement, not to ensure the epistemic reliability of the content they promote (Brady et al., 2023). Judged against their designed goals, such systems perform remarkably well. But when evaluated against the standard of factual accuracy, their outputs often fall short, because truth was never their direct optimization target.

Prioritizing Coherence Leads to LLM Hallucination

In LLMs, emphasizing coherence results in hallucinations—fluent but factually inaccurate statements (Ji et al., 2023). Hallucination arises throughout model training, probably because training and evaluation procedures reward guessing over acknowledging uncertainty (Kalai et al., 2025). LLMs are therefore incentivized to produce fluent and confident statements that are wrong over producing uncertain statements (Gibney, 2025; Huang et al., 2025; Kalai et al., 2025). Hallucination poses a range of applied problems. For example, in the case of AI companions designed to provide emotional support, fabricated content could mislead vulnerable users who rely on such systems for guidance, thereby undermining trust, exacerbating psychological distress, or prompting harmful decisions (Smith et al., 2025).

Hallucination is an especially common problem in domains where knowledge is rapidly evolving (e.g., recent political events), copyright restricted (e.g., new novels or paywalled research), or infrequently discussed (e.g., rare diseases) (Huang et al., 2025). In contrast, LLMs tend to hallucinate rarely when they describe well-documented events like the 9/11 terrorist attacks, which makes them good tools for breaking people's conspiracy theories involving these events (Costello et al., 2024). Hallucinations can also arise because large AI models do not reason. This lack of reasoning can lead LLMs to gravitate towards the most probabilistic reproductions in their training data rather than weighing evidence or reasoning toward a justified conclusion (Loru et al., 2025; Mitchell & Krakauer, 2023).

Prioritizing Engagement Leads to Misinformation on Social Media

Newsfeed-ranking algorithms can make similarly problematic prioritization decisions because of the mismatch between their training objectives and accuracy. The ultimate goal of these algorithms is to keep people engaging with the social media site so that they can maximize advertising revenue (Brady et al., 2020). To do this, newsfeed-ranking algorithms are trained to prioritize high-engagement posts, where engagement is measured through likes, shares, and time (i.e., spending time reading) (Brady et al., 2023). These high-engagement posts disproportionately feature three content themes: intergroup conflict, morality, and emotion, what are jointly called "IME" content (Brady et al., 2023). For example, a recent post on X by Marjorie Taylor Greene described COVID-19 vaccines as "deadly," claiming that "The American people deserve to know the truth." This post not only appealed to the vaccine skepticism in Greene's political in-group; it also framed vaccines as a moral peril, making it moralistic and emotional (Brady et al., 2017; McLoughlin et al., 2024).

By disproportionately promoting posts that have IME content, engagement-based algorithms may foster inaccurate perceptions, leading people to think IME content is more prevalent than it actually is. A recent study found evidence for this idea by tracking how one generation of participants engaged with social media posts through liking, sharing, and dwelling, and then training a simple engagement-based algorithm to sort a second generation's newsfeed based on engagement (Brady et al., 2025). The engagement-based feed-ranking algorithm increased in-group praise and out-

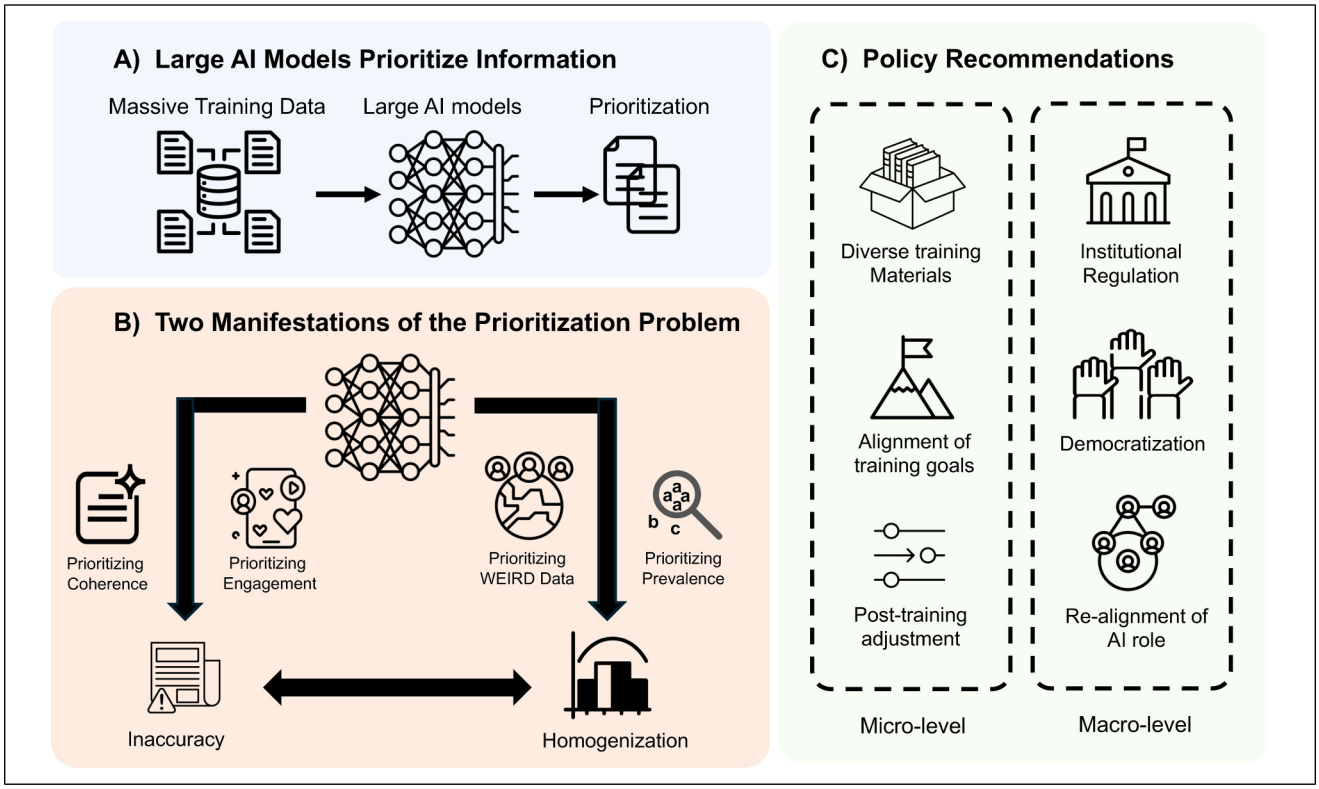

Figure 1. A Schematic of the Prioritization Problem and its Policy Solutions. A) Large AI models transform information by prioritizing certain information. B) Prioritization creates two problems—accuracy and heterogeneity—arising from misaligned training goals. C) Prioritization problems can be addressed via micro-level (e.g., improving the diversity of training materials, realigning training objectives, post-training adjustments) and macro-level reforms (e.g., institutional regulation, democratization, and re-aligning the role of AI).

group blame content by over 400% (from under 10% to about 40% of the feeds) from the first generation to the second generation when it was trained on users' likes and shares, even when it had no explicit intent to promote divisive content (Brady et al., 2025). A naïve user who infers social norms on the platform based on the content in the second generation will think that platform users are much more likely to engage in out-group blame than they really are.

By fostering misperceptions of IME discourse norms, newsfeed-ranking algorithms could also undermine adaptive norm change. Humans update their behavior through conformist transmission, copying what they perceive to be the prevailing norm (Henrich & Boyd, 1998). In many cases, this is adaptive. For example, driving on the same side of the road as everyone else will help someone avoid traffic accidents. But when norm perceptions are inaccurate, conformist transmission could undermine adaptation. In Brady et al. (2025), people who read the generation 2 newsfeed believed that dialogue was more moralistic, emotional, and divisive than it really was, leading them to misperceive the acceptability of expressing moral outrage (Brady et al., 2021). This could explain why moralistic language is rising on social media faster than it is rising in other online spaces (e.g., in online news) (Puryear, 2025).

Prioritizing engagement can also directly contribute to the spread of misinformation. When people post misinformation that evokes moral outrage, these posts drive greater engagement, and misinformation is subsequently promoted by feed-ranking algorithms, giving it more attention across the population of the platform (Brady et al., 2023). In a recent study of over 1 million Facebook links and 44,000 tweets, misinformation was more likely to spread than trustworthy information when it evoked moral outrage (McLoughlin et al., 2024).

Prioritizing Prevalence: Consequences for the Heterogeneity of Knowledge

Prioritization can risk compromising knowledge even when AI models are trained with no objective to promote engagement or fluency. Machine learning models are trained on loss functions that reward reproducing the most statistically probable outputs and penalize deviations from dominant patterns in the training data. By definition, this process creates a narrowing of content, as the most likely patterns are prioritized over less likely patterns. This narrowing has immediate problems for any kind of heterogeneity, including cultural diversity, and may have downstream consequences for innovation and problem-solving (Matz et al., 2025).

Prioritizing Prevalence Leads LLMs to Emphasize WEIRD Voices

LLM responses are a good example of this homogenization problem. Trained on loss functions that reward accurate token prediction (Brown et al., 2020; OpenAI et al., 2024), these models systematically favor common continuations: "have a nice" is almost always completed with "day," not "night." While this produces fluent and reliable text, it also means outputs are less heterogeneous than those of humans. Expressions that are majoritarian in human societies (e.g., 60% of people say A; 40% say B) can become unanimous in large language models. This is one barrier to replacing human subjects with LLMs in psychological studies (Dillion et al., 2023; Gerosa et al., 2024). Although LLMs can produce point estimates that resemble humans, and they underrepresent the heterogeneity of human opinions, beliefs, and attitudes.

Prioritizing prevalence does not just have implications for overall heterogeneity, but also has implications for whose voices are lost when LLMs distill information. Approximately 18% of the world's population lives in a country where English is an official language, and 5% of the population speaks English as a native language (Lewis, 2009). Yet 46% of the common crawl is English ("Common Crawl," 2025). More Wikipedia articles appear in English (~2.6 million) as in the next four languages combined (German, French, Polish, and Japanese) ("Common Crawl," 2025; Wikipedia, 2024). Since LLMs are mostly trained on online corpora, their training data is disproportionately English. Even the non-English training data mostly comes from other wealthy Western countries ("Wikipedia," 2024).

Because of biases in training data, content from non-Western non-English cultures will thus be disproportionately lost in LLMs that seek to prioritize prevalence. GPT's values most closely resembled those of populations in Western countries (Australia, Canada, and Germany), and far less those in countries (Libya, Pakistan, or Jordan) (Atari et al., 2023). Moreover, GPT displayed a Western analytic style of reasoning, favoring functional classifications over relational ones (Nisbett et al., 2001). As AI systems expand into education, religion, and therapy (Jackson et al., 2023a; Jackson et al., 2023b; Yam et al., 2023), such biases may not only homogenize cultural perspectives but also constrain how people conceptualize the world and themselves (Brinkmann et al., 2023; Choudhury, 2023; Glickman & Sharot, 2025). The rise of generative AI poses a real risk of making the world's psychology even more WEIRD (Western, Educated, Industrialized, Rich, and Democratic).

Prioritizing Prevalence Prevents New Ideas from Spreading on Social Media

This problem of homogenization also applies to recommender systems, though in a different form. Rather than predicting words within a context, these algorithms minimize loss functions tied to metrics such as clicks, likes, or dwell time. The result is that posts or videos that are already favored by a user are systematically elevated, and other content is sidelined—constituting a homogenization process (Brady et al., 2023, 2025; Glickman & Sharot, 2025).

A closely linked idea in the computer science of recommendation algorithms is the "cold start" problem, in which new content has little chance of being recommended because it does not have historical interaction data (Lika et al., 2014). This dynamic helps explain phenomena such as filter bubbles and echo chambers, where users are repeatedly exposed to similar content and gradually lose sight of the broader range of perspectives (Berman & Katona, 2020; Cinelli et al., 2021). For instance, on platforms such as Facebook and Twitter, algorithmically curated feeds foster strong homophilic clusters, so that information circulates mainly among like-minded users, amplifying polarization compared to platforms such as Reddit (Cinelli et al., 2021).

Cultural Homogenization Has Broader Implications

This loss of heterogeneity may also poses a knowledge problem (Messeri & Crockett, 2024). Multiple lines of research show that heterogeneity helps people solve complex problems and innovate on suboptimal solutions. Psychologists have long recognized that "groupthink" can impair decision-making because it reduces creativity and innovation (Janis, 1972). In sociology and cultural evolution, similar perspectives have conceptualized knowledge as a hill-climbing task where heterogeneity can help groups reach the global optimum (Lazer & Friedman, 2007; Smaldino et al., 2024). In simulations and experiments, when a group is able to break into smaller clusters that work independently to solve problems before pooling insights, they are better able to find the tallest hill in the landscape (Centola, 2022; Derex & Boyd, 2016). But when groups pool their insights from the start, they often get stuck on local optima and fail to find the best solutions (Smaldino et al., 2024).

However, in a world where large AI models are designed to expose people to a culturally homogenous system of knowledge, they systematically privilege majority patterns while filtering out minority voices, thereby reproducing a form of algorithmic monoculture (Kleinberg & Raghavan, 2021), discouraging cross-cultural innovation. Writing with LLMs reduces the diversity of generated content (Padmakumar & He, 2024), and scholars warn that such "machine culture" could erode cultural heterogeneity at scale (Brinkmann et al., 2023). Subsequently, this lack of heterogeneity risks diminishing functional diversity, inducing premature convergence in problem-solving, and stifling the very heterogeneity that underpins creativity and innovation (Burton et al., 2024). In this sense, the prioritization of Western-derived information in LLM outputs constitutes a knowledge problem that constrains the diversity on which effective problem-solving depends.

Fixing the Prioritization Problem: How to Incentivize Accuracy and Heterogeneity

How can we harness the benefits of large AI models for knowledge management while minimizing the costs associated with prioritization? At first glance, readers may perceive the prioritization problem as intractable: an inherent limitation of large AI models. This interpretation, however, may be overly pessimistic. The challenges described in this article are rooted in the current training regimes of large AI models. These regimes are not set in stone, but rather the consequence of the incentive structures and governance arrangements under which these systems have developed (Koch & Peterson, 2024).

For decades, the AI industry has been rewarded disproportionately for benchmark performance and user engagement, while accuracy, representativeness, and broader social welfare have been systematically undervalued (Dattakumar & Jagadeesh, 2003; Koch & Peterson, 2024). This misalignment of incentives has produced algorithms highly effective at optimizing narrow objectives, such as generating fluent text, maximizing clicks, or excelling on standardized benchmarks, yet poorly suited to ensuring reliability, preserving epistemic diversity, or fostering long-term innovation. Therefore, this misalignment is better understood as an institutional problem than a fundamental roadblock. Accordingly, changes in incentives, training regimes, and governance could have positive effects on the quality of AI-generated information.

Model-Level Policy Interventions

At the model level, one important avenue is to intervene on training materials. Current LLMs are primarily trained on data that overrepresent Western and English-dominant sources, a bias that narrows the epistemic base from the start (Brinkmann et al., 2023). Expanding training corpora to include underrepresented languages, cultural traditions, and minority perspectives would broaden the knowledge landscape for the models (Choudhury, 2023). Another strategy is to deliberately construct culturally specific sub-models or fine-tuned variants, each trained on distinct corpora (Li et al., 2024; Liu et al., 2025a; Xue et al., 2021). These sub-models could be offered to users either separately or in blended form, giving them the option to engage with multiple epistemic traditions rather than receiving a homogenized "average" output. Such diversification strategies would help counteract the monocultural tendencies that arise when models rely too heavily on a single global dataset.

A second set of interventions can take place at the level of optimization objectives. Current training procedures overwhelmingly reward models for producing the most probable continuation, which by design privileges majority patterns and sidelines minority ones. Adjusting the objective functions to incorporate diversity-sensitive penalties and rewards could help mitigate this tendency. For example, loss functions could be modified so that models are not only judged on accuracy in predicting common continuations but also rewarded for preserving variation or acknowledging their uncertainty (Kalai et al., 2025; Xu et al., 2018).

Beyond diversity and accuracy, optimization could involve reasoning (Binz et al., 2025). Recent work such as DeepSeek-R1 demonstrates that, by using reinforcement learning to introduce new reward signals favoring consistency, correctness, or strategic adaptation, models can acquire reasoning abilities that go beyond reproducing the frequency patterns of their training data (Guo et al., 2025). Interventions of optimization objectives could be equally useful for recommendation algorithms, which ensure that users are exposed to a richer range of perspectives without sacrificing usability.

The call to improve optimization objectives has been a focus in computer science, with the goal of maximizing fairness, trustworthiness, and responsibility in AI models. This entails developing models that not only maintain high accuracy but also uphold ethical and equitable standards. Studies in this vein have focused on three kinds of intervention: (1) Causal learning: designing loss functions that encourage models to identify causal relationships over mere statistical co-occurrences, thereby grounding predictions in causal reasoning (Liu et al., 2025b; Wang et al., 2021); (2) Fairness and robustness metrics: extending evaluation criteria beyond traditional accuracy measures to include metrics that capture model fairness and robustness (Agarwal et al., 2018; Gallegos et al., 2024); (3) Interpretable and transparent modeling: develop AI models that provide transparent and comprehensive explanations for their internal processes and outputs, promoting trustworthiness and responsible decision-making (Fujiwara et al., 2024).

Beyond training, post-training corrections can act as a further safeguard. Research on engagement-based algorithms suggests that small adjustments, such as reducing the influence of "super-posters" or deprioritizing toxic content at the end of current engagement-based algorithms, can improve information quality at low cost (Brady et al., 2023). Similarly, LLMs could integrate corrective layers after training. For example, filters could be designed to down-weight repetitive, homogenized outputs, or to ensure that generated responses reflect multiple cultural frames rather than converging on a single dominant narrative (Welbl et al., 2021). Recent evidence also suggests that even lightweight interventions such as prompt engineering can enhance the cultural and contextual sensitivity of LLM outputs, thereby increasing diversity without modifying model parameters (Lu et al., 2025). Together, these examples illustrate that technical fixes at the model level are not only possible but potentially highly effective.

Institutional Policy Interventions

Technical fixes at the model level can mitigate some challenges, but without macro-level governance reforms, such

improvements are unlikely to scale or endure. History shows that major technological revolutions, from the printing press to electricity to the internet, have never automatically translated into broad social benefits. Each required complementary institutional arrangements to ensure that private innovation served collective goals (Molho et al., 2024). AI is no exception (Farrell, 2025). Left solely to market forces, current incentive structures reward benchmark performance and user engagement while systematically undervaluing accuracy, representativeness, and broader social welfare, which can hardly be changed spontaneously without proper oversight or governance (Whittlestone et al., 2019). Therefore, deliberate policy and governance mechanisms, such as independent audits, transparency reports, embedded detection systems, and new laws (Birhane et al., 2022; Knott et al., 2023; Mökander et al., 2024) are needed to provide critical information about how models are trained, deployed, and used in practice, propelling the realignment of common goods with the long-term trajectory of AI.

A related problem is that the resources needed to develop and deploy large models, such as compute, data, and technical expertise, are increasingly concentrated in a handful of corporations and, in some cases, governments (Korinek & Vipra, 2025; Widder et al., 2023). This concentration heightens the risk of algorithmic monoculture and undermines democratic accountability (Chu et al., 2025; Kleinberg & Raghavan, 2021; Koch & Peterson, 2024; Shiiku et al., 2025; Summerfield et al., 2025). It also makes open models, transparency, participatory governance, and public investment in AI infrastructure essential for diversifying the epistemic base of AI and reducing dependence on a narrow set of powerful actors (Foffano et al., 2023; Gibney, 2022; Kapoor et al., 2024). Put simply, macro-level intervention is not optional but necessary to align AI development with pluralism, reliability, and social welfare.

A final strategy is to situate large AI models within a more appropriate social and epistemic niche. Current evidence does not suggest that large AI models can independently generate fundamentally new knowledge; rather, their comparative advantage lies in processing existing knowledge at unprecedented scale (Farrell, 2025; Farrell et al., 2025). For example, recent work shows that LLMs can take on multiple roles in collective cognition, including acting as participants in simulations, serving as interviewers to elicit responses, functioning as environments for interaction, facilitating consensus across groups, and analyzing large volumes of unstructured data into quantifiable formats (Sucholutsky, 2025). These functions highlight AI's strengths in organizing, synthesizing, and redistributing information, rather than in producing novel insights on their own.

Appropriately deployed, large models can enhance how humans manage, navigate, and recombine knowledge, thereby indirectly fostering innovation (Lazar et al., 2025; Shiiku et al., 2025). But this also means AI should not be treated as a substitute for human reasoning, moral judgment, or emotional support (Köbis et al., 2025; Montag et al., 2025; Wang et al., 2025). Instead, its proper role is as a human complement and collaborator: leveraging the speed and scale of AI while leaving the generation of new knowledge to human collectives. In short, societies should align AI's ecological niche with its strengths, using it to augment human knowledge production without displacing it.

Conclusion

Large AI models have an unprecedented ability to compress information into useful distillations. But this compression involves prioritization decisions, which, if poorly designed, can undermine the accuracy and heterogeneity of information. Here we have shown how the prioritization problem plays out in LLMs and Newsfeed-ranking algorithms on social media. Left unchecked, these problems could fuel misperceptions, bias, polarization, and stagnation of innovation. Although these consequences are serious, they are not insurmountable. We view them as largely engineering and governance problems. Through improved training materials, redesigned objectives, post-training safeguards, and supportive institutions such as open models, decentralized resources, and participatory governance, we can harness the power of large AI models to advance human welfare without undermining knowledge.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Joshua Conrad Jackson https://orcid.org/0000-0002-2947-9815

References

Open AI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., & Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*, *80*, 60–69. https://doi.org/10.48550/arXiv.1803.02453

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* https://osf.io/preprints/psyarxiv/5b26t/

Berman, R., & Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, *39*(2), 296–316. https://doi.org/10.1287/mksc.2019.1208

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., & Éltető, N. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. https://doi.org/10.1145/3551624.3555290

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Jackson, J. C., Doyle, M., & Baier, S. (2025). *Engagement-based algorithms disrupt human social norm learning*. https://osf.io/mgdwq/download

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947–960. https://doi.org/10.1016/j.tics.2023.06.008

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., & Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868. https://doi.org/10.1038/s41562-023-01742-2

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A.-M., Yasseri, T., Nickl, P., Almaatouq, A., & Hertwig, R. (2024). How large language models can reshape collective intelligence. *Nature Human Behaviour*, *8*(9), 1643–1655. https://doi.org/10.1038/s41562-024-01959-9

Centola, D. (2022). The network science of collective intelligence. *Trends in Cognitive Sciences*, *26*(11), 923–941. https://doi.org/10.1016/j.tics.2022.08.009

Choudhury, M. (2023). Generative AI has a language problem. *Nature Human Behaviour*, *7*(11), 1802–1803. https://doi.org/10.1038/s41562-023-01716-4

Chu, C. Y. C., Chang, J.-J., & Lin, C.-C. (2025). Why does AI hinder democratization? *Proceedings of the National Academy of Sciences*, *122*(19), e2423266122. https://doi.org/10.1073/pnas.2423266122

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Common Crawl. (2025). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Crawl&oldid=1272405656

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Dattakumar, R., & Jagadeesh, R. (2003). A review of literature on benchmarking. *Benchmarking: An International Journal*, *10*(3), 176–209. https://doi.org/10.1108/14635770310477744

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987. https://doi.org/10.1073/pnas.1518798113

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Farrell, H. (2025). AI As governance. *Annual Review of Political Science*, *28*(1), 375–392. https://doi.org/10.1146/annurev-polisci-040723-013245

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. *Science*, *387*(6739), 1153–1156. https://doi.org/10.1126/science.adt9819

Foffano, F., Scantamburlo, T., & Cortés, A. (2023). Investing in AI for social good: An analysis of European national strategies. *AI & SOCIETY*, *38*(2), 479–500. https://doi.org/10.1007/s00146-022-01445-8

Fujiwara, K., Sasaki, M., Nakamura, A., & Watanabe, N. (2024). Measuring the interpretability and explainability of model decisions of five large language models. *Open Science Framework: Charlottesville, VA, USA*. https://osf.io/d4ntw/download.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gerosa, M., Trinkenreich, B., Steinmacher, I., & Sarma, A. (2024). Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering*, *31*(1), 13. https://doi.org/10.1007/s10515-023-00409-6

Gibney, E. (2022). Open-source language AI challenges big tech's models. *Nature*, *606*(7916), 850–851. https://doi.org/10.1038/d41586-022-01705-z

Gibney, E. (2025). Can researchers stop AI making up citations? *Nature*, *645*(8081), 569–570. https://doi.org/10.1038/d41586-025-02853-8

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., & Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*(8081), 633–638. https://doi.org/10.1038/s41586-025-09422-z

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*(4), 215–241. https://doi.org/10.1016/S1090-5138(98)00018-X

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Jackson, J. C., Yam, K. C., Tang, P. M., Liu, T., & Shariff, A. (2023a). Exposure to robot preachers undermines religious commitment. *Journal of Experimental Psychology: General*, *152*(12), 3344. https://doi.org/10.1037/xge0001443

Jackson, J. C., Yam, K. C., Tang, P. M., Sibley, C. G., & Waytz, A. (2023b). Exposure to automation explains religious declines. *Proceedings of the National Academy of Sciences*, *120*(34), e2304748120. https://doi.org/10.1073/pnas.2304748120

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* https://psycnet.apa.org/record/1975-29417-000?utm_source=livewiremarkets.com&utm_medium=referral

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why Language Models Hallucinate* (arXiv:2509.04664). arXiv. https://doi.org/10.48550/arXiv.2509.04664

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., & Narayanan, A. (2024). *On the Societal Impact of Open Foundation Models* (arXiv:2403.07918). arXiv. https://doi.org/10.48550/arXiv.2403.07918

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Eyers, D., Trotman, A., Teal, P. D., Biecek, P., Russell, S., & Bengio, Y. (2023). Generative AI models should include detection mechanisms as a condition for public release. *Ethics and Information Technology*, *25*(4), 55. https://doi.org/10.1007/s10676-023-09728-4

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, *646*(8083), 126–134. https://doi.org/10.1038/s41586-025-09505-x

Koch, B. J., & Peterson, D. (2024). *From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution* (arXiv:2404.06647). arXiv. https://doi.org/10.48550/arXiv.2404.06647

Korinek, A., & Vipra, J. (2025). Concentrating intelligence: Scaling and market structure in artificial intelligence*. *Economic Policy*, *40*(121), 225–256. https://doi.org/10.1093/epolic/eiae057

Lazar, M., Lifshitz, H., Ayoubi, C., & Emuna, H. (2025). Would archimedes shout "eureka" with algorithms? The hidden hand of algorithmic design in idea generation, the creation of ideation bubbles, and how experts can burst them. *Academy of Management Journal*, *68*(5), 881–906. https://doi.org/10.5465/amj.2023.1307

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, *52*(4), 667–694. https://doi.org/10.2189/asqu.52.4.667

Lewis, M. P. (2009). *Ethnologue: Languages of the world*. SIL international.

Li, C., Teney, D., Yang, L., Wen, Q., Xie, X., & Wang, J. (2024). Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, *37*, 65183–65216. https://doi.org/10.52202/079017-2082

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073. https://doi.org/10.1016/j.eswa.2013.09.005

Liu, S., Jin, Y., Li, C., Wong, D. F., Wen, Q., Sun, L., Chen, H., Xie, X., & Wang, J. (2025a). *CultureVLM: Characterizing and Improving Cultural Understanding of Vision-Language Models for over 100 Countries* (arXiv:2501.01282). arXiv. https://doi.org/10.48550/arXiv.2501.01282

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., & Yu, T. (2025b). Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL*, *2025*, 7668–7684. https://doi.org/10.18653/v1/2025.findings-naacl.427

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, *122*(42), e2518443122. https://doi.org/10.1073/pnas.2518443122

Lu, J. G., Song, L. L., & Zhang, L. D. (2025). Cultural tendencies in generative AI. *Nature Human Behaviour*, *9*(11), 2360–2369. https://doi.org/10.1038/s41562-025-02242-1

Matz, S. C., Horton, C. B., & Goethals, S. (2025). *The Basic B*** Effect: The Use of LLM-based Agents Reduces the Distinctiveness and Diversity of People's Choices* (arXiv:2509.02910). arXiv. https://doi.org/10.48550/arXiv.2509.02910

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991–996. https://doi.org/10.1126/science.adl2829

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Mitchell, M. (2020). *Artificial Intelligence*. Picador Paper.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, *4*(4), 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Molho, C., Peña, J., Singh, M., & Derex, M. (2024). Do institutions evolve like material technologies? *Current Opinion in Psychology*, *60*, 101913. https://doi.org/10.1016/j.copsyc.2024.101913

Montag, C., Spapé, M., & Becker, B. (2025). Can AI really help solve the loneliness epidemic? *Trends in Cognitive Sciences*, *29*(10), 869–871. https://doi.org/10.1016/j.tics.2025.08.002

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291. https://doi.org/10.1037/0033-295X.108.2.291

Padmakumar, V., & He, H. (2024). *Does Writing with Language Models Reduce Content Diversity?* (arXiv:2309.05196). arXiv. https://doi.org/10.48550/arXiv.2309.05196

Puryear, C. (2025). *Rising Moralization in Social Media Discourse*. https://repository.uncw.edu/items/ac618300-6698-4696-835a-52819dec3fc1.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025). *The dynamics of collective creativity in human-AI hybrid societies* (arXiv:2502.17962). arXiv. https://doi.org/10.48550/arXiv.2502.17962

Smaldino, P. E., Moser, C., Pérez Velilla, A., & Werling, M. (2024). Maintaining transient diversity is a general principle for improving collective problem solving. *Perspectives on Psychological Science*, *19*(2), 454–464. https://doi.org/10.1177/17456916231180100

Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, *20*(6), 1081–1099. https://doi.org/10.1177/17456916251351306

Sucholutsky, I. (2025). Using LLMs to advance the cognitive science of collectives. *Nature Computational Science*, *5*(9), 704–707. https://doi.org/10.1038/s43588-025-00848-z

Summerfield, C., Argyle, L. P., Bakker, M., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., & Hadfield, G. K. (2025). The impact of advanced AI systems on democracy. *Nature Human Behaviour*, 1–11. https://doi.org/10.1038/s41562-025-02309-z

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. u., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, https://doi.org/10.48550/arXiv.1706.03762

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, *7*(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

Wang, Z., Shu, K., & Culotta, A. (2021). *Enhancing Model Robustness and Fairness with Causality: A Regularization Approach* (arXiv:2110.00911). arXiv. https://doi.org/10.48550/arXiv.2110.00911

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P.-S. (2021). *Challenges in Detoxifying Language Models* (arXiv:2109.07445). arXiv. https://doi.org/10.48550/arXiv.2109.07445

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Widder, D. G., West, S., & Whittaker, M. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (SSRN Scholarly Paper 4543807). *Social Science Research Network*, 1–27. https://doi.org/10.2139/ssrn.4543807

Wikipedia:Multilingual statistics. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Multilingual_statistics&oldid=1261711959.

Xu, J., Ren, X., Lin, J., & Sun, X. (2018). Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3940–3949. https://doi.org/10.18653/v1/D18-1428

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer* (arXiv:2010.11934). arXiv. https://doi.org/10.48550/arXiv.2010.11934

Yam, K. C., Tan, T., Jackson, J. C., Shariff, A., & Gray, K. (2023). Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence. *Management and Organization Review*, *19*(5), 859–875. https://doi.org/10.1017/mor.2023.21