Commentary

Embedding Societal Values into Social Media Algorithms

Michael S. Bernstein, Angèle Christin, Jeffrey T. Hancock, Tatsunori Hashimoto, Chenyan Jia, Michelle Lam, Nicole Meister, Nathaniel Persily, Tiziano Piccardi, Martin Saveski, Jeanne L. Tsai, Johan Ugander, and Chunchen Xu

1 Introduction

Social media influences what we see and hear, what we believe, and how we act—but artificial intelligence (AI) influences social media. By changing our social environments, AIs change our social behavior: as per Winston Churchill, “We shape our buildings; thereafter, they shape us.” Across billions of people on platforms from Facebook to Twitter to YouTube to TikTok, AI decides what is at the top of our feeds (Backstrom 2016; Fischer 2020), who we might connect with (Guy, Ronen, and Wilcox 2009), and what should be moderated, labeled with a warning, or outright removed (Gillespie 2018). These AI models change the social environment around us by amplifying or removing misinformation and radicalizing content (Hassan et al. 2015), by highlighting or suppressing antisocial behavior such as harassment (Lees et al. 2022), and by upranking or downranking content that might harm well-being (Burke, Cheng, and Gant 2020). How do we understand and engineer this sociotechnical ouroboros (Mansoury et al. 2020)?

As the traditional critique goes, these challenges arise because social media AIs are optimized for engagement (Backstrom 2016; Narayanan 2023). But this is not the full story: to help manage undesirable outcomes of engagement-based algorithms, platforms have long augmented their algorithms[1] with nonengagement outcomes (Eckles 2021). For instance, to help defeat clickbait, platforms such as Facebook began surveying users for their opinions on specific posts, and then building models that could predict and downrank posts that people dislike, even if they are likely to click on them (Backstrom and Mosseri 2015). To ensure that all users receive feedback, platforms designed algorithms weighing the effect of user feedback on other users who might otherwise get few replies (Eckles, Kizilcec, and Bakshy 2016). To diminish the prevalence of content that violates community standards, such as misinformation and gore, platforms built algorithms and paid moderation teams to flag and remove this content. This battery of surveys, moderation, downranking, peer effect estimation, and other models are all now components of many platforms (Eckles 2021).

[1] In this commentary, we refer to “AI” and “algorithm” interchangeably to refer to machine learning procedures that learn to predict from large-scale data. We are primarily concerned with social media algorithms focused on ranking and recommendation, especially feed algorithms, but we note that social media AIs play many other roles as well, including content moderation, (de)monetization, misinformation tagging, political content tagging, and toxicity judgments.

Journal of Online Trust and Safety, September 2023, page 1 of 13

doi:10.54501/jots.v2i1.148

Yet despite these moves, the values at the center of social media algorithms remain overridingly focused on the individual user's experience, especially their feelings of personal agency, enjoyment, and stimulation. This makes sense on one hand, since users typically join social media seeking positive experiences. But, as platforms have learned, focusing too heavily on individual values and experiences can lead to societal harms such as silencing, marginalization, and threats to democracy (Munn 2020; Klug et al. 2021). Efforts to broaden algorithms' objectives are undeniably valuable, but they are often ad hoc and limited to specific concerns or failures. Moreover, they only indirectly address the actual valued interest: for example, surely reducing misinformation is only a small part of the broader value of promoting a thriving democracy.

Could we create the means to embed societal values directly into these algorithms? A systematic method of organizing social media algorithms around values that impact us as groups or societies could offer a powerful toolkit for balancing the tensions at the core of social media. What would it mean for a social media algorithm to be able to optimize for prodemocratic attitudes, stronger community social capital, well-being, or increased social mobility? Could doing so produce more robust and general versions of existing feed ranking interventions? But even if societal values could add dimension and depth to feed algorithms, how would we build them concretely—and how would we measure and understand their impact?

Achieving this goal means creating a translational science that connects social scientific understanding of these issues to engineering and machine learning. This translational science might ask: what do we really mean when we tell an AI to estimate whether a post is "prodemocratic," or whether a post "harms well-being"? Although these concepts can seem slippery, social scientists have collectively operationalized, measured, refined, and replicated these constructs over decades. We argue that many of these measurements are now detailed and precise enough to translate into AI models, dramatically expanding the potential range of the societal values that could be encoded in social media algorithms. We envision a large repository of vetted, replicated, and well-tested social science models that can be integrated by any platform.

If embedding societal values into social media AIs is now possible, it also raises major questions for policy. Which societal values get included, and how do we negotiate cultural and ethnic differences in these values within and across societies? How do we trade off societal values against individual user experience, and when? Who gets to decide which of a society's values to prioritize? What happens when prioritizing one value in an algorithm inadvertently undermines another value? For instance, Facebook's prior attempts to prioritize the value of promoting communication between friends (by "sparking conversations and meaningful interactions between people") ended up increasing the spread of misinformation (Oremus et al. 2021), undermining the value of providing accurate information. Moreover, their efforts to downrank posts that Facebook users deemed "bad for the world" were canceled because the change reduced how much people used Facebook (Roose, Isaac, and Frenkel 2020). How do we navigate these trade-offs and competing goals?

This commentary lays out our proposed approach to centering societal values in social media AIs. We detail how all algorithms already embed values, explain how to embed societal values into these algorithms, and consider how explicitly modeling societal values toys with "opening up Pandora's Feed."

2 Feed Algorithms Already Embed Values

Social media algorithms *already* embed values. We argue that these values are typically focused on the individual by prioritizing their personal agency, enjoyment, and stimulation—and that we ought to be more explicit and expansive in the values that we embed.

When we say that feed algorithms already embed values, it is because they explicitly or implicitly encode notions of what content or behaviors are "good." Values are broadly defined as beliefs about desirable end states; they are considered as deriving from basic human needs and manifest in different ways across cultures (Rokeach 1973; Schwartz 1992; Schwartz and Bilsky 1987). Values guide our goals and decision-making (Schwartz and Boehnke 2004). Prevailing values are supported by strong and positive feelings attached to them, and are considered representations of cultural truisms (Ajzen 2001; Maio and Olson 1998).

Values are embodied in cultural practices and artifacts in a society. News feed algorithms, as artifacts, reflect and promote the values of their developers (Seaver 2017), prioritizing certain cultural values over others. Due to algorithms' "world-making" power in shaping the media landscape and their ability to affect society along multiple dimensions (Bucher 2018), designers have no options that avoid incorporating societal interests into their social media algorithms (Zhu et al. 2018).

Although today's social media algorithms incorporate a variety of considerations, concerns related to personal agency, enjoyment, and stimulation predominate. One basic substrata of many of these algorithms are observable engagement signals such as likes (Narayanan 2023), reflecting individual users' reactions to the content. Platforms then train a series of AI models based on other signals such as content, relationships between accounts, type of post, and models trained on satisfaction surveys, with each model predicting a different outcome. For example, in the case of Twitter's open sourced algorithm, AI models synthesize these signals into predictions of the probability that (1) the user will reply to the post, (2) the post's original author will engage with that reply, (3) the user will engage with the author's profile page, (4) the user will retweet and share the content, (5) the user will give negative feedback such as "not interested in this post," and others. On Facebook, YouTube, and other platforms, some of these models predict not just directly observable outcomes but also manual annotations from surveys of panels of users, or predict short-term proxies of longer-term outcomes. A linear combination of weights, or an omnibus learned model, then combines these individual predictions into a single score per post, as in Facebook's "Meaningful Social Interactions": a weighted combination of Likes, Reactions, Reshares, and Comments.

Changes to these feed algorithms are tested against behavioral metrics. These principal metrics take longer to measure, capturing notions such as "how often does a user, of their own volition, choose to open the app and consume our content?" These ideas are translated into metrics such as Daily Active Users (DAU), Monthly Active Users (MAU), and sessions per day, and any changes to the feed algorithm get A/B tested against changes to these outcomes. While secondary metrics exist—and they can take a more societal view—the principal metrics again embed values of personal agency, enjoyment, and stimulation, focusing mainly on outcomes concerning the individual.

One complication is that this approach of optimizing for individual experience can misfire: individual users can be drawn to content that conflicts with their values. In the United States, for example, people value positive emotions over negative emotions, and typically post more positive than negative content (Sims et al. 2015). However, US users are more likely to be influenced by others' posts that contain highly arousing negative content like anger, disgust, and fear (Hsu et al. 2021), and are more likely to share that content with others, perhaps because those states violate their values and therefore hijack their attention (Bellovary, Young, and Goldenberg 2021; Brady et al. 2017). This may explain why politically polarizing content and fake news that contain high arousal negative affect are more likely to spread in the US (Vosoughi, Roy, and Aral 2018).

One other mechanism that platforms use to support societal values is content moderation (Gillespie 2018), which can lead to takedowns, deplatforming, labeling, or demotion of content. Platforms often publish their community guidelines or policies that prompt moderation actions, including harmful content such as incitement and hate speech. To the extent that we view content moderation as part of "the algorithm," its filtering and downranking effects could certainly be seen as a lever for societal values. However, content moderation tends to focus on individual violations serious enough to remove, leaving significant room for other approaches that reason over more ambiguous cases.

3 Embedding Societal Values into Algorithms

Can we lift our gaze from individual to societal values? If so, a platform could prioritize content that promotes political participation, fosters mutual respect, and promotes peaceful transition of power, and deprioritize content that promotes political polarization and intergroup hate. Perhaps a platform might emphasize mental health, which would deprioritize content that encourages unhealthy social comparison or ways to self-harm. Or the platform may aim to increase social connection and local community-building. Some platforms already do a subset of these, but only at small scales.

The complexities

To answer this question, we have to start with: what does it mean to encode a value into an algorithm? If we cannot formulate an objective to say what we mean, when handed off to an AI, we are damned to mean what we say.

Fundamentally, AI systems work to maximize some outcome. For example, at their best, engagement measures should capture user experiences—presumably positive ones—that predict continued use. And just as operationalizing engagement is a subtle and often fraught art that must carefully weigh varied priorities, operationalizing societal values can be equally or even more fraught.

Then, how do we operationalize a societal value? Some values are well-operationalized by prioritizing or deprioritizing word lists, or other relatively elementary techniques from natural language processing (e.g., measuring valence) (Medhat, Hassan, and Korashy 2014), but these are rare and easily derailed by the vagaries of actual language use. Other values focus on how the scarce resources of user attention are allocated, e.g., whether feed algorithms should be configured to provide equal (or some minimal guarantee of) reach to all users, or whether some types of discussions should receive more attention than others. In configuring such allocation rules, platforms must navigate critiques of throttling, shadowbanning, and definition: What does "equal" mean? Equal access to each of my followers? Equal access to an equally sized audience? Operationalizing values inevitably leaves a gap between the value itself and its operational surrogate, a gap that can be susceptible to reward hacking (Skalse et al. 2022). Thus, implementing a prioritization of some societal value must be done with care to avoid significant unintended consequences.

Also challenging is that some pairs of values (or rather, their operationalizations) are countervailing, or even in direct opposition, while other pairs of values are synergistic or aligned, meaning that positive progress in terms of one value translates to positive progress in another. Few sets of values exist along independent axes, meaning that platforms can feasibly make more or less independent choices along those axes. In choosing to prioritize some constellation of societal values, it is crucial to understand which values are aligned, countervailing, or independent.

The approach

We argue that it is now possible to encode a much wider range of societal values into algorithms by connecting social science research with AI. Fields such as sociology, political science, law, communication, public health, science and technology studies, and psychology have long developed constructs to operationalize, describe, and measure complex social phenomena. These constructs have been proven reliable through repeated study and testing. In doing so, social scientists often develop measurement scales or codebooks to promote inter-rater reliability and replicability. We observe that the precision in these codebooks and constructs is now sufficient to translate into an artificial intelligence model. For example, Jia et al. (2023) translate a measure of antidemocratic attitudes into a prompt for in-context learning by a large language model such as ChatGPT. Their work draws on Voelkel et al.'s antidemocratic attitude scale (Voelkel et al. 2023), where, for instance, one variable captures "support for partisan violence." The original survey questions turn into a prompt for a large language model:

> *Please rate the following message's support for partisan violence from 1 to 3. Support for partisan violence is defined as "willingness to use violent tactics against outpartisans." Examples of partisan violence include sending threatening and intimidating messages to the opponent party, harassing the opponent party on the Internet, using violence in advancing their political goals or winning more races in the next election.*
>
> *Your rating should consider whether the following factors exist in the following message:*
>
> *A: Show support for partisan violence*
> *B1: Partisan name-calling*
> *B2: Emotion or exaggeration*
>
> *Rate 1 if doesn't satisfy any of the factors*
> *Rate 2 if doesn't satisfy A, but satisfies B1 or B2*
> *Rate 3 if satisfies A, B1 and B2*
>
> *After your rating, please provide reasoning in the following format:*
> *Rating: ### Reason: (### is the separator)*

The recipe for this translation is:

1. Identify a social science construct that measures the societal value of interest (e.g., reducing partisan violence).
2. Translate the social science construct into an automated model (e.g., adapting the qualitative codebook or survey into a prompt to a large language model).
3. Test the accuracy of the AI model against validated human annotations.
4. Integrate the model into the social media algorithm.

In the case of antidemocratic attitudes, each of the eight variables (e.g., support for partisan violence, support for undemocratic practices, opposition to bipartisanship) that comprise the social science construct (Voelkel et al. 2023) were turned into a labeling policy with expert annotators to operationalize how to manually rate posts. Then, this labeling policy served as the basis for an AI model that used a large language model (GPT-4) to replicate the expert annotations at scale. In a series of preregistered online experiments with partisans in the United States, the feed variants that embedded democratic values significantly reduced partisan animosity relative to traditional engagement-based alternatives, and the model accurately replicated expert manual annotation on these variables (Jia et al. 2023). These results chart a course for encoding other societal values into real-world social media AIs: building on social science theory and findings, we can adapt these constructs for social media content, use them to redesign algorithmic ranking and feed design, and scale up these ranking decisions to translate theory to product to effect the desired outcome—here, lower partisan animosity.

4 Opening Up Pandora’s Feed

Today’s status quo systems already encode certain values, but, in general, they are implicitly or indirectly incorporated. Explicitly embedding values such as mitigating antidemocratic attitudes could help platforms better reason over the tensions created by their ranking functions. However, this may also introduce new complications in deciding which values should be emphasized, and when. Here we illustrate some of these challenges and identify guiding principles for resolving these difficult decisions.

A first major question that arises if we embed societal values into feed algorithms: who gets to decide which values are included? When there are differences, especially in multicultural, pluralistic societies, who gets to decide how they should be resolved? Should a single CEO, who holds a particular set of values, control the fabric of a platform’s discourse by deciding which societal values should be weighed more than others? One benefit of intentional, explicit levers for societal values is that they might allow civil society, researchers, platform users, and others to articulate which values ought to be included in every platform, and at what level. But who can adjudicate differing opinions?

Second, because values are often in conflict, embedding some societal values will inevitably undermine other values. On one hand, certain values might seem like unobjectionable table stakes for a system operating in a healthy democracy: e.g., reducing content harmful to democratic governance by inciting violence, reducing disinformation and affective polarization, and increasing content beneficial to democratic governance via promoting civil discourse. But each of these is already complex. TikTok wants to be “the last happy corner of the internet,” which can imply demoting political content, and Meta’s Threads platform has expressly stated that amplifying political content is not their goal. If encoding societal values is at the cost of engagement or user experience, is that a pro- or antidemocratic goal?

Likewise, reducing misinformation comes with both democratic benefits and costs. Would downranking common sources of misinformation ultimately be a democratic move or an antidemocratic move? American legal precedent has come down firmly that the costs to society of chilling speech are far worse than the costs to society of allowing most distasteful speech. An algorithm that demotes content that has a high likelihood of containing disinformation may also be demoting content that a subsection of the population trusts and desires.

The jury remains out even on the basic question of whether maximizing individual agency and experience, as typefied by engagement signals, empowers or harms society. The traditional Habermasian view is that democracy and society flourish when there are public spheres for debate (Habermas 1991). However, Fraser (1990) argues that marginalized groups do not have, in practice, equal access to these public spheres, and that they must form counterpublics to gather and raise their voice. Maximizing engagement might be exactly what you want to do if you want to facilitate the creation of counterpublics, because recommendation systems can connect like-minded individuals to form these communities (Florini 2014). But it also enables harassment of those groups, since it can make it easier for a motivated individual to gain visibility for their harassment.

Third, as is well known in alignment problems, algorithms designed with the best intention and with apparently harmless behaviors may have unintended consequences that are hard to predict in advance. For example, an effect known in psychology as "institutional inversion" suggests that building institutions based on popular values may end up precisely discouraging behaviors consistent with that value (e.g., historical anti-debt attitudes in Protestant places have led to contemporary households in Protestant cultures now carrying the highest debt loads) (Cohen, Shin, and Lawless 2021). Embedding positive values without careful scrutiny of their impacts may even lead to unintended adverse side effects. We argue that the effect of changes must be investigated holistically by developing key metrics—using surveys, behavioral traces, and models—to regularly monitor the impact on the final user and society.

5 Charting a Path Forward

These are serious challenges, but the response ought not to be to back off. On the contrary, we think it will be absolutely critical to find a strong and defensible approach for embedding a larger set of societal values into algorithms. So how do we go about it?

A first step is to make sure that we model the values that are in conflict with each other. We cannot manage values if we cannot articulate them—for example, what might be harmed if an algorithm aims to reduce partisan animosity? Psychologists have developed a theory and measurement of basic values (Schwartz 2012), which we can use to sample values that are both similar and different across cultures and ethnicities.

A second step is to develop mechanisms for resolving value conflicts. Currently, platforms assign weights to each component of their ranking model, then integrate these weights to make final decisions. Our approach fits neatly into this existing framework. However, there is headroom for improved technical mechanisms for eliciting and making these trade-offs. One step might be increased participation, as determined through a combination of democratic participation and a bill of rights. An additional step is to elicit tensions between these values and when each one ought to be prioritized over another: for example, under what conditions might speech that increases partisan animosity be downranked, and under what conditions might it be amplified instead? What combination of automated signals and procedural processes will decide this?

On the product side, platforms need to be concerned with the behavior of algorithms that embed societal values. What content is upranked compared to the current feed? What content is downranked? Which content creators benefit? Which are harmed? What impact does this change have on traditional metrics? What are the second-order behavioral impacts of this kind of change? Although we encourage platforms to take stances on which values to integrate, and to what extent, it is also possible to consider a middleware or end-user customization approach where the platforms expose levers, dials, or controls to end users. These controls might allow end users to manipulate the strength of each of these values in their feed's algorithm. For example, platforms (e.g., Bluesky) might offer a "prodemocratic feed" and allow users to opt in. This approach would provide the option, while leaving untouched key democratic interests in free speech and free choice. Here too we would have to assess what the societal impact of these personal choices is.

On the research side, what are the social implications of encoding values at the platform level versus allowing customization at the user level? What theories are able to be operationalized into algorithmic objectives? Social science has built many theories over the years—are there patterns in which are more or less amenable to this approach? Even many robust theories are not designed to operate across the wide range of content that appears on social media, and may need to be adjusted. On the engineering side, it remains to be seen what we can, and can't, optimize for in practice. Are there theories where the current generation of large language models are simply not up to the task? How do we make these inferences robust?

On the policy side: how should, or should not, the government be involved in regulating the values that are implicitly or explicitly encoded into social media AIs? Are these decisions considered protected speech by the platforms? In the United States context, would the First Amendment even allow the government to weigh in on the values in these algorithms? Is being transparent about the values that a platform decides to encode—or allows people to customize—a way forward to support auditing, or would it lead to further gaming of the algorithms by bad actors? Does a choice by a platform to prioritize its own view of democracy raise its own free speech concerns? Given the global reach of social media platforms, should we try to consider how US policies will interact with those of other governments across the world?

Ultimately, we contend that many of the societal challenges facing social media platforms today are failures of our own imagination—of a lack of viable alternatives—rather than irreconcilable differences in value. Our goal, and one that we think research is best positioned to contribute, is to expand this horizon of the imagination by articulating approaches to expand the set of values in our algorithms.

References

Ajzen, Icek. 2001. “Nature and operation of attitudes.” *Annual Review of Psychology* 52 (1): 27–58. https://doi.org/10.1146/annurev.psych.52.1.27.

Backstrom, Lars, and Adam Mosseri. 2015. “How News Feed Works.” Presented at F8, 2015. https://www.youtube.com/watch?v=8-Yhpz_SKiQ.

Backstrom, Lars. 2016. “Serving a billion personalized news feeds.” In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining,* 469–69. https://doi.org/10.1145/2835776.2835848.

Bellovary, Andrea K., Nathaniel A. Young, and Amit Goldenberg. 2021. “Left-and right-leaning news organizations use negative emotional content and elicit user engagement similarly.” *Affective Science* 2:391–96. https://doi.org/10.1007/s42761-021-00046-w.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. “Emotion shapes the diffusion of moralized content in social networks.” *Proceedings of the National Academy of Sciences* 114 (28): 7313–18. https://doi.org/10.1073/pnas.1618923114.

Bucher, Taina. 2018. *If...Then: Algorithmic Power and Politics.* Oxford University Press. https://doi.org/10.1177/1461444819832541.

Burke, Moira, Justin Cheng, and Bethany de Gant. 2020. “Social comparison and Facebook: Feedback, positivity, and opportunities for comparison.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–13. https://doi.org/10.1145/3313831.3376482.

Cohen, Dov, Faith Shin, and Robert M. Lawless. 2021. “Attitudes, behavior, and institutional inversion: The case of debt.” *Journal of Personality and Social Psychology* 120 (5): 1117. https://doi.org/10.1037/pspa0000265.

Eckles, Dean. 2021. “Algorithmic transparency and assessing effects of algorithmic ranking.” *Testimony before the Senate Subcommittee on Communications, Media, and Broadband,* https://doi.org/10.31235/osf.io/c8za6.

Eckles, Dean, René F Kizilcec, and Eytan Bakshy. 2016. “Estimating peer effects in networks with peer encouragement designs.” *Proceedings of the National Academy of Sciences* 113 (27): 7316–22. https://doi.org/10.1073/pnas.1511201113.

Fischer, Sara. 2020. “Inside TikTok’s killer algorithm.” *Axios,* https://www.axios.com/2020/09/10/inside-tiktoks-killer-algorithm.

Florini, Sarah. 2014. “Tweets, Tweeps, and Signifyin’ Communication and Cultural Performance on ‘Black Twitter’.” *Television & New Media* 15 (3): 223–37. https://doi.org/10.1177/1527476413480247.

Fraser, Nancy. 1990. “Rethinking the public sphere: A contribution to the critique of actually existing democracy.” *Social Text,* nos. 25/26, 56–80. https://doi.org/10.2307/466240.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press. https://doi.org/10.12987/9780300235029.

Guy, Ido, Inbal Ronen, and Eric Wilcox. 2009. “Do you know? Recommending people to invite into your social network.” In *Proceedings of the 14th International Conference on Intelligent User Interfaces,* 77–86. https://doi.org/10.1145/1502650.1502664.

Habermas, Jurgen. 1991. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society.* The MIT Press. ISBN: 978-0-262-58108-0. https://mitpress.mit.edu/9780262581080/the-structural-transformation-of-the-public-sphere.

Hassan, Naeemul, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. "The quest to automate fact-checking." In *Proceedings of the 2015 Computation+ Journalism Symposium.* October. https://www.researchgate.net/publication/301801279_The_Quest_to_Automate_Fact-Checking.

Hsu, Tiffany W., Yu Niiya, Mike Thelwall, Michael Ko, Brian Knutson, and Jeanne L. Tsai. 2021. "Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values." *Journal of Personality and Social Psychology* 121 (5): 969. https://doi.org/10.1037/pspa0000282.

Jia, Chenyan, Michelle S. Lam, Minh Chau Mai, Jeff Hancock, and Michael S. Bernstein. 2023. "Embedding Democratic Values into Social Media AIs via Societal Objective Functions." arXiv: 2307.13912 [cs.HC].

Klug, Daniel, Yiluo Qin, Morgan Evans, and Geoff Kaufman. 2021. "Trick and please. A mixed-method study on user assumptions about the TikTok algorithm." In *Proceedings of the 13th ACM Web Science Conference 2021,* 84–92. https://doi.org/10.1145/3447535.3462512.

Lees, Alyssa, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. "A new generation of perspective API: Efficient multilingual character-level transformers." In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 3197–207. https://doi.org/10.1145/3534678.3539147.

Maio, Gregory R., and James M. Olson. 1998. "Values as truisms: Evidence and implications." *Journal of Personality and Social Psychology* 74 (2): 294. https://doi.org/10.1037/0022-3514.74.2.294.

Mansoury, Masoud, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. "Feedback loop and bias amplification in recommender systems." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management,* 2145–48. https://doi.org/10.1145/3340531.3412152.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5 (4): 1093–113. https://doi.org/10.1016/j.asej.2014.04.011.

Munn, Luke. 2020. "Angry by design: toxic communication and technical architectures." *Humanities and Social Sciences Communications* 7 (1): 1–11. https://doi.org/10.1057/s41599-020-00550-7.

Narayanan, Arvind. 2023. "Understanding Social Media Recommendation Algorithms." *Knight First Amendment Institute,* https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

Oremus, Will, Chris Alcantara, Jeremy B. Merrill, and Artur Galocha. 2021. "How Facebook shapes your feed." *Washington Post,* https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works.

Rokeach, Milton. 1973. *The Nature of Human Values.* Free Press. https://doi.org/10.1086/226092.

Roose, Kevin, Mike Isaac, and Sheera Frenkel. 2020. "Facebook Struggles to Balance Civility and Growth." *New York Times,* https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

Schwartz, Shalom H. 1992. "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries." In *Advances in Experimental Social Psychology,* edited by Mark P. Zanna, 25:1–65. Elsevier. https://doi.org/10.1016/S0065-2601(08)60281-6.

———. 2012. "An overview of the Schwartz theory of basic values." *Online Readings in Psychology and Culture* 2 (1): 11. https://doi.org/10.9707/2307-0919.1116.

Schwartz, Shalom H., and Wolfgang Bilsky. 1987. "Toward a universal psychological structure of human values." *Journal of Personality and Social Psychology* 53 (3): 550. https://doi.org/10.1037/0022-3514.53.3.550.

Schwartz, Shalom H., and Klaus Boehnke. 2004. "Evaluating the structure of human values with confirmatory factor analysis." *Journal of Research in Personality* 38 (3): 230–55. https://doi.org/10.1016/S0092-6566(03)00069-2.

Seaver, Nick. 2017. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." *Big Data & Society* 4 (2): 2053951717738104. https://doi.org/10.1177/2053951717738104.

Sims, Tamara, Jeanne L Tsai, Da Jiang, Yaheng Wang, Helene H. Fung, and Xiulan Zhang. 2015. "Wanting to maximize the positive and minimize the negative: implications for mixed affective experience in American and Chinese contexts." *Journal of Personality and Social Psychology* 109 (2): 292. https://doi.org/10.1037/a0039276.

Skalse, Joar, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. "Defining and characterizing reward gaming." *Advances in Neural Information Processing Systems* 35:9460–71.

Voelkel, Jan G., Michael Stagnaro, James Chu, Sophia Pink, Joseph Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi Allen, et al. 2023. "Megastudy identifying effective interventions to strengthen Americans' democratic attitudes." https://www.strengtheningdemocracychallenge.org/paper.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359 (6380): 1146–51. https://doi.org/10.1126/science.aap9559.

Zhu, Haiyi, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. "Value-sensitive algorithm design: Method, case study, and lessons." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–23. https://doi.org/10.1145/3274463.

Authors

Authors are alphabetized.

Michael S. Bernstein is an Associate Professor of Computer Science at Stanford University.

(msb@cs.stanford.edu)

Angèle Christin is an Associate Professor of Communication at Stanford University.

(angelec@stanford.edu)

Jeffrey T. Hancock is a Professor of Communication at Stanford University.

(hancockj@stanford.edu)

Tatsunori Hashimoto is an Assistant Professor of Computer Science at Stanford University.

(thashim@stanford.edu)

Chenyan Jia is a postdoctoral scholar at the Stanford Cyber Policy Center at Stanford University, and an incoming Assistant Professor at Northeastern University.

(chenyanj@stanford.edu)

Michelle Lam is a PhD student in Computer Science at Stanford University.

(mlam4@stanford.edu)

Nicole Meister is a PhD student in Electrical Engineering at Stanford University.

(nmeist@stanford.edu)

Nathaniel Persily is a Professor of Law at Stanford University.

(npersily@law.stanford.edu)

Tiziano Piccardi is a postdoctoral scholar in Computer Science at Stanford University.

(piccardi@stanford.edu)

Martin Saveski is a postdoctoral scholar in Management Science & Engineering at Stanford University, and an incoming Assistant Professor at the University of Washington.

(msaveski@stanford.edu)

Jeanne L. Tsai is a Professor of Psychology at Stanford University.

(jltsai@stanford.edu)

Johan Ugander is an Associate Professor of Management Science & Engineering at Stanford University.

(jugander@stanford.edu)

Chunchen Xu is a postdoctoral scholar in Psychology at Stanford University.

(cxu66@stanford.edu)

Acknowledgements

Our research was supported by the Hoffman-Yee Research Grants at the Stanford Institute for Human-Centered Artificial Intelligence (HAI). Michael Bernstein discloses

an ongoing part-time consulting relationship with Larian Studios. Tatsunori Hashimoto discloses an ongoing part-time consulting relationship with Microsoft.

Keywords

social media algorithms, societal values