DOI:10.1145/3624715

Henry Farrell and Cosma Shalizi

Opinion

Bias, Skew, and Search Engines Are Sufficient to Explain Online Toxicity

Social media would still be a mess even without engagement algorithms.

IMAGERY FROM SHUTTERSTOCK.COM

U.S. POLITICAL DISCOURSE seems to have fissioned into discrete bubbles, each reflecting its own distorted image of the world. Many blame machine-learning algorithms that purportedly maximize "engagement"—serving up content that keeps YouTube or Facebook users watching videos or scrolling through their feeds—for radicalizing users or strengthening their partisanship. Sociologist Shoshana Zuboff[15] even argues that "surveillance capitalism" uses optimized algorithmic feedback for "automated behavioral modification" at scale, writing the "music" that users then "dance" to.

There is debate whether such algorithms in fact maximize engagement (their objective functions also typically contain other desiderata). More recent research[3] offers an alternative explanation, suggesting that people consume this content because they want it, independent of the algorithm. It is impossible to tell which is right, because we cannot readily distinguish the consequences of machine learning from users' preexisting proclivities. How much demand comes from algorithms that maximize on engagement or some other commercially valuable objective function, and how much would persist if people got information some other way?

Even if we cannot answer this question in any definitive way, we need to do the best we can. There are many possible interface technologies that can help organize vast distributed repositories of knowledge and culture like the Web. These include:

- Traditional systems of categorization (such as the Dewey Decimal System, or the original Yahoo!)
- Systems such as Wikipedia and Reddit, in which human volunteers collate, organize, present, and revise information, providing an information resource, and a means for searching it, and human-selected links to external sources.
- "Traditional" search algorithms like Google's original PageRank algorithm[7] that rank items in terms of relevance, estimated as a function of the text of the options and the query, the number and "quality" of inbound links, and so forth.
- Modern social media algorithms: machine-learning driven systems that rank content to maximize some observable notion of users' engagement with it or other profit-related measure, updating the ranking model depending on user responses to the options presented.
- Large language model-driven interfaces that generate outputs based on a set of statistical weights that loss-

ily summarize some larger corpus of text and associated data.

If some of these interfaces lead to the kinds of toxicity (most particularly, distorted or false beliefs) that plague online political discussion in the U.S. we really want to know it. For example, if Zuboff is right, our politics would be *much* better if we had not adopted the kinds of social media algorithms that she worries about, and might be dramatically improved if we reverted to earlier, simpler interfaces.

If social media algorithms are primarily to blame for fractured discourse, then curbing them might make the Internet safer for democracy. If people still find distorted information when "algorithmic rabbit holes"[3] are not there, then curbing such algorithms would have less benefit, and perhaps even none at all. Answering such questions involves comparing different interfaces with each other, to figure out which kinds of social and political consequences might be associated with each kind of interface.

A Thought Experiment: The Internet without Modern Algorithms

Without good data (and appropriate statistical tools: social networks can seem designed to impede causal inference), we will resort to a thought experiment. How would the Internet affect democracy if modern social media algorithms *were not* a key interface through which people find content? Specifically, what would have happened if machine learning had not been used, and we had remained in the Internet circa 2012?

A thought experiment like this uses a simple model to compare the likely outcomes associated with different interfaces. Such models have obvious limitations. They strip out most of the features of complex phenomena, focusing on some causal relationships rather than others. But they also force modelers to clarify their intuitions, and can have considerable explanatory benefits if they focus on the right causal relationships. Scholars of complexity such as Scott Page[8] advocate acquiring a rich portfolio of models, but urge that each individual model, to be useful, must be "simple enough that within it we can apply logic."

We want our thought experiment to be psychologically realistic. Whether we are in the real world, or our imaginary counterfactual, human beings will follow predictable psychological patterns. They will look for information that tells them what they want to believe, rather than discomfiting contradictory evidence. Hugo Mercier and Dan Sperber[5] argue that we reason less to understand the world than to find seemingly convincing justifications for what we already want to believe. Equally, we are far better at spotting the holes in others' ideas than our own. That implies we think better in groups than alone, *if* we listen to criticism, and group members who disagree *do* have to argue with each other. If everyone in a group agrees, it may spin its shared ideas into increasingly convoluted yarns no outsider will touch. Similarly, when people provide information, they will typically be less motivated by disinterested truth-telling than their wish to persuade others, and to have their influence and wisdom socially recognized.

If social media algorithms are primarily to blame for fractured discourse, then curbing them might make the Internet safer for democracy.

Understanding that allows us to construct our counterfactual world on three pillars.

First: an information resource, which people both learn from and easily add to. This is the Internet of "Web 2.0," in Tim O'Reilly's phrase—technologically unsophisticated people can produce their own content, via Facebook, Twitter/X, and other platforms.

Second: an interface through which people discover plausibly relevant information from the resource. This interface would not be personalized, as machine learning allows. Instead, like early search engines, it would draw on the underlying link structure of the information resource as a proxy for interestingness and quality.[7]

Third: while the information resource is shared, it is easy to ignore other users, and the mechanics of the resource neither enforce nor reward consensus. There is no constraint on people's ability to find congenial information, or to share it with others to gain social recognition.

This provides a minimal model of how the Internet would look if modern social media algorithms did not exist. If this model also predicts a world awash in misinformation, and if our model focuses on important causal relations, we can surmise such algorithms are likely not the root of our trouble (though they may worsen it). If, alternatively, people in that world end up better informed, perhaps things would be better if the ML revolution had not happened.

Search Engines and Skew Distributions

Cutting to the chase: our alternative world also fills up with misinformation. People who reason as Mercier and Sperber describe will not use a Web 2.0 Internet to find truly objective knowledge on controversial topics, but to look for **rationalizations**—"information created or selected to provide epistemic support for beliefs that agents want to hold for non-epistemic reasons.[14] These rationalizations need not be pure trash, and may contain genuine facts and logic. But they need not be *right*, if they are *plausible* to their consumers.

People who dislike the consensus worldview will turn to the Internet for rationalizations to help justify and support their (possibly true) belief "they are being taken advantage of." Search—and ye shall find. They may not care *what* they find, so long as it is minimally plausible, even if they retrospectively contrive stories about how their new views were the only possible satisfying ones.

What is crucial is that dissatisfied searchers will gradually converge on *shared* rationalizations. As search engines direct people toward links that other searchers have linked to, stochastic perturbations may get locked in.[9] People will be increasingly likely to find well-followed sources of rationalizations, through search or links. As people use Web 2.0 technologies to publish their own rationalizations,

they will link to other rationalizations that they find attractive. In both the Web 2.0 and reality, they may be encouraged to develop and spread these rationalizations by platform metrics. For example, Twitter "gamifies" people's desire for social recognition in "addictive" ways by providing quantified measures of the influence of tweets and individuals.[6] All this creates a self-reinforcing dynamic *without* engagement-maximizing machine learning, through which already-influential sources of rationalizations become more influential over time.

These very simple features can support a system of **cumulative advantage** or **preferential attachment**, in which some sources of rationalizations become far more prominent than others, *because* they were already more prominent, generating a right-skewed distribution of influence, as described long ago by Herbert Simon.[12] The abstract model can be readily applied as follows.

- Every morning, some dissatisfied person wakes up to seek rationalizations for their dissatisfaction.
- With probability ρ, this person joins an existing group, a community sharing compatible rationalizations.
  - The probability of joining an existing group of k members is proportional to the number of people in such groups ("preferential attachment").
- With probability $1-\rho$, this person finds no satisfying group and begins their own.

This process leads to a heavy-tailed distribution of group sizes. There are, say, $N_k(t)$ groups of size k after the t^{th} searcher, and these numbers will all grow with t, but we can hope the *distribution* stabilizes, so that $N_k(t) \to p_k t$. (More exactly, $N_k(t)/t \to p_k$.) With each new searcher, $N_k(t)$ can either increase by 1, stay the same, or decrease by 1. It increases by 1 if the new searcher joins a group of size $k-1$ (so the new size of that group is k); similarly $N_k(t)$ decreases by 1 if the new searcher joins a group of size k; otherwise $N_k(t)$ stays the same. So the expected change in $N_k(t)$ (given the current size distribution) is

$$\mathbb{E}\left[N_k(t+1)\right] - N_k(t) = \rho \frac{(k-1)N_{k-1}(t) - kN_k(t)}{t}.$$

Substituting in $N_k(t) = p_k t$ and solving gives, at equilibrium,

$$\frac{p_k}{p_{k-1}} = \frac{\rho(k-1)}{1+\rho k}.$$

Defining $\alpha = 1/\rho$,

$$p_k = \frac{k-1}{k+\alpha} p_{k-1}.$$

Recursing, and telescoping factors together with the gamma function, gives

$$p_k = \frac{\Gamma(k)\Gamma(\alpha+1)}{\Gamma(k+\alpha+1)} p_1,$$

so, to ensure $\sum_k p_k = 1$, we need

$$p_k = \alpha \frac{\Gamma(k)\Gamma(\alpha+1)}{\Gamma(k+\alpha+1)}.$$

Using asymptotics for the gamma function, we get that for large k,

$$p_k = O(k^{-\alpha-1}).$$

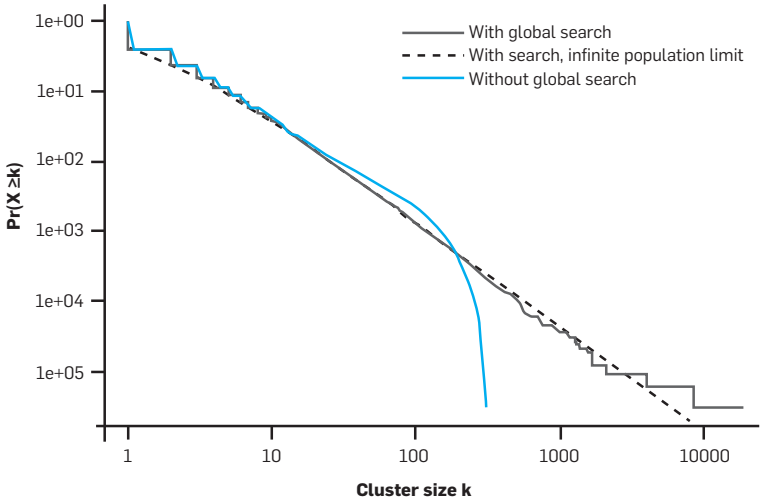
This distribution's right tail will be an approximate power law: a few groups absorb most searchers, surrounded by a vast sea of tiny groups. As the probability ρ of joining an existing group increases, the exponent α decreases, making the right tail heavier, increasing the skew, and increasing the share of searchers in the very largest groups.

Web 1.0, *with* search engines, makes this relevant. *Without* the Web, people would still be eager to rationalize their dissatisfactions, but those offering rationalizations would find it hard to broadcast them to searchers. Search engines connect searchers to rationalizations based on content, not spatial proximity. This makes it possible for searchers to find and join groups regardless of location, which is crucial to the asymptotics above. The Internet moved us toward a "mean field" regime, where dissent condenses into fewer, larger, and more consequential blobs of rationalization. This contrasts with research findings on pre-Web unconventional beliefs and identities (for example, Showalter[11]), which often emphasize how they disseminated *locally*, through geographically or professionally bounded social networks, leading to a world of inhibited preferential attachment where dissent is diffused through an immense number of very small groups or isolated individuals, pulling away from the mainstream in different directions and perhaps canceling each other out (see the accompanying figure).

Web 1.0, with search, made the asymptotics of preferential attachment relevant. Improving search engines increased ρ, by using better syntactic and topological cues to link searchers to congenial rationalizations. Search engines flatten the Web into lists that prioritize more popular results or those that seem more authoritative given the existing link structure. This is invisible to users, some of whom treat search engine prioritizations, nearly literally, as Gospel truth.[13] More subtly, Web 2.0, with its user-generated content that fed back in to search rankings, increased ρ still further, making it more likely that searchers will find, and join, large groups in proportion to their size.

In short, our thought experiment

With global search for clusters, a preferential attachment process ($\rho = 2/3$, $\alpha = 3/2$) in a population of 10^6 (black line) closely approximates the infinite-population limit (dashed). When the searchers are broken into 10^3 groups of 10^3, the cluster-size distribution cannot form a heavy tail (blue).



suggests that rationalizations and the communities around them will have a highly skewed distribution, a few very large, but surrounded by an immense number of miniscule groups. That does not seem much different, or better, than the high-misinformation Internet we inhabit.

Institutions before Algorithms

This does not prove search is responsible for the Internet's democratic problems, or that social media platforms' algorithms are irrelevant to online toxicity. Nor does it account for the interactions between Web-based information discovery and other media, for example, cable television.[2] But it does predict that even without machine learning, our online space would be populated by self-reinforcing communities of specious rationalizations. Under plausible assumptions, previously dominant interfaces (such as search) can explain people's separation into self-reinforcing bubbles.

Our thought experiment suggests that our current problems (or something very like them) were built into Web 2.0, even into Web 1.0, if not so glaring at first. There were many examples of deranged cognition (for example, communities of people convincing each other that they were the victims of mind control[1]) back then. We should not fall into the trap of thinking all this toxicity will go away if we can just rein in the engagement algorithms.

It provides a plausible but simple account of how individual psychological propensities may interact with specific interfaces, with large-scale collective consequences. There is an existing literature that highlights how people's desire for recognition and plausible rationalizations may drive them to seek out toxic material and behave in toxic ways,[6,14] but it does not explain the large-scale dynamics of how toxicity aggregates. Our model provides an account that connects individual desires to large scale outcomes.

That in turn points to other ways forward. Some are prior to search: Few people who are fairly satisfied with their lives will search for rationalizations explaining why everything is wrong. But our account also suggests that interfaces affect the ways in which people aggregate, and who they aggregate with. Put more simply, different interfaces will likely be associated with different group-level dynamics, and some group dynamics may be healthier than others. Left alone, people tend to seek out rationalizations for what they already think. This may be worsened by interfaces that actively or tacitly guide them toward such rationalizations. But other interfaces might *oblige them* to engage with those who they disagree with, so that they might have to respond to criticisms of their flabby arguments and specious assumptions.

This does not prove search is responsible for the Internet's democratic problems, or that social media platforms' algorithms are irrelevant to online toxicity.

This is not a purely theoretical argument. There is some empirical evidence that certain human-moderated platform interfaces mitigate online toxicity by reducing homophily and building disagreement in. Shi et al. find that Wikipedia articles are, on average, higher quality when written by people with sharp political disagreements.[10] Wikipedia's structures force them to engage with each other, so that their arguments are improved through mutual criticism. As Shi and co-authors describe it:

Editors ... said, "We have to admit that the position that was echoed at the end of the argument was much stronger and balanced." Did they begrudgingly come to that? They did, and that's the key.

Like democratic politics,[4] Wikipedia forces people with different perspectives to work together and reach acceptable if unhappy compromises. Like democratic politics, it has its ugly side. Yet it also provides relatively high-quality information. Not all of the Internet should be like Wikipedia, just as grudging consensus is not the only thing we want from politics. But if we want an Internet with less misinformation, one where polarization sharpens rather than corrodes our thinking, we can learn how to build better interfaces from Wikipedia. Its combination of evidentiary standards, and requirements that people argue out differences under these standards can be vexing and inefficient, but has also built one of the most robust structures of reasonably reliable information on the Internet. C

References

1. Bell, V., Maiden, C., Muñoz-Solomando, A., and Reddy, V. "Mind control" experience on the Internet: Implications for the psychiatric diagnosis of delusions. *Psychopathology 39* (2006), 87–91; 10.1159/000090598
2. Benkler, Y., Faris, R., and Roberts, H. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, Oxford, 2018; 10.1093/oso/9780190923624.001.0001
3. Chen, A.Y. et al. Subscriptions and external links help drive resentful users to alternative and extremist youtube videos. *Science Advances 9* (2023), eadd8080; 10.1126/sciadv.add808
4. Farrell, H. and Rohilla Shalizi, C. Pursuing cognitive democracy. In *From Voice to Influence: Understanding Citizenship in a Digital Age*. D. Allen and J.S. Light (Eds.). University of Chicago Press, Chicago, 211–231; http://bactra.org/weblog/917.html
5. Mercier, H. and Sperber, D. *The Enigma of Reason*. Harvard University Press, Cambridge, Massachusetts.
6. Thi Nguyen, C. How Twitter gamifies communication. In *Applied Epistemology*, J. Lackey (Ed.). Oxford University Press, Oxford, 2021, 410–436; https://philpapers.org/rec/NGUHTG
7. Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. *Technical Report 1999-66*. Stanford University InfoLab, 1999; http://ilpubs.stanford.edu:8090/422/ Previous number = SIDL-WP-1999-0120.
8. Page, S.E. *The Model Thinker: What You Need to Know to Make Data Work for You*. Basic Books, New York, (2018).
9. Salganik, M.J., Dodds, P.S., and Watts, D.J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science 311* (2006), 854–856; http://www.princeton.edu/~mjs3/musiclab.shtml
10. Shi, F., et al. The wisdom of polarized crowds. *Nature Human Behaviour 3* (2019), 329–336; 10.1038/s41562-019-0541-6
11. Showalter, E. *Hystories: Hysterical Epidemics and Modern Culture*. Columbia University Press, New York, 1997.
12. Simon, H.A. On a class of skew distribution functions. *Biometrika 42* (1955), 425–440; 10.2307/2333389
13. Bolla Tripodi, F. *The Propagandists' Playbook: How Conservative Elites Manipulate Search and Threaten Democracy*. Yale University Press, New Haven, CT, (2022).
14. Williams, D. The marketplace of rationalizations. *Economics and Philosophy 39* (2023), 99–123; 10.1017/S0266267121000389
15. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, 2019.

Cosma Shalizi (cshalizi@cmu.edu) is a professor of statistics at Carnegie Mellon University, Pittsburgh, PA, USA.

Henry Farrell (henry.farrell@gmail.com) is Stavros Niarchos Foundation Agora Institute Professor of International Affairs at Johns Hopkins University, Washington, D.C., USA.

The authors are grateful to two anonymous reviewers, to the editors of *Communications*, and to Carl Bergstrom, Michael Bernstein, and Nathan Mathias for comments on earlier versions of this Opinion column.