Safe Mode: Security Vocabulary and the Platforms That Execute Belief

Flyxion
Independent Researcher

Abstract

Contemporary technological culture repeatedly confuses the visible performance of control with control itself. A computer is judged fast because it boots quickly, secure because numerous services have been disabled, and private because an "AI" switch has been turned off. Information is judged credible because it arrives through familiar people, attracts a large audience, invokes technical language, or resembles a responsible warning. Social platforms intensify this confusion by treating engagement as a proxy for value and by allowing inexpensive claims to impose expensive verification work on their recipients. This essay argues that the comparison between computer malware and so-called "mind viruses" is more than metaphorical: both exploit trusted channels, disguise their operative purpose, acquire permissions from users, reproduce through existing infrastructure, and resist removal. Informational security nonetheless differs from computer security in one decisive respect, namely that the human recipient participates in interpretation and transmission rather than merely executing instructions. The appropriate response is therefore not a universal ideological firewall but a discipline of provenance, friction, bounded trust, corrigibility, and a defensible right of refusal.

1 When the Interface Says "Safe"

Consumer security guidance of the kind circulated for older operating systems provides a convenient starting point, not because any particular guide deserves scrutiny, but because its structure recurs across decades of consumer computing advice. Such guides typically recommend disabling unused services, removing unnecessary accounts, restricting network protocols, editing configuration files directly, exposing hidden file extensions, and modifying local name-resolution files to block known malicious hosts. Considered individually, several of these steps genuinely reduce an attack surface: a service that is not running cannot be exploited through a vulnerability in that service, and an account that does not exist cannot be compromised [1].

The difficulty lies not in the individual steps but in the inferential leap that customarily follows them. A list of completed hardening actions is quietly treated as proof of a secured system, even though no threat model has been specified, no adversarial testing has been performed, and no distinction has been drawn between exposures that were closed and exposures that were never assessed. Undetected compromise is thereby treated as equivalent to the absence of compromise, which is not a security claim at all but an admission of ignorance dressed as an achievement [2].

This produces a substitution that recurs throughout the domains examined in this essay:

$$\text{security-associated operations} \not\Rightarrow \text{verified security}.$$

A procedure can employ authentic technical vocabulary, and even manipulate the correct object, while still misunderstanding the mechanism it claims to control. One recurring instance of this

pattern is the conflation of IP protocol numbers with port numbers in firewall configuration instructions. A procedure of this kind enters values such as 6 and 53 under a field labeled "IP Protocols," apparently treating them as service ports. Yet IP protocol 6 denotes TCP itself, not any single service running over TCP, and IP protocol 53 does not denote DNS port 53 at all. The resulting configuration therefore permits a far broader class of traffic than its numerical precision suggests, even though the interface, the vocabulary, and the underlying values are all technically authentic. The error concerns the relation between the vocabulary and the mechanism, not the presence or absence of either one. The lesson generalizes well beyond operating systems: vocabulary that names a desirable property is not evidence that the property has been produced.

2 Performance by Conspicuous Proxy

The same structural error appears in popular treatments of computer performance. Boot time, shutdown time, and application-launch latency are legitimate and measurable quantities, but they capture only a narrow slice of what "performance" could mean. A system that reaches its desktop in eighteen seconds may nonetheless require days or weeks to transcribe a collection of lectures that a modern GPU processes in minutes. The disagreement in such cases is not really about which computer is faster; it concerns which workloads have been admitted into the definition of speed in the first place.

It is useful to separate at least three distinct quantities that popular usage collapses into one:

$$T_{\text{interactive}} \neq T_{\text{compute}} \neq T_{\text{total workflow}}.$$

Interactive latency governs the subjective experience of using a machine: how quickly menus open, windows respond, and the desktop appears. Computational throughput governs how quickly a well-defined task, such as video encoding or model training, is completed once initiated. Total workflow time governs how long a person's actual goal takes from start to finish, including setup, waiting, error recovery, and any manual steps a slow interactive experience makes more tedious. A narrow benchmark drawn from the first category becomes misleading exactly when it is promoted, whether by a reviewer, an advertisement, or a habit of speech, into a universal ranking that implicitly claims to speak to the second and third. This is proxy capture: an easily observed and cheaply produced indicator displaces the harder, more expensive, and more relevant property it was originally meant to represent.

3 Safety as a Vocabulary Effect

These two examples motivate a broader concept: safety vocabulary. Terms such as secure, trusted, protected, private, encrypted, restricted, intelligent, verified, and safe communicate desirable states of affairs, but their mere presence in a sentence, a settings menu, or a marketing page does not demonstrate that the state they name has actually obtained. A checkbox labeled "enhanced protection" changes what a user believes about their exposure regardless of whether it changes their exposure.

The same substitution operates in reverse. Removing antivirus software, disabling automatic updates, opting out of cloud services, blocking telemetry, or switching off an "AI" feature can each produce a strong subjective feeling of independence and control without necessarily producing greater security or privacy. A control can remove one risk while quietly reinstating another: disabling automatic updates reduces the risk of an unwanted change to a working configuration, but may leave known vulnerabilities uncorrected unless another reliable update practice replaces it.

The corrective is to insist that safety claims be decomposed into their constituent relations before they are accepted or rejected:

$\mathrm{Safe}(x)$ for whom, against what threat, during which interval, and under what assumptions?

A claim of safety that cannot answer these four questions is not yet an engineering claim. It is closer to marketing copy that has borrowed engineering's vocabulary without adopting engineering's obligations of specification and testing.

4 From Executable Malware to Executable Claims

The comparison between malicious software and misleading information is best drawn at the level of mechanism rather than at the level of any single author's psychology or intent. Much malware enters through trusted channels or exploits permissions already granted to trusted software: an email attachment from a known contact, a plausible software update, or a familiar-looking application. Not all of it works this way; worms and exploit chains can propagate through unpatched services without any meaningful user consent at all. But a large and socially significant class of malware does depend on trust and granted permission, and it is this class that the informational comparison tracks most closely. Such software obscures its actual behavior behind an innocuous name or interface. It requests, or silently inherits, permissions beyond what its stated purpose requires. It modifies the behavior of subsequent operations on the infected system. And it attempts to reproduce, whether by mailing itself to further contacts, embedding itself in shared files, or exploiting the same trust relationships that let it in.

A further distinction sharpens the analogy rather than weakening it. Malware is often classified partly by its unauthorized function: it does something the system's owner never agreed to. Misleading claims, by contrast, are usually transmitted with the recipient's full voluntary participation. The informational analogue of malicious execution is therefore not unauthorized execution but misinformed authorization. The claim succeeds by inducing the recipient to grant permission, in the form of belief, trust, or onward sharing, under a false description of what is actually being admitted. This can be written as

$$\mathrm{Admission}(m) = \mathrm{Permission}\big(m \mid \widehat{\mathrm{Function}}(m)\big),$$

where the recipient grants permission according to an estimate of the message's function, $\widehat{\mathrm{Function}}(m)$, that may differ sharply from the message's actual function. A conspiratorial post presents itself as a warning or as suppressed evidence while its actual function is to modify the recipient's trust relations, identity commitments, and future admissibility criteria. The permission is real; the description under which it was granted is not.

Certain informational objects display closely analogous properties. They arrive through friends, family members, or recognized public figures rather than through anonymous strangers. They disguise persuasion as warning, framing themselves as protective disclosure rather than argument. They activate fear, anger, or moral urgency more readily than they activate deliberation. They modify the recipient's subsequent trust relations, often by recasting mainstream institutions as untrustworthy and the message's source as uniquely reliable. And they encourage, sometimes explicitly, immediate onward sharing. Their success depends less on being true than on being memorable, identity-compatible, and easy to transmit intact.

The analogy should nevertheless be bounded rather than taken as a literal identity. A belief is not executable code, and a human recipient is neither deterministic nor passive in the way

a vulnerable program is. Informational infection is better modeled as a conditional interaction between message, recipient, and context than as a fixed payload with the same effect on every host:

$$\text{Effect}(m, u, c) = f(\text{message}, \text{recipient}, \text{context}).$$

The "virus," if the term is to be used at all, resides partly in the message's construction, partly in the psychological and social state of the receiving individual, and partly in the network incentives that surround and reward transmission. None of the three is sufficient on its own to explain why a given claim propagates.

5 Trusted Sources and Counterfeit Certificates

Computer systems have long distinguished the provenance of code from the behavior of code. A valid cryptographic signature establishes that a piece of software originated from a particular publisher; it does not, by itself, establish that the software is harmless, well written, or free of vulnerabilities [3]. Informational environments routinely collapse this distinction. "This came from someone I know" establishes social provenance and nothing more. "This person has five hundred thousand followers" establishes popularity and nothing more. "This person calls himself a technology entrepreneur" establishes a claimed role and nothing more. None of these facts, individually or jointly, establishes the truth of any particular technical assertion the person makes.

The relevant chain of inference can be written explicitly, precisely because ordinary discourse tends to skip its middle term:

$$\text{Identity}(s) \not\Rightarrow \text{Competence}(s, d) \not\Rightarrow \text{Truth}(p).$$

A source can be reliably identifiable without possessing competence in the relevant domain, and a source can possess genuine competence in one domain while asserting a falsehood about a different one, or even about the domain in which it is ordinarily competent. Social media interfaces, through their emphasis on follower counts, verification badges, and algorithmically amplified reach, compress these separate evaluations into a single undifferentiated impression of credibility, and that compression is precisely what allows the first link in the chain to stand in for all three.

6 Conspiracy Before the Algorithm

Any historical account of these dynamics should resist a simple technological determinism in which platforms are treated as the origin of the phenomena they merely accelerate. Conspiratorial reasoning, sectarian politics, propaganda, rumor, nationalism, and manufactured group identity all predate the contemporary internet by centuries. The platform did not invent the underlying human susceptibility to these patterns; it altered the cost and speed of their expression.

The early web reduced the marginal cost of publication to nearly zero, allowing anyone to produce content that had previously required a printing press, a broadcast license, or an institutional affiliation. Social media then layered onto this reduced cost a set of additional mechanisms: algorithmic recommendation, visible peer endorsement, automatic redistribution through shares and reposts, identity-based communities organized around shared belief, and continuous, granular measurement of audience response.

This infrastructure did not need to invent distrust from nothing. Modern history supplies recurring episodes in which genuine institutional secrecy was followed, sometimes years later, by confirmed evidence of government or corporate deception. Each such episode produces two things

at once: a legitimate, independently verifiable basis for skepticism toward the specific claim that was later shown to be false, and a much larger supply of general suspicion that can be, and often is, generalized far beyond the evidence that actually warranted it. Conspiratorial networks specialize in that generalization, extending a narrow and justified doubt about one claim into a categorical doubt about entire classes of institutions. The technological transformation, on this account, was not the creation of false belief from nothing, but the industrialization of its discovery, reinforcement, and transmission: a set of processes that previously operated at the speed of pamphlets and rumor now operate at the speed of a notification [6, 7].

7 Engagement Is Not Epistemic Value

Platform ranking systems require signals that can be measured cheaply and at scale. Clicks, comments, reactions, time spent viewing, and shares are all available in this form; truth, social value, and long-term understanding are not, and arguably cannot be, measured with comparable ease or speed. This mismatch produces a further instance of proxy capture. A schematic engagement score for a message m might be written as

$$E(m) = \alpha C + \beta R + \gamma S + \delta T,$$

a weighted combination of clicks, reactions, shares, and time spent. This is offered only as a schematic, not as a description of any actual platform's implementation: real ranking systems combine many more signals, nonlinear interactions, learned prediction models, and user-specific features that no simple linear expression could capture. What survives the simplification is the structural point that

$$E(m) \not\approx \text{Truth}(m),$$

and that the correlation between the two may be weak, absent, or even negative for certain categories of content [7].

Content that provokes correction can perform, on such a metric, just as successfully as content that earns agreement, since a ranking system built primarily on engagement need not distinguish persuasion from disgust when both extend the length of a session. Outrage is an especially efficient input to such a system because it simultaneously produces attention, emotional response, public identity display through sharing, and further distribution, all from a single stimulus [12, 10]. Real systems can and do contain integrity classifiers, downranking rules, and explicit safety objectives alongside engagement optimization, so the claim cannot be that engagement optimization always wins; the actual behavior of a feed algorithm is a complex, empirically contingent matter rather than a simple function of engagement alone [11]. The narrower and more defensible claim is this: unless accuracy and usefulness enter through independent objectives or constraints, maximizing $E(m)$ supplies no intrinsic reason to disfavor content that achieves engagement through outrage rather than through accuracy or usefulness.

8 Manufactured Opposition

The general pattern of institutionally manufactured group identity, familiar from school spirit campaigns, color-coded team competitions, and overt political polarization, offers a useful model for how arbitrary distinctions generate authentic attachment. Institutions can divide participants into groups on essentially arbitrary grounds, assign symbols and colors to each group, keep score, and generate sincere identification with remarkable speed. The resulting emotion need not be

considered fake merely because the original partition was itself arbitrary; institutional structure can produce genuine attachment around a distinction that has no independent basis [5].

Competition of this kind contains a dual structure that recurs across very different domains:

$$\text{internal cooperation} + \text{external opposition}.$$

Teams generate cooperation among their own members, but frequently by defining a second group as the obstacle to be overcome. Social platforms adapt this same structure to the transmission of political and factual claims. A claim can function simultaneously as an assertion about the world and as a signal of team membership, and under sufficient polarization, successful communication of the claim becomes difficult to distinguish from scoring a point against an opposing group [10]. The alternative to this dynamic is not the wholesale elimination of comparison, disagreement, or group identity, all of which serve legitimate social functions. It is the deliberate organization of cooperation around an external task, such as solving a shared problem, rather than around the symbolic defeat of another group.

9 Epistemic Content Farms

High-volume advice accounts, of the sort that combine professional-looking imagery, large follower counts, corporate-style logos, numbered lists of instructions, warnings of concealed dangers, and claims of little-known tricks, exhibit a structure worth examining independently of any particular creator's identity or motives. Their content is characteristically organized around a repeatable formula:

$$\text{familiar service} + \text{hidden fact} + \text{personal danger or advantage} + \text{simple corrective ritual}.$$

One post from such an account may warn that an artificial intelligence system is secretly inspecting private information; the next may recommend an AI-enabled feature as an indispensable productivity tool. This is not necessarily experienced by the audience as contradiction, because each post is optimized independently against the same engagement metric rather than against a coherent underlying position.

Consistency across claims is, in fact, unnecessary when the stable product being sold to the audience is neither advocacy nor education but recurring revelation: the repeated experience of being told that something important has been concealed, followed by the reassurance that this particular creator can restore a sense of control. The business model rewards the frequency and emotional intensity of the revelation far more than it rewards the truth or coherence of its content.

It is worth distinguishing inconsistency from incoherence here, since the two are often conflated. An account that warns against one use of a technology while recommending another is not, by itself, being inconsistent in any objectionable sense; reasonable positions frequently draw exactly this kind of distinction, approving of some applications of a technology while rejecting others. The revealing problem is not the mere coexistence of a warning and a recommendation. It is the absence of any stated criterion that would explain why one form of processing is objectionable and the other acceptable. Without such a criterion, the account is not holding a differentiated position; it is producing alternating affect, tuned independently on each occasion to whichever emotional register performs best.

10 When "AI Off" Ceases to Be an Operation

The discussion of misleading posts about artificial intelligence motivates a broader and more genuinely difficult problem: the practical impossibility, for most users, of refusing AI processing alto-

gether. "AI" no longer names one separable feature that can be toggled off. It can refer, depending on context, to spam filtering, content ranking, autocomplete, image enhancement, fraud detection, product recommendation, generative summarization, retrieval-augmented search, automated decision-making, or the training of future models on a user's own data.

As a consequence,

AI Off

is not a well-defined operation in the way that a single switch implies. Turning off a visible assistant may leave automated classification of the same content fully intact elsewhere in the system. Disabling a set of features marketed as "smart" may remove desirable functionality from the user's own experience while leaving server-side processing, over which the user has no visibility or control, entirely untouched. A person may refuse, on principle, to use generative tools directly, while still having their emails summarized by a recipient's assistant, their job application ranked by an employer's automated system, or their photographs classified by a platform they did not choose to leave. The important distinction, then, is not a binary between AI and no AI, but a differentiated set of questions about which forms of processing are occurring, for what purpose, under what permissions, with what data retention, feeding what consequential decisions, and subject to what route of appeal. Treating AI risk as contextual, sociotechnical, and distributed across design, deployment, use, and evaluation, rather than as a property of a single feature that can be switched off, is itself the direction current risk-management guidance has taken [4].

11 The Economics of Contamination and Repair

Perhaps the essay's most consequential formal claim concerns the asymmetry of cost between producing a claim and verifying one. A sensational post can be produced or shared at negligible cost in time, effort, or expertise. Responsible evaluation of the same claim may require locating its original source, defining ambiguous terminology precisely, checking primary documents, reading technical or legal material closely, locating court or regulatory records, and then communicating the resulting distinctions clearly enough that they are actually usable by someone else.

Let C_p denote the cost of producing or propagating a claim, and let C_v denote the cost of adequately verifying it. In the setting of contemporary social media,

$$C_p \ll C_v.$$

Generative AI tools plausibly widen this inequality further, since they reduce C_p closer to zero without producing any comparable reduction in the institutional labor required for verification, which still depends on documents, records, and expert judgment that generative tools do not themselves supply. The recipient who cares most about accuracy is thereby assigned the largest, and the least compensated, share of the total cost of the exchange. This makes universal correction an unstable long-term strategy for any individual or institution: if every low-cost claim can successfully demand a high-cost response, then conscientiousness itself becomes an exploitable vulnerability, since it can be triggered cheaply and repeatedly by anyone willing to produce more claims than any single verifier can address.

12 Reweighting Is Not Refusal

A distinction between two superficially similar operations supplies the essay's strongest structural conclusion. Most platforms offer reweighting controls of some kind: options to see less of a source,

to hide its content temporarily, to snooze it, or to adjust the recommendation algorithm's treatment of it. These controls modify the probability that a given source will appear in a feed while leaving that source fully present within the relevant space of competing attention.

Blocking, by contrast, performs a categorically different operation. If Ω denotes the set of sources currently admitted to compete for a person's attention, then muting approximates a continuous reduction,

$$P(s) \to 0,$$

whereas blocking performs a discrete removal,

$$A_s^-(\Omega) = \Omega \setminus \{s\}.$$

Blocking a source is not a metaphysical declaration that the blocked person is permanently incapable of change, nor is it an attempt to erase the fact that the earlier exchange occurred. It is, more narrowly, a refusal to grant a particular channel continuing authority over one's own attention going forward. The earlier material need not be deleted from memory or record, and the blocked party need not be denied any further existence, whether online or off. What changes is the live admissibility of that channel to the person's ongoing attention, not the historical record of what that channel previously said. This is epistemic boundary-setting rather than the rewriting of history, and the distinction between the two is exactly the distinction this essay treats as most load-bearing: a system can refuse a source without pretending the source, or its past claims, never existed.

13 Toward an Epistemic Security Model

A limited and workable epistemic security model can be proposed in place of any fantasy of total immunity from false or manipulative information. Such a model would preserve provenance, so that the origin of a claim remains traceable. It would distinguish popularity from authority, so that a large audience is never treated as evidence of domain competence. It would introduce friction before redistribution, so that sharing requires at least a moment's deliberate attention to accuracy rather than a single frictionless gesture [9]. It would expose uncertainty rather than presenting contested claims with unwarranted confidence. It would retain correction histories, so that a claim's revision is visible rather than silently overwritten [8]. It would limit cross-context recommendation, so that engagement with one topic does not automatically route a person toward unrelated and more extreme material. And it would allow decisive, discrete refusal of specific channels, distinct from the continuous reweighting that platforms currently emphasize.

It is important that this condition attach to the surrounding information environment rather than to any individual message. A single message does not itself possess revisability or refusability, and its verification cost is not an intrinsic property of the message; these are properties of the system that delivers, ranks, and preserves it. A provisional condition for an information environment Σ, evaluated with respect to a recipient u, can be written as

$$\text{EpiSecure}(\Sigma, u) = \text{Traceable}(\Sigma) \wedge \text{Contextual}(\Sigma) \wedge \text{Corrigible}(\Sigma) \wedge \text{CostBounded}(\Sigma, u) \wedge \text{Refusable}(\Sigma, u).$$

This correctly locates the obligations at the level of the system rather than requiring each encountered message to pass some truth-admission test before it can be seen at all. $\text{EpiSecure}(\Sigma, u)$ does not require that every message be verified before it can be encountered; such a requirement would be both practically impossible and normatively objectionable in an open information environment. It requires only that a system designed with epistemic security in mind should preserve a recipient's ability to identify a claim's origin, understand why it was surfaced, correct a durable error

once it is discovered, avoid an unbounded burden of verification, and indefinitely exclude a channel that has proven persistently destructive to their own attention and judgment, unless the recipient deliberately reverses that decision.

14 Conclusion

Across every case examined here, a visible and cheaply produced proxy gradually displaces the underlying property it was originally meant to indicate:

$$\begin{aligned} \text{boot speed} &\to \text{performance}, \\ \text{security vocabulary} &\to \text{safety}, \\ \text{familiarity} &\to \text{trustworthiness}, \\ \text{followers} &\to \text{expertise}, \\ \text{engagement} &\to \text{value}, \\ \text{settings} &\to \text{consent}. \end{aligned}$$

A computer is not safe merely because a user has disabled many of its settings. A person is not epistemically secure merely because familiar sources happen to produce a coherent worldview. A platform is not socially valuable merely because its users remain engaged with it. An AI service is not meaningfully optional merely because a settings page contains several switches related to it.

The essay's ultimate claim is that epistemic security requires more than the identification of individual false propositions. It requires examining the systems, technical and social alike, that determine what receives permission to appear, to reproduce, to demand a response, and to return once it has been refused. The defining vulnerability of the social web is not merely that false ideas can enter a person's informational environment. It is that the systems governing attention profit whenever those ideas compel the recipient to keep them running.

References

[1] Ron Ross, *Guide for Conducting Risk Assessments*, NIST Special Publication 800-30, Revision 1, National Institute of Standards and Technology, 2012. `https://doi.org/10.6028/NIST.SP.800-30r1`

[2] Karen Scarfone, Murugiah Souppaya, Amanda Cody, and Angela Orebaugh, *Technical Guide to Information Security Testing and Assessment*, NIST Special Publication 800-115, National Institute of Standards and Technology, 2008. `https://doi.org/10.6028/NIST.SP.800-115`

[3] David Cooper, Andrew Regenscheid, Murugiah Souppaya, Christopher Bean, Mike Boyle, Dorothy Cooley, and Michael Jenkins, *Security Considerations for Code Signing*, NIST Cybersecurity White Paper, National Institute of Standards and Technology, 2018. `https://doi.org/10.6028/NIST.CSWP.01262018`

[4] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023. `https://doi.org/10.6028/NIST.AI.100-1`

[5] Henri Tajfel, Michael G. Billig, Robert P. Bundy, and Claude Flament, Social categorization and intergroup behaviour, *European Journal of Social Psychology*, 1(2):149–178, 1971. `https://doi.org/10.1002/ejsp.2420010202`

[6] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain, The science of fake news, *Science*, 359(6380):1094–1096, 2018. https://doi.org/10.1126/science.aao2998

[7] Soroush Vosoughi, Deb Roy, and Sinan Aral, The spread of true and false news online, *Science*, 359(6380):1146–1151, 2018. https://doi.org/10.1126/science.aap9559

[8] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook, Misinformation and its correction: Continued influence and successful debiasing, *Psychological Science in the Public Interest*, 13(3):106–131, 2012. https://doi.org/10.1177/1529100612451018

[9] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand, Shifting attention to accuracy can reduce misinformation online, *Nature*, 592:590–595, 2021. https://doi.org/10.1038/s41586-021-03344-2

[10] Steve Rathje, Jay J. Van Bavel, and Sander van der Linden, Out-group animosity drives engagement on social media, *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021. https://doi.org/10.1073/pnas.2024292118

[11] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Christina de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker, How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023. https://doi.org/10.1126/science.abp9364

[12] Killian L. McLoughlin, William J. Brady, Aden Goolsbee, Ben Kaiser, Kate Klonick, and Molly J. Crockett, Misinformation exploits outrage to spread online, *Science*, 386(6725):991–996, 2024. https://doi.org/10.1126/science.adl2829