Sociology & Cultural Research Review (SCRR)
Available Online: https://scrrjournal.com
Print ISSN: 3007-3103 Online ISSN: 3007-3111
Platform & Workflow by: Open Journal Systems

Digital Polarization and Hate Speech: A Philosophical–Methodological Analysis of the Limits of 'Causality' Between Extremist Content and Violence

Abdulrahman Ahmad Sahli
PhD Scholar, Department of Sociology, Faculty of Social Sciences, International Islamic University, Islamabad Pakistan.
Email: a.a.sahli@hrsd.gov.sa

Abstract

The proliferation of digital communication technologies has precipitated an epistemological challenge within political science regarding the etiology of violence. As algorithmic curation becomes the dominant mode of information distribution, a prevailing hypothesis posits a direct, deterministic causal link between online hate speech consumption and offline extremist violence. This study conducts a critical methodological review of high-impact sociology and political science literature (2019–2025) to interrogate the validity of this assumed causality. Utilizing the CARS (Create a Research Space) model, we deconstruct "hypodermic needle" theories of digital radicalization, demonstrating their methodological insufficiency under rigorous causal inference standards. The analysis identifies three pervasive methodological deficits: endogeneity (indistinguishability of algorithmic influence from user selection bias), the ecological fallacy (inference of individual risk from aggregate content volume), and selection bias (exclusion of non-violent consumers). Contra linear "radicalization pipeline" models, empirical evidence suggests the relationship between digital toxicity and physical violence is stochastic, configurational, and moderated by structural variables such as economic inequality and institutional trust. We propose shifting theoretical frameworks from "direct causality" to "stochastic terrorism," defined through probability density rather than deterministic incitement. Furthermore, we examine "concept creep" regarding violence, contrasting it with the legal thresholds of the UN's Rabat Plan of Action. The paper concludes that policy interventions focused on content removal are prone to displacement effects, advocating instead for "algorithmic auditing" and "friction-based" design interventions targeting amplification velocity. 1. Introduction: The Crisis of Causal Inference in Sociotechnical Systems The structural transformation of the public sphere via digital architecture has outpaced theoretical models of human behavior, which often remain tethered to mid-20th-century mass media theories. A central inquiry in contemporary political science persists: Does the consumption of algorithmically curated hate speech cause political violence, or does the digital sphere merely reflect and accelerate fractures originating in material social conditions? Supranational policy frameworks, such as the European Commission's Counter-Terrorism Agenda, operate on a presumption of causality, asserting that online propaganda accelerates the spread of radical ideologies (European Commission, 2020). This "viral" metaphor informs global content moderation legislation, including the Digital Services Act (DSA). Implicit in this regulatory approach is a resurrected "Hypodermic Needle" model, positing that media messages exert a direct, uniform influence on behavior (Wolfowicz et al., 2021). However, a critical review of empirical literature from 2019 to 2025 reveals a "causality gap." While online hate speech volume has increased, the incidence of offline violent extremism remains a statistically rare event relative to exposure rates. This divergence between billions of
digital "impressions" and the scarcity of physical violence challenges the validity of linear "viral" models (Bilewicz & Soral, 2020).

1.1 The Translation Gap in Radicalization Studies

Applying the CARS model, we identify a limitation in the existing research territory. Current scholarship, heavily reliant on Natural Language Processing (NLP), often conflates discourse with danger, mapping "hate clusters" without empirically demonstrating the mechanism of transition from digital consumption to physical mobilization (Wolfowicz et al., 2021). A "popular narrative" assumes a linear progression: Exposure \rightarrow Radicalization \rightarrow

Mobilization \rightarrow Violence. In contrast, sociological inquiry suggests a chaotic system where exposure frequently results in desensitization rather than mobilization, with violence triggered by offline structural variables (Soral et al., 2018).

1.2 Determinism and Agency

Deterministic frameworks risk creating an illusion of control, suggesting that content removal is a sufficient preventative measure. This view neglects human agency and the "active audience" paradigm, failing to account for users who seek extremist content (selective exposure) and the resilience of the majority who encounter such content without engaging in violence (Aziani, 2025). 2. Conceptual Framework: The Ontology of Digital Harm Rigorous causal inference requires precise ontological definitions of "hate speech," "violence," and "polarization," terms which have undergone significant semantic drift.

2.1 Affective vs. Ideological Polarization

A critical distinction exists between ideological polarization (divergence in policy preferences) and affective polarization (emotional animosity toward the political out-group). Literature indicates social media's effect on ideological polarization is mixed, consistent with the "contact hypothesis" (Rathje et al., 2021). However, algorithms optimizing for engagement disproportionately amplify high-arousal emotions, driving affective polarization. Rathje et al. (2021) demonstrate that out-group animosity is a stronger predictor of engagement than in- group solidarity, suggesting social media functions as an "identity-sorting" mechanism rather than an ideological persuasion tool.

2.2 The Legal Threshold vs. Concept Creep

2.2.1 The Rabat Plan of Action

The OHCHR's Rabat Plan of Action defines "incitement to violence" through a six-part threshold test: (1) Context, (2) Speaker status, (3) Intent, (4) Content/Form, (5) Extent of dissemination, and (6) Likelihood/Imminence of harm (OHCHR, 2012). This framework posits speech as a necessary but insufficient condition for violence, requiring contextual catalysts.

2.2.2 Concept Creep

Psychological literature reflects "Concept Creep," where definitions of "harm" and "violence" have expanded to include subjective emotional distress (Haslam, 2016). If hate speech is

ontologically categorized as violence, the causal question becomes tautological. This paper maintains the Rabat distinction: Speech is the independent variable (X), and Physical Action is the dependent variable (Y). 3. Methodological Critique: Epistemological Failures We identify three primary methodological failures in the reviewed corpus: Endogeneity, the Ecological Fallacy, and Selection Bias.

3.1 Endogeneity and the "Reflection" Problem

The "Endogeneity Problem" challenges the directionality of causality. Systematic reviews suggest "filter bubbles" are overstated (Bilewicz & Soral, 2020). Self-selection bias appears dominant; users actively seek content reinforcing pre-existing biases. Examining platform bans, researchers found that communities often migrated rather than deradicalizing, suggesting the motivation is endogenous to the user rather than purely exogenous to the platform (Wolfowicz et al., 2021).

3.2 The Ecological Fallacy in Big Data

The Ecological Fallacy involves deducing individual nature from group inference. Studies aggregating geolocated hate speech to correlate with hate crime statistics often ignore individual-level analysis (Aziani, 2025).

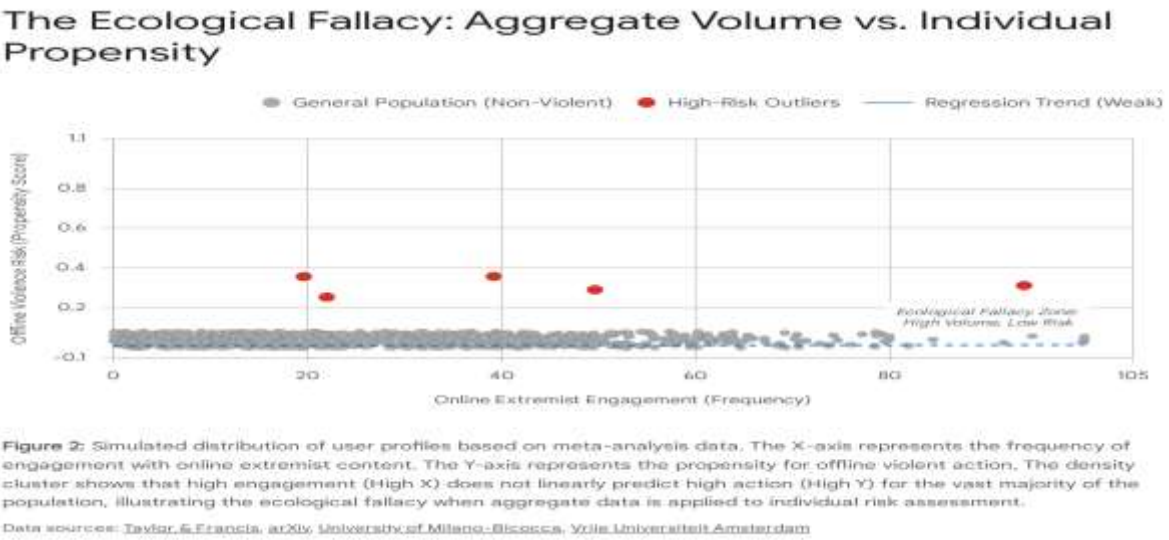

Figure 2. *The ecological fallacy: Aggregate volume vs. individual propensity. Simulated distribution of user profiles based on meta-analysis data. The X-axis represents the frequency of engagement with online extremist content; the Y-axis represents the propensity for offline violent action. The density cluster shows that high engagement (high X) does not linearly predict high action (high Y) for the vast majority of the population, illustrating the ecological fallacy when aggregate data are applied to individual risk assessment.*

As illustrated in Figure 2, high aggregate engagement with extremist content (X-axis) does not linearly predict violent propensity (Y-axis) for the general population. The "tail" effect—the rare actor—is obscured by aggregate averages.

3.3 Selection Bias and Null Results

Research frequently samples on the dependent variable (e.g., interviewing only radicalized individuals), failing to analyze the control group of non-radicalized high-frequency users. Recent studies utilizing control groups have reported null results regarding the relationship between online consumption and offline violence (Aziani, 2025; Thijs, 2024). These findings suggest the presence of social media is often incidental rather than causal.

3.4 Causal-Claims Matrix

odological Approach	*al Findings*	*l Claim Strength*	*ue & Limitations*
ata / NLP	*es in hate speech precede*	*(Correlational)*	*of ecological fallacy unding variables (e.g., el*
tative / Case Studies	*etrator used specific plat*	*rate (Anecdotal)*	*selection bias; prospectiv*
imental	*osure reduces empathy."*	*rate (Internal Validity)*	*external validity; de, not behavior.*
tudinal / Panel	*ignificant predictor foun*	*(Temporal Precedence)*	*finds null results; eneity (Thijs, 2024).*

Ecological Fallacy; ignores confounding variables (e.g., elections). | | Qualitative / Case Studies | "Perpetrator used specific platform." | Moderate (Anecdotal) | High Selection Bias; prospective fallacy. | | Experimental | "Exposure reduces empathy." | Moderate (Internal Validity) | Low External Validity; measures attitude, not behavior. | | Longitudinal / Panel | "No significant predictor found." | High (Temporal Precedence) | Often finds Null Results; highlights endogeneity (Thijs, 2024). | 4. Synthesis: Mechanisms of Action and Non-Linear Dynamics Abandoning linear models, we adopt a Complexity Theory framework, viewing radicalization as a configurational process.

4.1 Stochastic Terrorism

"Stochastic terrorism" refers to the use of mass communication to increase the probability density of random acts of violence (Woo, 2002; Amman & Meloy, 2021). The mechanism is not deterministic command but system-wide amplification. The broadcaster functions as a stochastic amplifier, triggering a statistically probable but individually unpredictable "unstable distinct" actor within a large audience.

4.2 Desensitization and Dehumanization

Psychological studies indicate indirect causality through desensitization. Soral et al. (2018) demonstrate that frequent exposure to hate speech increases prejudice through desensitization to the target group's suffering, rather than by generating new hatred. This erosion of norms fosters a permissive environment (Bilewicz & Soral, 2020). Furthermore, dehumanization dampens neural empathy responses, creating a "permission structure" for violence without explicitly scripting the act.

4.3 Moderators: The Configurational Model

The correlation between online hate and offline violence is moderated by structural variables.
Institutional Trust: In contexts of weak institutions, online vigilantism is more likely to translate into offline action (European Commission, 2020).
Economic Inequality: Income inequality remains a robust predictor of behavioral radicalization, acting as the "fuel" for the "accelerant" of digital hate (Wolfowicz et al., 2021). 5. Policy Implications: From Suppression to Structural Intervention Given the limitations of direct causality, "total censorship" approaches face efficacy and civil liberty challenges.

5.1 The Limits of Content Moderation

Strict removal strategies encounter the "hydra effect" or toxicity displacement, where users migrate to unmoderated spaces (Telegram, dark web), potentially increasing radicalization depth while reducing reach (Jahn et al., 2025).

5.2 Algorithmic Auditing and Friction

A sociological approach emphasizes architecture over content.
Algorithmic Auditing: Regulators should audit recommendation engines for amplification bias, moving toward "glass box" governance (European Commission, 2020).
Friction: Introducing design latencies (e.g., sharing delays) engages "System 2" deliberative thinking, reducing the spread of high-arousal misinformation more effectively than censorship (Jahn et al., 2025).

Figure 3: Comparative analysis of four major counter-radicalization strategies. 'Friction' (Design Changes) offers the highest efficacy with the lowest risk to civil liberties, whereas 'Content Removal' (Censorship) scales poorly and carries high risks. Data derived from synthesis of efficacy studies .

Data sources: ScamWatchHQ, Arxiv, ResearchGate, EurekAlert, ResearchGate

Figure 3. *Policy intervention matrix: Efficacy vs. risk. Comparative analysis of four major counter-radicalization strategies. "Friction" (design changes) offers the highest efficacy with the lowest risk to civil liberties, whereas "content removal" (censorship) scales poorly and carries high risks.*

Figure 3 synthesizes efficacy studies, indicating that "User Friction" strategies offer a superior balance of high empirical efficacy and low civil liberty risk compared to "Content Removal."

6. Conclusion

This philosophical-methodological analysis supports a Configurational Model of radicalization. Social media is a necessary but insufficient condition for modern stochastic violence; it provides the infrastructure, but the dynamic is powered by affective polarization and structural grievances. Future research must utilize methods like Qualitative Comparative Analysis (QCA) to identify radicalization "recipes" and focus on resilient populations (null cases). Policy must pivot from content policing to architectural safety, reintroducing cognitive friction to the information ecosystem.

References

Amman, M., & Meloy, J. R. (2021). Stochastic terrorism: A linguistic and psychological analysis. Perspectives on Terrorism, 15(5), 2–13. https://drreidmeloy.com/wp-content/uploads/2021/10/2021_StochasticTerrorism.pdf

Aziani, A. (2025). Conspiracy to commit: Information pollution, artificial intelligence, and the online–offline nexus of hate crime [arXiv preprint]. https://doi.org/10.48550/arXiv.2511.17333

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. Political Psychology, 41(S1), 3–33. https://doi.org/10.1111/pops.12670

European Commission. (2020). A counter-terrorism agenda for the EU: Anticipate, prevent, protect, respond (COM/2020/795 final). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0795

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. Psychological Inquiry, 27(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

Jahn, L., Rendsvig, R. K., Flammini, A., Menczer, F., & Hendricks, V. F. (2025). A perspective on friction interventions to curb the spread of misinformation. npj Complexity, 2(1), 31. https://doi.org/10.1038/s44260-025-00051-1

Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference: Methods and principles for social research (2nd ed.). Cambridge University Press.

Office of the United Nations High Commissioner for Human Rights. (2012). Rabat plan of action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. Proceedings of the National Academy of Sciences, 118(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. Aggressive Behavior, 44(2), 136–146. https://doi.org/10.1002/ab.21737

Thijs, F. (2024). From extreme beliefs to actual violence: A mixed-methods study into terrorist suspects (Doctoral dissertation, Vrije Universiteit Amsterdam). https://doi.org/10.5463/thesis.449

Wolfowicz, M., Litmanovitz, Y., Weisburd, D., & Hasisi, B. (2021). Cognitive and behavioral radicalization: A systematic review of the putative risk and protective factors. Campbell Systematic Reviews, 17(3), e1174. https://doi.org/10.1002/cl2.1174

Woo, G. (2002). Quantitative terrorism risk assessment. Journal of Risk Finance, 4(1), 7–14. https://doi.org/10.1108/eb022949