Wah Academia Journal of Social Sciences
Volume 5, Issue 1, Mar 2026, pp. 627-646
ISSN – E 2958-8731 P 2958-8723
https://doi.org/10.63954/WAJSS.5.1.33.2026

Weaponization of Social Media Algorithms: Computational Propaganda, Manufactured Consent, and the Subversion of Democratic Communication

Dr. Taha Shabbir[1], Dr. Muhammad Aftab Madni[2] and Dr. Usman Farooq[3]

Abstract

Social media platforms have fundamentally restructured the architecture of public communication by routing information through proprietary algorithmic systems designed to maximise user engagement. These systems, originally conceived as neutral tools for content discovery, have been systematically exploited by state actors, political operatives, commercial interests, and extremist organisations to manufacture consent, amplify disinformation, suppress dissent, and polarise publics. This paper examines the weaponization of social media algorithms as a multi-dimensional phenomenon encompassing computational propaganda, coordinated inauthentic behaviour, micro-targeted political advertising, and algorithmic radicalisation. Drawing on political communication theory, platform studies, critical data studies, and documented case evidence from multiple geopolitical contexts, the paper argues that algorithmic weaponization constitutes a structural threat to democratic communication that cannot be resolved by content moderation alone. The analysis interrogates the political economy of attention-maximising platforms, the technical mechanisms through which algorithms are exploited, the role of disinformation actors ranging from troll farms to nation-state intelligence agencies, and the asymmetric vulnerabilities of developing democracies including Pakistan. The paper concludes by proposing an integrated framework of platform regulation, algorithmic transparency, civic literacy, and international coordination as necessary conditions for reclaiming the democratic potential of networked communication.

Keywords: algorithmic weaponization, computational propaganda, social media, disinformation, coordinated inauthentic behaviour, political communication, platform regulation, echo chambers, filter bubbles, democratic communication

[1]Department of Computing, Faculty of Engineering Sciences & Technology, Hamdard University, Karachi– Pakistan
[2]Department of Media and Communication Studies, Shaheed Benazir Bhutto University, Shaheed Benazirabad, Nawabshah – Pakistan
[3]Department of Media Studies & Design, Faculty of Communication & Design, Indus University, Karachi – Pakistan

Introduction

The relationship between communication technology and democratic governance has always been contested. From the printing press to radio to television, each successive medium has simultaneously expanded the possibilities for democratic deliberation and introduced new vulnerabilities to manipulation, monopolisation, and propaganda. Social media represents the latest and, in many respects, the most consequential iteration of this pattern. The apparent democratisation of communication that social media promised universal access to a global public sphere, citizen journalism, horizontal peer-to-peer information sharing has been systematically undermined by the logic of the platforms that host it. Those platforms are not neutral conduits. They are algorithmic architectures whose design logic, business incentives, and technical mechanisms can be and have been weaponised against the democratic purposes they ostensibly serve.

The concept of algorithmic weaponization denotes the deliberate exploitation of social media recommendation and amplification systems to advance political, ideological, or commercial objectives through illegitimate means means that distort, rather than facilitate, informed democratic participation. This encompasses a wide range of actors, tactics, and objectives: authoritarian governments deploying troll armies and bot networks to flood the information environment with pro-regime content; political campaigns using psychographic micro-targeting to manipulate voter behaviour; extremist networks exploiting recommendation algorithms to radicalise vulnerable users; and commercial disinformation-for-hire industries that treat influence operations as a business service. What unites these disparate practices is the common exploitation of a shared technical substrate: the engagement-maximising algorithms that determine what billions of people see, when they see it, and how prominently it is displayed.

This paper offers a comprehensive analysis of the weaponization of social media algorithms across its principal dimensions. Following an examination of the theoretical foundations of algorithmic media and their political implications, the paper analyses the specific mechanisms of algorithmic weaponization, including coordinated inauthentic behaviour, micro-targeted advertising, and recommendation-driven radicalisation. The paper then situates this phenomenon within the global landscape of computational propaganda, drawing on documented case studies from Russia, China, the United States, and South Asia. A critical section examines the structural conditions that make contemporary platforms susceptible

to weaponization, focusing on the political economy of surveillance capitalism. The paper then addresses the particular vulnerabilities of developing democracies before proposing a framework for resistance and remediation.

Theoretical Foundations: Algorithms, Power, and Democratic Communication

❖ The Political Economy of Engagement Maximisation

To understand algorithmic weaponization, it is necessary to first understand why social media algorithms exist and what purposes they serve. The dominant business model of social media platforms advertising-funded, free-to-access requires the monetisation of user attention. Advertising revenue is a function of time spent on platform, which in turn is maximised by serving content that provokes strong emotional responses and compels continued engagement. This design logic, sometimes called the attention economy, creates a systematic bias in favour of content that is emotionally arousing, morally provocative, novel, and identity-affirming regardless of its accuracy, civic value, or contribution to informed deliberation (Zuboff, 2019; Wu, 2017).

The consequences of this design logic for the information ecosystem have been extensively documented. Vosoughi et al. (2018) demonstrated in a landmark Science study that false news spreads significantly faster and more broadly on Twitter than accurate news, because false stories are on average more novel and emotionally provocative precisely the qualities that engagement-maximising algorithms reward. Brady et al. (2017) showed that the use of moral-emotional language in political tweets substantially increases their spread, suggesting that algorithmic amplification creates selective pressure for the most inflammatory political communication. Berger and Milkman (2012) established that content arousing high-activation emotions anger, anxiety, awe spreads more virally than content arousing low-activation emotions such as sadness or contentment. The cumulative effect of these dynamics is an information environment systematically tilted toward outrage, polarisation, and disinformation.

❖ Gillespie's Politics of Platforms and Algorithmic Governance

Tarleton Gillespie's (2014) influential framework of the 'politics of platforms' provides an essential theoretical resource for understanding algorithmic weaponization. Gillespie argued that platforms embed normative judgements about

relevance, credibility, and appropriate expression in their technical architectures, presenting these judgements as neutral and objective while they in fact reflect the commercial interests, cultural assumptions, and political vulnerabilities of platform operators. The algorithm is not a mirror of social reality but a constitutive force that shapes what counts as important, credible, or worthy of attention. This reconceptualisation of platforms as political actors rather than passive conduits is the precondition for understanding how their algorithms can be weaponised.

Bucher (2018) extended this framework through her concept of the 'threat of invisibility' the algorithmic penalty of non-exposure that disciplines platform users toward forms of communication that maximise engagement. Users who produce content optimised for algorithmic amplification are rewarded with reach; those who produce content that does not provoke strong reactions are rendered invisible. This dynamic creates incentive structures that systematically favour sensationalism, partisanship, and emotional provocation over accuracy, nuance, and deliberative engagement. Political actors who understand these incentive structures whether authoritarian governments, radical movements, or sophisticated political campaigns can exploit them to amplify their messages far beyond what their organic audience or the quality of their content would otherwise warrant.

- **Manufactured Consent in the Algorithmic Age**

Herman and Chomsky's (1988) propaganda model of media described how elite ownership, advertising dependency, sourcing conventions, and ideological consensus operate as 'filters' that shape news production in the interests of dominant economic and political elites. Writing before the internet era, the model could not anticipate the specific mechanisms of algorithmic amplification, but its core insight that apparently free and diverse media systems can function as instruments of manufactured consent remains profoundly relevant. The weaponization of social media algorithms represents a new, distributed, and far more personalised form of manufactured consent in which individual users receive customised information environments shaped by a combination of commercial design logic and deliberate manipulation.

Where Herman and Chomsky's filters operated at the level of media institutions, algorithmic filters operate at the level of individual experience. Each user receives a bespoke information environment, or 'personal propaganda environment,' in which the content they encounter has been shaped by their past behaviour, the behaviour

of algorithmically similar users, and the strategic interventions of actors who understand how to exploit the platform's ranking systems. This personalisation makes manipulation simultaneously more powerful because it is precisely targeted to individual vulnerabilities and preexisting beliefs and more difficult to detect because there is no common information environment against which individual experience can be calibrated (Pariser, 2011; Zuboff, 2019).

Mechanisms of Algorithmic Weaponization

❖ Coordinated Inauthentic Behaviour: Bots, Trolls, and Astroturfing

The most extensively documented form of algorithmic weaponization is coordinated inauthentic behaviour (CIB) the use of fake accounts, automated bots, and organised human actors to simulate organic public sentiment, amplify selected messages, and suppress or harass opposing voices. The Oxford Internet Institute's Computational Propaganda Project documented CIB operations in 81 countries between 2010 and 2020, identifying government and political party operatives as primary actors in the majority of cases (Bradshaw & Howard, 2019). These operations exploit the engagement signals that social media algorithms use to determine content prominence: likes, shares, comments, and follower counts. By artificially inflating these signals through coordinated inauthentic activity, manipulators cause platforms' recommendation algorithms to interpret manufactured engagement as evidence of genuine popularity and to amplify the target content to wider audiences.

The Internet Research Agency (IRA), the Russian government-linked organisation that conducted extensive CIB operations targeting the 2016 United States presidential election, exemplifies the sophistication that state-level actors can bring to algorithmic weaponization. The IRA operated networks of fake accounts across Facebook, Twitter, Instagram, and YouTube, creating authentic-seeming communities organised around divisive political identities racial justice, gun rights, immigration, evangelical Christianity and then using these communities to amplify content designed to exacerbate existing social divisions (Mueller, 2019; DiResta et al., 2018). The IRA's operations were not primarily designed to promote specific candidates or policies but to degrade the epistemic environment to increase political distrust, social division, and cynicism about democratic institutions. This strategy of epistemic sabotage is arguably more effective and more durable than

conventional propaganda because it targets the preconditions for democratic deliberation rather than the outcomes of any specific debate.

Bot networks operate by automating the creation and operation of fake accounts at scales that would be impossible with human labour alone. Automated accounts can post content, follow and unfollow users, like and share posts, and manufacture trending hashtags creating the appearance of grassroots movements, or 'astroturfing,' that can influence both platform algorithms and the perception of social consensus among genuine users. Ferrara et al. (2016) estimated that between 9% and 15% of active Twitter accounts in the period around the 2016 US election were bots, with significant concentrations in political hashtags and discussions. The presence of bots in political conversations not only amplifies specific messages but also corrupts the social information environment from which genuine users draw inferences about public opinion, normalising positions that are in reality minority views amplified by artificial activity.

- **Micro-Targeted Political Advertising**

The weaponization of social media's advertising targeting capabilities represents a qualitatively different form of algorithmic exploitation than CIB, one that operates through officially sanctioned platform features rather than policy violations. Social media advertising platforms particularly Facebook's offer advertisers the ability to target audiences with extraordinary specificity, combining demographic variables, interest categories, behavioural data, and psychographic profiles derived from platform behaviour. Political advertisers have exploited these capabilities to deliver precisely crafted messages to audiences selected for their susceptibility to specific emotional appeals, a practice enabled by and entirely dependent on the data infrastructure of surveillance capitalism (Zuboff, 2019; Cadwalladr & Graham-Harrison, 2018).

The Cambridge Analytica scandal brought the political weaponization of psychographic micro-targeting to global attention. Cambridge Analytica, a political data firm with ties to the Trump campaign and the Leave.EU Brexit campaign, harvested psychological profiles of up to 87 million Facebook users through a third-party quiz application and used these profiles to deliver targeted political advertising calibrated to individual psychological vulnerabilities specifically the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). The firm claimed to be able to identify 'persuadable' voters in key

constituencies and deliver content specifically designed to move them toward desired political positions (Cadwalladr & Graham-Harrison, 2018). While independent assessments of Cambridge Analytica's actual effectiveness have been mixed, the scandal exposed the potential for political actors to exploit platforms' advertising infrastructure in ways that are invisible to targeted individuals and to the broader public.

Dark posts Facebook advertisements visible only to their intended targets are particularly susceptible to weaponization because they cannot be monitored by journalists, opposition campaigns, or regulators. A political actor can simultaneously deliver contradictory messages to different audience segments, telling one group that their candidate will protect gun rights while telling another that the same candidate supports common-sense gun reform, with no accountability for the contradiction because neither audience can see what the other is receiving. This micro-targeted message inconsistency is not merely a matter of emphasis or framing; it represents a fundamental deception of the electorate that is enabled and facilitated by platform advertising infrastructure (Howard, 2020).

- **Algorithmic Radicalisation and the Recommendation Funnel**

One of the most consequential and extensively debated forms of algorithmic weaponization is the role of recommendation algorithms in facilitating ideological radicalisation. YouTube's recommendation system, which accounts for over 70% of the platform's total viewing time, operates by predicting the next video a user is likely to engage with based on their watch history, the behaviour of users with similar histories, and real-time engagement signals. Ribeiro et al. (2020) conducted an audit of YouTube's recommendation system and documented what they described as a 'radicalization pathway': a pattern in which the algorithm progressively recommends content from mainstream political channels to more ideologically extreme content in the same broad category, as extreme content tends to be more engaging and thus more algorithm-rewarding.

The recommendation funnel dynamic was also documented by journalist Kevin Roose (2019) through his 'rabbit hole' experiment, in which he created a fresh YouTube account and allowed the algorithm to guide his viewing from mainstream news content through increasingly extreme political material. Within a short period, the algorithm had led him from network news to far-right conspiracy content. Algorithmic radicalisation of this kind is not necessarily the result of deliberate

manipulation by external actors but is an emergent property of engagement-maximising design: extremist content is, by design of the platforms, disproportionately amplified because it provokes strong emotional responses and high engagement. Bad actors who understand this dynamic can position their content to be algorithmically discovered by users who begin with mainstream political interests, exploiting the recommendation system to recruit audiences they could not reach through direct appeal.

Whittaker et al. (2021) analysed the role of YouTube's recommendation system in the spread of conspiracy theories, including QAnon content, demonstrating that the platform's algorithmic architecture systematically exposed users to increasingly radical content when they engaged with gateway conspiracy theories. The implications extend beyond individual platforms: radicalization pathways that begin on YouTube may drive users toward more extreme platforms such as Telegram, Gab, or 4chan, where even less moderation exists and where operational planning for real-world violence has been documented. The algorithmic recommendation system thus functions as a pipeline from mainstream political engagement to extremist communities, with profound consequences for political violence, democratic stability, and social cohesion.

The Global Landscape of Computational Propaganda

❖ State-Sponsored Information Operations

State-sponsored computational propaganda has emerged as a central instrument of contemporary geopolitical competition. Russia's information operations, conducted primarily through the Internet Research Agency and the GRU's military intelligence apparatus, have targeted elections and social cohesion in the United States, the United Kingdom, France, Germany, and multiple other democracies (Mueller, 2019; European External Action Service, 2020). China's 'Fifty Cent Army' named for the alleged fee paid per post to online commenters conducts sophisticated information operations aimed both at domestic audiences (promoting pro-government narratives and discrediting critics) and at international audiences in contested geopolitical spaces including Hong Kong, Taiwan, and the South China Sea (King et al., 2017).

China's approach to computational propaganda is distinctive in its emphasis on what King et al. (2017) termed 'distraction' rather than direct argumentation. Rather

than attempting to persuade audiences of specific positions through argument, Chinese government-linked information operations flood the information environment with cheerful, positive, nationalistic content during periods of social tension, effectively overwhelming critical discourse with pro-regime noise and making it harder for opposition voices to be heard. This strategy of informational flooding exploiting algorithmic amplification through coordinated volume is arguably more suited to an engagement-maximising platform environment than direct refutation, which risks amplifying the very criticisms it seeks to counter.

Iran, Saudi Arabia, and other states in the Middle East and South Asia have conducted documented information operations through social media, targeting both domestic and diaspora audiences (Stanford Internet Observatory, 2020). The Computational Propaganda Project's cross-national comparative research identified common tactical templates across state-sponsored operations in different countries, including the use of fake accounts to inflate follower counts and engagement metrics, coordinated posting campaigns to manufacture trending topics, and the targeted harassment of journalists and political opponents (Bradshaw & Howard, 2019). These commonalities suggest a global diffusion of computational propaganda techniques, with authoritarian and semi-authoritarian governments adapting and sharing methods across borders.

- **Commercial Disinformation-for-Hire**

Alongside state-sponsored operations, a commercial disinformation-for-hire industry has developed that provides influence operation services to political clients, corporations, and any actor with sufficient resources and sufficiently few ethical constraints. Firms such as Cambridge Analytica (now defunct), various 'black PR' agencies, and a proliferating ecosystem of boutique political data consultancies offer services ranging from audience psychographic profiling and micro-targeted advertising to the creation and management of fake account networks. Ong and Cabanes (2018) documented a particularly sophisticated commercial disinformation ecosystem in the Philippines, where multiple tiers of operators from macro-influencers and bloggers to low-paid click workers managing bot networks were hired by political clients to manufacture online consensus during the 2016 elections that brought Rodrigo Duterte to power.

The commercial disinformation industry represents a particularly difficult governance challenge because it diffuses the operational capacity for algorithmic

weaponization beyond the reach of state-level intelligence and law enforcement. A political actor who wishes to conduct an influence operation no longer needs state resources or specialised intelligence apparatus; they can purchase the service on an open market, with deniability afforded by the commercial intermediary. The emergence of AI-powered tools for content generation, account management, and audience targeting further lowers the barrier to entry, enabling even relatively small political actors to conduct sophisticated influence operations at modest cost (Bradshaw & Howard, 2019).

- **Platform Complicity and the Accountability Deficit**

A critical dimension of the computational propaganda landscape that is frequently underemphasised is the role of platforms not merely as passive victims of weaponization but as active, if unwitting, co-producers of the conditions that make weaponization possible. Platform revenue is partially generated by the very same political advertising that enables micro-targeted manipulation; platforms' engagement-maximising algorithms are the mechanism through which coordinated inauthentic behaviour achieves amplification; and platforms' data practices create the psychographic profiles that political advertisers exploit. Facebook's internal research, leaked by whistleblower Frances Haugen in 2021, revealed that the company's own studies had identified its algorithms' role in amplifying hate speech, political extremism, and health misinformation, and that the company had repeatedly deprioritised remediation of these harms in favour of engagement metrics (Horwitz, 2021).

The Haugen disclosures confirmed what platform studies scholars had argued theoretically: that platforms' commercial incentives are structurally misaligned with their social responsibilities, and that this misalignment is not an aberration or the result of insufficient attention but a constitutive feature of the advertising-funded platform business model. Platforms profit from engagement; engagement is maximised by emotionally arousing content; emotionally arousing content disproportionately includes disinformation, outrage, and extremism. This structural complicity does not exonerate external actors who deliberately weaponise algorithms, but it does establish that algorithmic weaponization is not purely an exogenous threat but is enabled and amplified by endogenous platform design choices.

Surveillance Capitalism and the Structural Conditions of Weaponization

❖ The Data Infrastructure of Manipulation

Shoshana Zuboff's (2019) theory of surveillance capitalism provides the most comprehensive critical framework for understanding the structural conditions that make social media algorithmic weaponization possible and, in a sense, inevitable. Zuboff argued that the dominant technology platforms have pioneered a new economic logic in which human behaviour in its most granular, intimate, and predictive dimensions is claimed as raw material for the production of 'behavioural prediction products' sold to advertisers and other institutional buyers. The massive data infrastructure required to produce these prediction products the surveillance apparatus that tracks users' every click, pause, like, share, search, and interpersonal interaction creates precisely the detailed psychographic map of individual vulnerabilities that political manipulators require to conduct targeted influence operations.

The data infrastructure of surveillance capitalism is not a byproduct of social media platforms; it is their core product. Everything else the social networking functionality, the content sharing, the messaging is the bait that attracts users to the surveillance apparatus. This reframing has critical implications for the governance of algorithmic weaponization: proposals that treat manipulation as a moderation problem to be addressed by removing bad content or bad accounts leave the underlying data infrastructure intact, and with it the structural conditions that enable weaponization. As long as platforms collect the detailed behavioural data that enables psychographic profiling, and as long as they sell access to this profiling capability through advertising systems, the potential for political manipulation through micro-targeted advertising will persist regardless of content moderation policies.

❖ Attention as a Political Resource

Tim Wu's (2017) historical account of the attention economy situated social media's information dynamics within a longer history of the commercialisation of human attention as a political and economic resource. Wu documented how each major communications technology from early newspapers through radio and television was initially celebrated for its democratising potential before being captured by commercial and state interests that harnessed it for manipulation and propaganda. The pattern is remarkably consistent: the medium that empowers ordinary citizens in its early phase becomes, through the logic of scale and commercialisation, an instrument of concentrated power.

Social media has followed this pattern with exceptional speed and thoroughness. The Twitter of the Arab Spring which genuinely facilitated horizontal, leaderless political coordination by marginalised citizens is not obviously the same Twitter/X of the mid-2020s, in which the platform's algorithm and content moderation policies are directly shaped by its owner's political preferences and commercial relationships. The Facebook that connected isolated individuals and communities in its early phase became, within a decade, the most powerful advertising surveillance machine in history and the primary vector for computational propaganda operations targeting elections on multiple continents. The attention that social media commands from billions of users is a political resource of extraordinary value, and its weaponization by actors who understand the technical and psychological mechanisms of the platform is a predictable consequence of this value.

Asymmetric Vulnerabilities: Developing Democracies and the Global South

- **Institutional Fragility and Information Environment Degradation**

The weaponization of social media algorithms does not affect all democratic systems equally. Countries with strong democratic institutions, independent media, high levels of civic literacy, and robust regulatory frameworks are substantially better equipped to absorb and resist algorithmic manipulation than those with fragile institutions, partisan media landscapes, and underdeveloped regulatory capacity. The structural features of democratic vulnerability political polarisation, distrust of institutions, weak fact-checking ecosystems, encrypted messaging as primary information infrastructure are disproportionately present in developing democracies across South Asia, Southeast Asia, Sub-Saharan Africa, and Latin America (Roozenbeek et al., 2020).

For Pakistan, the weaponization of social media algorithms intersects with pre-existing features of the information environment to produce conditions of exceptional susceptibility. The country's media landscape is characterised by acute partisan polarisation, commercial media ownership linked to political interests, a tradition of state pressure on independent journalism, and a WhatsApp-centric information ecosystem in which unverified content circulates at scale through encrypted, effectively unmoderable channels. The combination of high social media penetration, low average digital literacy, weak fact-checking infrastructure, and intense political conflict creates conditions in which algorithmic weaponization by

domestic political actors, foreign state operatives, and commercial influence-for-hire services can achieve substantial effects.

❖ Language, Localisation, and the Moderation Gap

A critical dimension of algorithmic weaponization's disproportionate impact on developing democracies is the systematic failure of platforms' content moderation and safety systems in non-English languages. Platform moderation infrastructure both AI classifiers and human review teams has been overwhelmingly developed for English-language content, with substantially less investment in languages such as Urdu, Sindhi, Pashto, Punjabi, Arabic, Hindi, Swahili, and Myanmar's Burmese. Facebook's internal assessments, revealed through the Haugen disclosures, acknowledged that the platform's AI systems were significantly less effective at detecting hate speech and disinformation in non-English languages (Horwitz, 2021).

This moderation gap has had catastrophic consequences in several documented cases. In Myanmar, Facebook's platform served as the primary distribution mechanism for the anti-Rohingya hate speech and incitement to violence that preceded the 2017 genocide; the platform's AI moderation systems were unable to effectively detect Burmese-language content, and human moderation resources were grossly insufficient for the market's size (UN Human Rights Council, 2018). In Ethiopia, Facebook's inadequate moderation of content in Amharic, Oromo, and Tigrinya contributed to the spread of incitement during the Tigray conflict. In India, WhatsApp-distributed disinformation in Hindi and regional languages triggered mob lynchings. These cases establish a pattern in which platforms' moderation gaps in non-English languages create conditions for the most severe forms of algorithmic weaponization the mobilisation of violence against targeted populations with effects concentrated in the communities with the least institutional capacity to resist them.

❖ State Weaponization and Authoritarian Capture

Developing democracies face not only the threat of external algorithmic weaponization but also the risk of domestic state actors weaponising social media against their own citizens. Authoritarian and semi-authoritarian governments have demonstrated sophisticated capacity to exploit platform algorithms for surveillance, targeted harassment, and the suppression of political opposition. In Pakistan, credible accounts of state-linked social media operations targeting opposition politicians, journalists, and civil society activists have been documented by Freedom Network, Digital Rights Foundation, and Amnesty International, though attribution

of specific operations to specific state actors remains contested (Amnesty International, 2021; Digital Rights Foundation, 2023).

The tools of state algorithmic weaponization include coordinated harassment campaigns targeting critics designed to make online spaces hostile to opposition voices and to impose psychological costs on dissent as well as the deployment of fake account networks to manufacture the appearance of public support for state positions. PECA 2016 and subsequent amendments have been criticised by press freedom organisations for creating legal mechanisms that can be used to silence critics of the state under the guise of combating disinformation, effectively mobilising the law itself as an instrument of information control in a context where algorithmic amplification and legal suppression operate in tandem (Reporters Without Borders, 2023).

Resistance and Remediation: Towards a Framework for Democratic Reclamation

❖ Regulatory Approaches: Structural Intervention vs. Content Moderation

The dominant policy response to algorithmic weaponization has been content moderation the removal of individual pieces of harmful content and the suspension of accounts that violate platform policies. While content moderation is necessary, it is structurally insufficient as a primary remediation strategy for several reasons. First, it is reactive rather than preventive, addressing manifestations of algorithmic weaponization after they have already achieved amplification effects. Second, it requires platforms to make inherently political judgements about what constitutes harmful disinformation versus legitimate political speech judgements that are susceptible to both error and partisan bias. Third, it leaves intact the underlying algorithmic architecture and data infrastructure that enables weaponization, addressing symptoms rather than causes. Fourth, at scale, it is impossible to implement consistently across the volume of content generated on major platforms (Gillespie, 2022).

More fundamental regulatory interventions are required. The European Union's Digital Services Act (DSA), adopted in 2022 and fully implemented by 2024, represents the most ambitious regulatory framework yet developed for platform governance. The DSA requires very large online platforms to conduct and publish

annual systemic risk assessments, including assessments of algorithmic amplification risks; to provide independent researchers with access to data necessary for auditing algorithmic systems; to allow users to access content ranked chronologically rather than algorithmically; and to submit to external audits of risk mitigation measures. While the DSA applies only within the EU, its extraterritorial effects as platforms modify their global systems to comply with European requirements and its role as a model for other jurisdictions make it a significant landmark in the governance of algorithmic weaponization (European Commission, 2022).

- **Algorithmic Transparency and Auditability**

Algorithmic transparency understood not as the publication of proprietary code but as meaningful disclosure of how recommendation and amplification systems work, what inputs they use, and what effects they produce is a necessary precondition for democratic accountability of platform systems. Diakopoulos (2016) proposed a graduated framework of algorithmic transparency obligations, ranging from disclosure of the existence of algorithmic curation to disclosure of input variables, internal logic, and measurable outputs. Different levels of transparency are appropriate for different stakeholder groups: users deserve to know that content is being selected for them by an algorithm and to have some control over its operation; regulators require access to audit systems' compliance with legal obligations; researchers require data access to independently study algorithmic effects; and the public deserves aggregate transparency about how algorithmic systems shape public discourse.

Independent algorithmic auditing systematic examination of platform systems by researchers, regulators, or civil society organisations has emerged as a practical mechanism for achieving accountability in the absence of voluntary platform transparency. Audit methodologies include sock puppet studies, in which researchers create controlled fake accounts to probe recommendation behaviour; API analysis, examining patterns in publicly available data; and leaked internal document analysis, as exemplified by the Haugen disclosures. The limitations of these approaches their reliance on inference rather than direct observation, their susceptibility to platform countermeasures, and their dependence on continued data access underscore the importance of regulatory mandates for research access of the kind provided by the DSA.

❖ Platform Redesign: From Engagement Maximisation to Civic Value

The most fundamental remediation of algorithmic weaponization would require redesigning the incentive structures of social media platforms moving from engagement maximisation as the primary design objective to the optimisation of civic value, information quality, or user wellbeing. Researchers at the Center for Humane Technology and elsewhere have proposed alternative algorithmic objectives including 'time well spent' metrics that weight user satisfaction rather than time-on-platform, 'bridging-based ranking' that prioritises content that achieves engagement across political divides rather than within filter bubbles, and friction-based interventions that slow the spread of unverified content by introducing deliberative prompts before sharing (Ovadya & Thorburn, 2022; Bail, 2021).

Twitter's 2023 experiment with community notes a crowdsourced fact-checking system applied to viral tweets and Meta's various friction-based interventions (such as prompts encouraging users to read articles before sharing them) represent partial steps toward civic-value optimisation, though they leave the underlying engagement-maximising architecture intact. More fundamental redesign would require platforms to accept reduced engagement metrics and thus reduced advertising revenue an outcome that the current political economy of surveillance capitalism makes unlikely without regulatory compulsion. The case for public interest obligations on dominant communications platforms, analogous to the public interest obligations imposed on broadcast licensees, is strengthened by the scale of the harm that engagement-maximising design has demonstrably enabled.

❖ Media Literacy and Prebunking

Alongside structural and regulatory interventions, investment in media literacy and what researchers call 'prebunking' the proactive inoculation of audiences against manipulation techniques before they are exposed to specific disinformation is an important component of a comprehensive anti-weaponization strategy. Inoculation theory, originally developed in social psychology to explain resistance to attitude change, has been applied to disinformation contexts with promising results. Roozenbeek et al. (2020) demonstrated that brief 'prebunking' interventions exposing participants to weakened forms of manipulation techniques such as emotional appeals, false dichotomies, and ad hominem attacks significantly increased their subsequent ability to identify and resist these techniques in real disinformation content.

The Go Viral! and Bad News games, developed by researchers at the University of Cambridge, operationalise prebunking through interactive simulations in which players take the role of disinformation producers, learning manipulation techniques by practising them in a game environment. Large-scale randomised experiments with these tools have demonstrated significant improvements in participants' ability to identify disinformation, with effects that persist over time. Google has partnered with researchers to deploy prebunking video advertisements on YouTube, reaching millions of users with brief inoculation content. While media literacy cannot substitute for structural interventions the scale of algorithmic weaponization exceeds what individual-level resilience alone can address it is a necessary complement that equips citizens to exercise more critical engagement with the information environments they inhabit.

- **International Coordination and the Governance of Cross-Border Operations**

The international dimension of computational propaganda state-sponsored operations that cross-national borders, commercial influence-for-hire industries that operate across jurisdictions, and platform systems that function globally requires international governance responses that no single national regulatory framework can provide. The Paris Call for Trust and Security in Cyberspace, the EU's Code of Practice on Disinformation, and various bilateral and multilateral initiatives to share intelligence about influence operations represent nascent steps toward international coordination, though they remain voluntary, limited in scope, and largely confined to relationships among Western democracies (European Commission, 2022).

For countries like Pakistan, whose information environment is shaped by both domestic political actors and external state-sponsored operations, participation in international information-integrity coalitions offers potential benefits in terms of shared intelligence, technical assistance, and diplomatic leverage over platform behaviour. The development of regional frameworks potentially through SAARC or bilateral cooperation with other South Asian democracies for sharing evidence of cross-border influence operations and coordinating responses could enhance the capacity of individually vulnerable states to resist algorithmic weaponization. Academic research institutions, including universities with expertise in computational propaganda and NLP applied to South Asian languages, have an important role to play in building the evidentiary base that policy responses require.

Conclusion

The weaponization of social media algorithms represents one of the defining challenges of contemporary democratic governance. The exploitation of engagement-maximising algorithmic systems by state actors, political operatives, commercial disinformation industries, and extremist networks has produced measurable harms to democratic deliberation, political trust, social cohesion, and in the most severe cases, human safety. These harms are not incidental byproducts of otherwise beneficial technologies; they are enabled and amplified by the structural design choices, business model incentives, and data practices of platforms whose commercial success depends on the same attention-capturing mechanisms that manipulators exploit.

A comprehensive response to algorithmic weaponization must be proportionate to the structural nature of the problem. Content moderation, while necessary, is insufficient as a primary strategy: it is reactive, politically fraught, and leaves intact the underlying conditions of weaponization. Structural interventions regulatory frameworks that mandate algorithmic transparency and auditability, restrict the most invasive forms of political micro-targeting, impose public interest obligations on dominant platforms, and provide researchers and regulators with meaningful data access are necessary complements to moderation. These must be accompanied by investment in media literacy and prebunking programmes, support for independent fact-checking and forensic journalism infrastructure, and international coordination to address the cross-border dimensions of computational propaganda.

For developing democracies with asymmetric vulnerabilities including Pakistan, where institutional fragility, a polarised media landscape, and a WhatsApp-centric information ecosystem create conditions of exceptional susceptibility these challenges are particularly acute and the stakes particularly high. The integrity of democratic processes, the safety of journalists and political minorities, and the quality of public deliberation all depend on the capacity of states, civil societies, and academic communities to understand, expose, and resist the weaponization of the algorithmic systems through which increasing majorities of citizens encounter the political world. The moment for complacency has long passed; the moment for sustained, serious, and coordinated action is now.

References

Amnesty International. (2021). Pakistan: Targeting dissent PECA and the suppression of online free expression. Amnesty International. https://www.amnesty.org/en/documents/asa33/4503/2021/en/

Bail, C. A. (2021). Breaking the social media prism: How to make our platforms less polarizing. Princeton University Press.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? Journal of Marketing Research, 49(2), 192–205. https://doi.org/10.1509/jmr.10.0353

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

Bradshaw, S., & Howard, P. N. (2019). The global disinformation order: 2019 global inventory of organised social media manipulation. Oxford Internet Institute. https://demtech.oii.ox.ac.uk/research/posts/the-global-disinformation-order/

Bucher, T. (2018). If...then: Algorithmic power and politics. Oxford University Press.

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

Digital Rights Foundation. (2023). Annual report on digital rights and online freedom in Pakistan. Digital Rights Foundation. https://digitalrightsfoundation.pk/

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2018). The tactics and tropes of the Internet Research Agency. New Knowledge. https://digitalcommons.unl.edu/senatedocs/2/

European Commission. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council on a single market for digital services (Digital Services Act). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065

European External Action Service. (2020). First report on the implementation of the action plan against disinformation. EEAS. https://eeas.europa.eu/topics/disinformation_en

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104. https://doi.org/10.1145/2818717

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), Media technologies: Essays on communication, materiality, and society (pp. 167–194). MIT Press.

Gillespie, T. (2022). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Herman, E. S., & Chomsky, N. (1988). Manufacturing consent: The political economy of the mass media. Pantheon Books.

Horwitz, J. (2021, September 14). Facebook knows Instagram is toxic for teen girls, company documents show. The Wall Street Journal. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

Howard, P. N. (2020). Lie machines: How to save democracy from troll armies, deceitful robots, junk news operations, and political operatives. Yale University Press.

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. American Political Science Review, 111(3), 484–501. https://doi.org/10.1017/S0003055417000144

Mueller, R. S. (2019). Report on the investigation into Russian interference in the 2016 presidential election (Vol. I). U.S. Department of Justice. https://www.justice.gov/archives/sco/file/1373816/dl

Ong, J. C., & Cabanes, J. V. A. (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. Newton Tech4Dev Network. https://doi.org/10.7275/2cq4-5396

Ovadya, A., & Thorburn, L. (2022). Bridging-based ranking: A proposal to reduce societal division while preserving democratic discourse. Berkman Klein Center for Internet & Society. https://cyber.harvard.edu/publication/2022/bridging-based-ranking

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.

Reporters Without Borders. (2023). Pakistan: Press freedom index 2023. RSF. https://rsf.org/en/country/pakistan

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W., Jr. (2020). Auditing radicalization pathways on YouTube. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 131–141. https://doi.org/10.1145/3351095.3372879

Roose, K. (2019, March 29). The making of a YouTube radical. The New York Times. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Roozenbeek, J., Schaewitz, L., Fabian, M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science, 7(10), 201199. https://doi.org/10.1098/rsos.201199

Stanford Internet Observatory. (2020). Influence operations in South Asia: A cross-platform analysis. Stanford Internet Observatory, Stanford University. https://fsi.stanford.edu/research/influence_operations

UN Human Rights Council. (2018). Report of the independent international fact-finding mission on Myanmar (A/HRC/39/64). United Nations. https://www.ohchr.org/en/hr-bodies/hrc/myanmar-ffm/report

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aao2998

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. Internet Policy Review, 10(2). https://doi.org/10.14763/2021.2.1565

Wu, T. (2017). The attention merchants: The epic scramble to get inside our heads. Knopf.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

Article Information:

Received	*7-Jan-2025*
Revised	25-Feb-2026
Accepted	10-Mar-2026
Published	30-Mar-2026

Declarations:

Authors' Contribution:

- [1]**Conceptualization, and intellectual revisions**
- [2-3]**Data collection, interpretation, and drafting of manuscript**
- The authors agree to take responsibility for every facet of the work, making sure that any concerns about its integrity or veracity are thoroughly examined and addressed

- **Conflict of Interest**: NIL

- **Funding Sources**: NIL

Correspondence:

Dr. Taha Shabbir

taha.shabbir@hamdard.edu.pk