BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 1, pages: 591-610
Submitted: December 14th, 2025 | *Accepted for publication:* February 1st, 2026

Algorithmic Deviance and Radicalisation in Digital Platform Societies: Neurocognitive Reinforcement and AI Recommendation Systems

Ionut Virgil Serban
University of Craiova, A. I. Cuza Str., No. 13, Craiova, 200585, Romania; fellow at the University of Chieti-Pescara; the University "Kore", Enna; and the University of International Studies in Rome (Unint), Italy.
ionut.serban@edu.ucv.ro,
johnutzserban@yahoo.com,
https://orcid.org/0000-0001-7240-9989

Bogdan Pătruț
Department of Computer Science, Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania.
bogdan@edusoft.ro
https://orcid.org/0000-0003-1756-6468

Valer Nimineț
Department of Mathematics and Informatics, Faculty of Science, Vasile Alecsandri University of Bacau, Romania.
valern@ub.ro,
https://orcid.org/0009-0002-0098-8066

Abstract: *Artificial intelligence–driven recommender systems increasingly shape how information circulates within digital platforms and how users encounter political and social narratives. As a result, processes of radicalization, extremist mobilization, and digitally mediated deviance can no longer be explained solely by social strain or ideological indoctrination, but must also be understood within algorithmically curated environments designed to maximize user engagement. This research develops an interdisciplinary framework explaining how recommendation algorithms interact with neurocognitive reward mechanisms to reinforce and amplify radicalization pathways. Bringing together criminological theory, digital sociology, and cognitive neuroscience, the study draws on General Strain Theory, Social Learning Theory, and Actor–Network Theory, alongside research on dopaminergic reward systems, emotional salience processing, predictive coding, and neuroplasticity. Within this framework, the article introduces the concept of Algorithmic Strain Environments (ASEs), defined as digitally mediated ecosystems in which engagement-optimized recommendation systems repeatedly amplify grievance narratives, emotional arousal, and identity polarization through recursive feedback loops. To translate these dynamics into measurable signals, the study proposes four analytical indicators: the Extremity Drift Index (EDI), the Engagement Volatility Score (EVS), the Homophily Density Metric (HDM), and the Narrative Convergence Rate (NCR). These indicators are designed not only for retrospective analysis but also for early detection of radicalization trajectories, thereby positioning the model as a predictive rather than purely descriptive framework. A simulation based on a hypothetical dataset illustrates how such indicators can be integrated into a quantitative approach for analyzing algorithmically mediated radicalization dynamics. Finally, the article examines the governance implications of these processes within emerging regulatory frameworks, including the European Union Artificial Intelligence Act, the Digital Services Act, the United Kingdom Online Safety Act, and ongoing regulatory debates in the United States. It proposes a neuro-algorithmic governance framework that integrates algorithmic auditing, cognitive risk modeling, and systemic platform accountability. Overall, the findings suggest that radicalization in platform societies is increasingly shaped through the interaction between human cognitive vulnerabilities and engagement-driven algorithmic infrastructures, highlighting the need for governance approaches capable of addressing both technological design and neurocognitive reinforcement mechanisms.*

Keywords: *artificial intelligence; algorithmic radicalisation; recommendation algorithms; neurocognitive reinforcement; digital deviance; platform societies; echo chambers; general strain theory; social learning theory; algorithmic governance.*

How to cite: Șerban, I. V, Pătruț, B., Nimineț, V. (2026). Algorithmic Deviance and Radicalisation in Digital Platform Societies: Neurocognitive Reinforcement and AI Recommendation Systems. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(1), 591-610. https://doi.org/10.70594/brain/17.1/41

1. Introduction

The rapid expansion of artificial intelligence–driven digital platforms has transformed how information circulates and how social influence operates within contemporary societies. In algorithmically curated environments, recommendation systems shape user exposure by ranking and promoting content predicted to maximise engagement. Although these systems have improved the efficiency of information distribution, they may also create conditions that intensify ideological polarisation and facilitate processes of radicalisation. Existing research on digital extremism has produced valuable insights, yet it often examines technological infrastructures or psychological vulnerability separately. This article argues that radicalisation in digital platform societies is best understood as emerging from the interaction between engagement-optimised algorithmic systems and neurocognitive reinforcement processes that shape how individuals perceive and respond to emotionally salient information.

Classical criminological perspectives have long explained deviant behaviour through mechanisms such as structural strain (Merton, 1938), differential association (Akers, 1973), labelling processes (Becker, 1963), and weakened social bonds (Hirschi, 2002). Digital platforms, however, reshape the informational environments in which these processes unfold. Unlike traditional communication systems, algorithmically curated ecosystems continuously rank, filter, and amplify content according to predicted engagement probability. Because these systems adapt dynamically to user behaviour, they progressively structure exposure patterns over time. Individuals experiencing frustration or perceived marginalisation may therefore encounter increasingly homogeneous streams of emotionally charged narratives that reinforce perceptions of injustice or collective victimisation. In such contexts, recommendation algorithms do more than mediate communication; they help structure the informational environments in which deviant identities can develop.

While classical criminological frameworks remain analytically valuable, they largely assume social environments in which influence occurs through interpersonal interaction or geographically bounded communities. Contemporary digital platforms introduce a different form of social mediation. Algorithmically driven exposure systems can continuously amplify emotionally arousing material at a global scale. Research on online radicalisation suggests that exposure to progressively extreme content is often facilitated by recommender systems designed to optimise engagement (Neumann, 2016; Zuboff, 2019). Yet much of the literature continues to treat technological architectures and psychological vulnerability as separate explanatory domains.

Recent interdisciplinary work has begun to bridge this gap. Studies in cognitive neuroscience show that high-arousal digital stimuli can activate dopaminergic reward pathways and stress-response systems associated with emotional salience and behavioural reinforcement (Schultz, Dayan, & Montague, 1997). At the same time, many contemporary recommendation systems rely on reinforcement-learning architectures that update predictions based on behavioural feedback. Taken together, these developments suggest an alignment between three interacting systems: human reward circuitry, algorithmic reinforcement learning, and engagement-driven monetisation strategies. Within such environments, emotionally salient content generates engagement signals that feed algorithmic optimisation processes, increasing the probability that similar or more extreme material will be recommended in subsequent exposure cycles. Over time, this recursive dynamic may contribute to gradual ideological escalation.

Despite growing research on digital extremism and algorithmic amplification, existing studies rarely integrate criminological deviance theory with neurocognitive reinforcement mechanisms and algorithmic recommendation dynamics within a unified analytical framework.

Against this background, the present study pursues four primary objectives:

- First, it extends classical theories of deviance to algorithmically mediated information environments.
- Second, it integrates neurocognitive reinforcement mechanisms into theoretical explanations of digital radicalisation.

- Third, it develops measurable indicators capable of detecting radicalisation trajectories within algorithmically curated platforms.
- Finally, it examines the implications of these dynamics for emerging governance frameworks regulating artificial intelligence systems.

To address these objectives, the article **proposes an interdisciplinary analytical framework linking criminological theory, digital sociology, and contemporary neuroscience**. Radicalisation is conceptualised as an emergent outcome of recursive interactions between structural strain conditions, engagement-optimised recommendation architectures, and reward-based cognitive reinforcement processes. Building on this framework, the study introduces a set of operational indicators designed to capture algorithmically mediated radicalisation dynamics, including the ***Extremity Drift Index (EDI), the Engagement Volatility Score (EVS), the Homophily Density Metric (HDM), and the Narrative Convergence Rate (NCR)***. A conceptual modelling approach is then developed to illustrate how these indicators may be used to detect escalating exposure trajectories. The analysis also considers governance implications within emerging regulatory regimes, including the European Union Artificial Intelligence Act and the Digital Services Act (European Parliament and Council, 2022; European Parliament and Council, 2024).

Through this interdisciplinary approach, the article contributes to the growing literature on digital deviance by showing that radicalisation in platform societies increasingly emerges from the interaction between human cognitive vulnerabilities and algorithmic amplification infrastructures. This research contributes to the literature on digital radicalisation by developing a neuro-algorithmic framework that integrates criminological deviance theory, cognitive neuroscience, and algorithmic recommendation systems. The study also proposes measurable indicators capable of operationalising radicalisation trajectories within algorithmically mediated environments.

2. Materials and Methods

This study adopts an interdisciplinary conceptual and modelling approach designed to examine how algorithmically mediated environments may contribute to radicalisation dynamics. Rather than relying on proprietary platform datasets—which remain largely inaccessible to independent researchers—the analysis develops a theoretically grounded analytical framework capable of operationalising radicalisation processes within digital platform ecosystems. The indicators and modelling structure proposed in this study are therefore intended as analytical tools for future empirical research and algorithmic auditing, rather than as definitive measurements derived from platform-level data.

The methodological strategy integrates insights from criminology, digital sociology, cognitive neuroscience, and artificial intelligence research. Through this interdisciplinary synthesis, the study constructs an analytical framework for examining how engagement-driven recommendation systems interact with human cognitive reinforcement mechanisms in shaping exposure to increasingly extreme content.

2.1. Theoretical Synthesis

The first stage of the analysis consists of a structured conceptual review of several theoretical traditions that explain deviance, social influence, and technologically mediated interaction.

Particular attention is given to:

• **General Strain Theory** (Agnew, 1992; Agnew, 2006), which links deviant behaviour to persistent frustration and blocked social opportunities.

• **Social Learning Theory** (Akers, 1973; Akers & Jensen, 2006), which explains how deviant norms and behaviours can be acquired through interaction and reinforcement within peer groups.

• **Actor–Network Theory** (Latour, 2005), which conceptualises technological systems as actors that participate in shaping social processes.

• Research in **digital sociology** examining platform ecosystems and algorithmic mediation.
• Studies of **AI recommender systems**, particularly engagement-based ranking architectures.
Synthesising these perspectives makes it possible to identify how algorithmic environments reshape traditional pathways through which deviant identities and beliefs may emerge.

2.2. Neurocognitive Integration

The second stage incorporates findings from cognitive neuroscience and behavioural psychology, in order to examine how digital stimuli interact with human learning and emotional processing systems.

The analysis draws on research addressing:
• dopaminergic reward circuitry and reinforcement learning mechanisms
• emotional salience processing and amygdala activation
• amplification of cognitive biases in digitally mediated environments
• predictive coding models of belief formation and belief rigidity
• neuroplastic adaptation associated with repeated exposure to emotionally salient stimuli

Integrating these findings allows established theories of deviance to be interpreted alongside contemporary insights into how individuals process reward, emotion, and social feedback in digitally mediated environments.

2.3. Conceptual Modelling Framework

The final stage develops a conceptual modelling framework that translates these theoretical mechanisms into measurable indicators of algorithmic radicalisation dynamics. Rather than focusing solely on ideological content, the framework identifies behavioural and structural signals that may indicate escalating exposure trajectories within algorithmically curated environments.

Examples of such signals include:
• acceleration in user engagement patterns
• progressive shifts towards more extreme content exposure
• increasing network homophily within interaction clusters
• volatility in emotional valence across engagement patterns
• convergence between user-generated language and grievance-based narratives

Although the model does not rely on proprietary platform data, it proposes a replicable analytical structure that can be applied in future empirical studies and algorithmic auditing initiatives. Because large-scale platform datasets remain largely inaccessible to independent researchers, conceptual modelling provides an important step towards developing measurable indicators that can later be tested empirically.

3. Theoretical Framework

3.1. Algorithmic Strain Environments

General Strain Theory suggests that deviance may emerge when individuals experience persistent frustration, blocked opportunities, or negative social relationships (Agnew, 1992, 2006). In digitally mediated environments, these forms of strain can take on new characteristics. Social media platforms expose users to constant comparisons, narratives of perceived injustice, and emotionally charged interpretations of social events. Such dynamics are consistent with Bauman's (2000) notion of "liquid modernity," in which individuals experience heightened insecurity, fragmented identities, and continuous exposure to shifting social benchmarks.

Digital platform architectures do not simply host these narratives. Recommendation systems rank and prioritise content based on predicted engagement, which often favours emotionally intense material. As a result, users who already feel marginalised or frustrated may encounter streams of information that repeatedly reinforce grievance-based interpretations of social reality.

Such exposure environments can gradually narrow the range of perspectives users encounter. Over time, repeated encounters with similar narratives may strengthen feelings of

resentment and reinforce interpretive frameworks that frame social events through conflict or injustice.

This study refers to such digitally mediated exposure environments as **Algorithmic Strain Environments (ASEs)**.

*An **Algorithmic Strain Environment** can be defined as a digital ecosystem in which algorithmic ranking systems systematically amplify emotionally salient grievance narratives. Within such environments, repeated reinforcement of these narratives may increase the probability that frustration becomes integrated into deviant identity formation.*

Because engagement signals guide algorithmic optimisation, emotionally charged content is more likely to circulate widely and persist in recommendation cycles. This dynamic may create feedback loops in which grievance narratives receive increasing visibility, potentially accelerating ideological escalation.

3.2. Neurocognitive Reinforcement Loops

Interaction with digital platforms also engages neural systems involved in reward processing and learning. Stimuli associated with novelty, social approval, or emotional intensity can activate dopaminergic reward pathways that reinforce attention and memory (Schultz, Dayan, & Montague, 1997).

Emotionally charged content—particularly narratives involving anger, outrage, or perceived injustice—tends to attract stronger attention than neutral information. These responses make such content especially effective at generating behavioural engagement.

At the same time, modern recommendation systems operate through optimisation processes similar to reinforcement learning. Within these systems:

- user engagement functions as a behavioural signal,
- algorithms update ranking predictions based on interaction patterns,
- and exposure patterns are continually adjusted to maximise future engagement.

When users interact with emotionally salient narratives and receive social validation through likes, shares, or comments, biological reward mechanisms and algorithmic reinforcement processes begin to align. Engagement signals are interpreted by recommendation systems as indicators of relevance, increasing the likelihood that similar content will be recommended in future exposure cycles.

Through repeated iterations, this interaction can create a recursive exposure dynamic in which emotionally charged narratives become progressively more prominent. Over time, such reinforcement cycles may strengthen ideological alignment and stabilise grievance-oriented interpretive frameworks.

The interaction between user engagement, neurocognitive reward responses, and algorithmic amplification is illustrated in Figure 1.

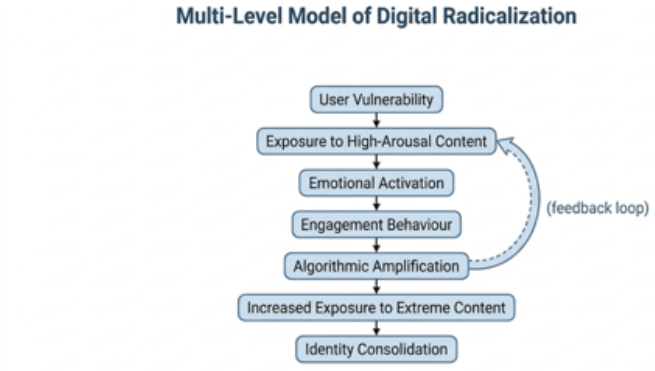

Figure 1. Algorithmic Reinforcement Cycle in the Multi-level Model of Digital Radicalisation
Conceptual representation of the interaction between user engagement behaviour and algorithmic recommendation systems. Engagement signals function as feedback inputs that influence subsequent content ranking, increasing the probability that emotionally salient narratives reappear in future exposure cycles.

As illustrated in Figure 1, radicalisation trajectories may develop through a recursive reinforcement process. Exposure to emotionally salient content can trigger engagement behaviours such as viewing, liking, or sharing. Recommender systems interpret these behaviours as signals of relevance, increasing the likelihood that similar content will appear in subsequent recommendation cycles. Over time, this feedback dynamic may intensify exposure to grievance-oriented narratives and strengthen ideological alignment.

3.3. Social Learning in Digital Echo Systems

Social Learning Theory argues that deviant behaviour may be acquired through interaction with peers who provide definitions favourable to rule violation and reinforce such interpretations (Akers, 1973; Akers & Jensen, 2006; Bandura, 1991). Digital platforms can facilitate similar processes by clustering users with comparable engagement patterns and ideological preferences.

Algorithmic recommendation systems often connect individuals with others who consume and interact with similar content. Over time, these clusters may develop into highly homogeneous networks characterised by strong ideological alignment.

Within such environments, users repeatedly encounter interpretations of events that reinforce shared grievances while alternative viewpoints become less visible. Social reinforcement mechanisms—including approval signals, reposting behaviour, and supportive commentary—may normalise increasingly radical interpretations of social reality.

As these interactions accumulate, moral disengagement and dehumanising representations of out-groups may become socially acceptable within particular network clusters. These dynamics suggest that algorithmically structured communities can accelerate the social learning processes traditionally associated with deviant group formation.

The result may be a self-reinforcing interpretive ecosystem in which shared narratives are continuously validated while competing interpretations gradually disappear from the information environment.

3.4. Actor–Network Reinterpretation

Actor–Network Theory expands the concept of agency beyond human actors to include technological systems and institutional infrastructures (Latour, 2005). From this perspective, algorithmic recommendation systems function as non-human actors that actively participate in shaping informational flows.

The architecture of recommender systems, interface design choices, and platform monetisation strategies all influence which narratives become visible and how widely they circulate. Radicalisation processes therefore cannot be attributed solely to individual psychological predispositions. Instead, they emerge from networks of interacting human and technological actors.

Within digital platform ecosystems, these networks include:

- human users
- recommender algorithms
- automated amplification mechanisms
- monetisation architectures
- interface design structures

Together, these elements contribute to shaping the informational environment within which ideological identities develop.

The interaction between human vulnerabilities and algorithmic amplification mechanisms can be conceptualised across multiple analytical levels, as summarised in Table 1.

Table 1. Multi-Level Radicalisation Model

Level	**Human Factor**	**AI Factor**	**Outcome**
Psychological	Frustration	Engagement ranking	Polarisation
Cognitive	Moral disengagement	Content prioritisation	Dehumanisation
Social	Peer validation	Echo clustering	Identity radicalisation
Economic	Attention incentives	Monetisation logic	Outrage amplification
Political	Grievance narratives	Algorithmic visibility	Extremist mobilisation

By integrating General Strain Theory, Social Learning Theory, Actor–Network Theory, and insights from cognitive neuroscience, a multi-level framework emerges for understanding radicalisation within algorithmically mediated environments. Figure 2 presents the conceptual model proposed in this study.

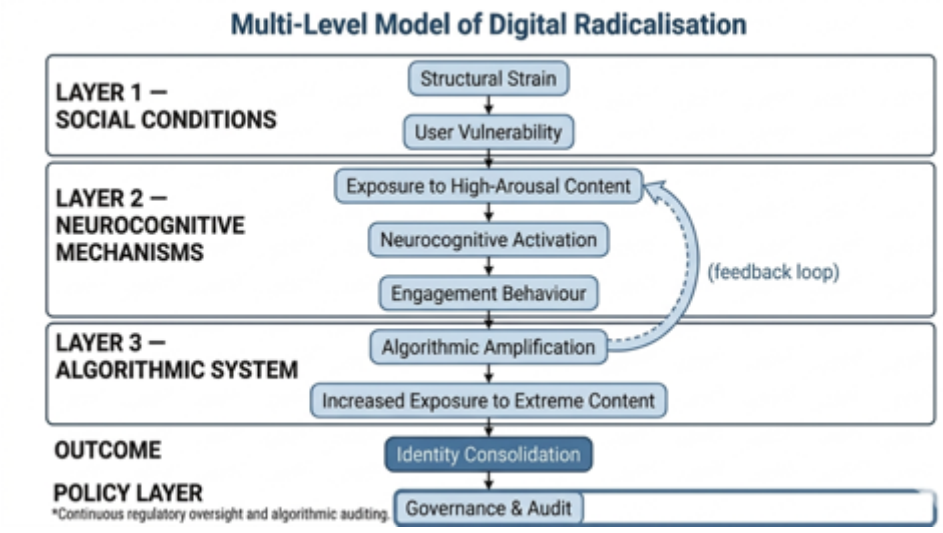

Figure 2. Multi-Level Model of Digital Radicalisation.
Conceptual framework illustrating how structural social conditions, neurocognitive reinforcement mechanisms, and engagement-driven recommendation systems interact within digital platforms. Exposure to emotionally salient content can trigger engagement behaviours that generate feedback signals for algorithmic ranking systems, increasing the likelihood of similar content appearing in subsequent exposure cycles. Governance and auditing mechanisms represent potential intervention points designed to mitigate systemic amplification dynamics.

Figure 2 illustrates the multi-layer structure of the proposed model. Structural social conditions create vulnerability contexts that increase the likelihood of exposure to emotionally salient narratives. Neurocognitive reinforcement processes transform such exposure into engagement behaviours that generate feedback signals for recommendation systems. These signals guide algorithmic amplification mechanisms, increasing the visibility of similar narratives. Over time, this interaction may contribute to identity consolidation and ideological escalation. Governance and auditing mechanisms represent potential intervention points designed to mitigate these amplification dynamics.

Within this framework, radicalisation does not appear solely as an individual psychological process. Rather, it emerges as a distributed outcome produced through the interaction of human cognitive responses, algorithmic optimisation processes, and networked social environments.

Taken together, these mechanisms suggest that algorithmic radicalisation involves not only social interaction dynamics but also underlying neurocognitive processes. The following section therefore examines the neural systems that may contribute to these reinforcement dynamics.

4. Neuroscientific Mechanisms

4.1. Neurocognitive Mechanisms of Algorithmically Reinforced Radicalisation

Radicalisation within algorithmically mediated environments can be interpreted not only as a socio-structural process but also as a form of neurocognitive adaptation. The mechanisms described in the previous section—algorithmic amplification, social reinforcement, and grievance validation—interact with neural systems involved in reward learning, emotional salience, and belief formation.

Repeated exposure to emotionally charged digital content can influence attention allocation, memory consolidation, and behavioural persistence. When engagement-driven recommendation systems systematically prioritise high-arousal narratives, the interaction between algorithmic amplification and neurocognitive reinforcement mechanisms may gradually strengthen ideological commitment.

Digital engagement patterns also resemble intermittent reinforcement schedules, a well-known mechanism in behavioural psychology associated with persistent conditioning. Notifications, likes, shares, and comment feedback are typically delivered unpredictably, generating variable reward expectations that intensify dopaminergic signalling in reward-processing circuits. Over time, repeated interaction with emotionally salient narratives may strengthen neural associations between grievance-based interpretations and rewarding social feedback.

The following subsections outline several neural processes that may contribute to these reinforcement dynamics.

4.2. Dopaminergic Reward Circuitry and Engagement Optimisation

The mesolimbic dopamine system—particularly the ventral tegmental area (VTA) and the nucleus accumbens—plays a central role in reinforcement learning and motivational salience. Dopamine release increases when individuals encounter novel, emotionally engaging, or socially validating stimuli.

Recent neuroimaging studies suggest that unpredictable social rewards generate stronger dopaminergic responses than predictable reinforcement (Schultz, Dayan, & Montague, 1997). Digital platforms frequently employ variable reward structures through notifications, likes, shares, and comment feedback. These mechanisms closely resemble intermittent reinforcement paradigms known to sustain behavioural engagement.

When users encounter emotionally charged grievance narratives and receive social validation through engagement signals, dopaminergic reinforcement may strengthen the neural encoding of those narratives. High-arousal content may engage emotional and attentional systems that facilitate learning processes and memory consolidation, which in turn may contribute to stronger long-term potentiation (LTP) and the reinforcement of stimulus–response associations linking emotional narratives with rewarding social feedback.

4.3. Emotional Salience and Amygdala Activation

The amygdala plays a crucial role in threat detection and emotional salience processing. Stimuli associated with anger, moral outrage, or perceived injustice often trigger heightened amygdala activation, increasing attentional capture and strengthening memory encoding.

Because engagement-driven algorithms prioritise content that generates strong reactions, users may be disproportionately exposed to narratives that activate threat perception and defensive cognition. Repeated activation of these neural pathways may heighten sensitivity to perceived injustice and reinforce distinctions between in-groups and out-groups.

Some research suggests that sustained exposure to emotionally provocative content may engage neural systems associated with emotional salience, which may in turn influence the regulatory balance between the amygdala and the prefrontal cortex and affect inhibitory control processes (Ochsner & Gross, 2005; Arnsten, 2009).

4.4. Stress Neurobiology and Chronic Strain

General Strain Theory proposes that persistent frustration can generate negative emotional states that increase the likelihood of deviant behaviour (Agnew, 1992). From a neurobiological perspective, chronic exposure to grievance-oriented narratives may activate stress-response systems involving cortisol release and sympathetic nervous system arousal.

Sustained stress exposure has been associated with reduced prefrontal inhibitory control, increased impulsivity, and heightened reactive aggression. If algorithmic exposure patterns repeatedly activate stress pathways, individuals may experience diminished capacity for reflective evaluation of emotionally charged narratives.

Under such conditions, grievance-based interpretations of events may be processed through reactive rather than deliberative cognitive pathways, increasing susceptibility to radical ideological framing.

4.5. Predictive Coding and Cognitive Confirmation Loop

Contemporary neuroscience often conceptualises perception through predictive coding frameworks, in which the brain continuously generates predictions about the environment and updates them when confronted with error signals (Friston, 2010; Friston et al., 2017). Exposure to contradictory information normally produces prediction errors that facilitate belief revision (Clark & Hohwy, 2024).

Digital echo chambers may reduce exposure to such corrective signals. When users encounter predominantly confirming narratives within ideologically homogeneous networks, prediction errors become less frequent and belief updating slows. This dynamic can lead to increased cognitive rigidity and stronger confidence in existing interpretations.

Algorithmic clustering based on engagement patterns may therefore reinforce existing predictive priors and contribute to ideological entrenchment.

4.6. Neuroplasticity and Identity Consolidation

Neural pathways strengthen through repeated activation, a process known as experience-dependent neuroplasticity. Identity formation is therefore not fixed but continuously shaped by patterns of exposure and reinforcement.

Within algorithmically curated environments, repeated engagement with emotionally salient narratives—combined with social validation and reward signalling—may strengthen neural networks associated with particular ideological schemas. Over time, these reinforcement patterns can increase the automaticity with which individuals interpret events through those schemas.

From this perspective, radicalisation in digital environments may develop gradually through repeated reinforcement and neural adaptation rather than through sudden ideological conversion.

These neurocognitive mechanisms suggest that radicalisation trajectories may emerge from the interaction between biological reinforcement processes and algorithmically structured exposure environments. The following section examines how these mechanisms become observable within platform-level amplification dynamics.

5. Algorithmic Amplification Dynamics

The theoretical and neuroscientific mechanisms discussed in the preceding sections suggest that digital platforms do not merely host information but can structure exposure in ways that gradually reinforce particular interpretations of social reality. Within algorithmically mediated environments, amplification processes, neurocognitive reinforcement, and network clustering interact to produce feedback dynamics capable of shaping radicalisation trajectories. Rather than operating independently, these mechanisms form interconnected loops. Engagement behaviour influences algorithmic ranking decisions, which in turn affect future exposure patterns. Over time,

this interaction may progressively intensify the visibility of emotionally charged narratives and reduce exposure to competing perspectives.

5.1. The Architecture of Amplification

AI-driven recommender systems organise digital information visibility by predicting which content is most likely to generate engagement. These systems rely on behavioural signals—including click-through rates, viewing duration, sharing activity, and comment patterns—to refine ranking algorithms.

Emotionally charged content often produces stronger engagement responses than neutral material. As a result, recommendation systems may disproportionately prioritise narratives that evoke anger, outrage, or perceived injustice. When such content repeatedly receives high engagement signals, it becomes more likely to appear in subsequent recommendation cycles.

Three amplification mechanisms are particularly relevant.

First, engagement escalation occurs when users who interact with emotionally salient material begin receiving recommendations containing increasingly intense or polarised narratives. Second, controversy bias emerges because content that provokes strong emotional reactions tends to generate higher interaction rates and therefore receives greater algorithmic visibility.

Third, reinforcement filtering gradually reduces informational diversity by clustering users with similar behavioural profiles and engagement histories.

Taken together, these mechanisms can narrow the range of perspectives visible to users while simultaneously amplifying grievance-oriented narratives.

5.2. Neurocognitive Reinforcement Dynamics

The reinforcement loop described in Section 3 becomes observable through engagement patterns within algorithmically mediated environments. When emotionally salient content activates reward-processing systems, users respond through interactions such as clicking, sharing, or commenting. These behaviours function as feedback signals that recommendation systems interpret as indicators of relevance.

Through repeated iterations, engagement signals reinforce algorithmic predictions. Content that successfully captures attention becomes more likely to reappear in future recommendation cycles. In this way, behavioural responses generated by neurocognitive reward mechanisms contribute directly to algorithmic amplification processes.

Over time, such interactions may produce exposure trajectories characterised by increasing emotional intensity, reduced informational diversity, and progressive ideological alignment.

5.3. Echo-System Social Learning

Digital platforms also facilitate social learning processes that reinforce ideological alignment within networked communities. Social Learning Theory suggests that deviant behaviour develops through interaction with peers who provide definitions favourable to rule violation and reinforce such interpretations.

Algorithmically mediated networks frequently cluster users with similar engagement patterns. These clusters may evolve into dense communities in which shared narratives are repeatedly reinforced while dissenting perspectives become less visible.

Within such environments, three structural characteristics frequently appear:

1. **Network homophily**, in which users interact primarily with others who share similar beliefs or behavioural patterns.
2. **Narrative convergence**, where community members increasingly adopt similar interpretations of events.
3. **Collective grievance validation**, in which shared emotional reactions reinforce group identity and ideological cohesion.

Through these processes, individual exposure patterns can develop into socially reinforced interpretive frameworks capable of stabilising radical narratives.

5.4. Observable Radicalisation Trajectories

When algorithmic amplification, neurocognitive reinforcement, and network clustering interact, identifiable exposure trajectories may emerge. These trajectories often involve gradual escalation in which users encounter increasingly extreme narratives over time.

Several observable signals may indicate such developments. Content exposure may display extremity drift, reflecting a gradual shift towards more polarised material. Engagement behaviour may show volatility spikes, particularly during interactions with emotionally charged narratives. Linguistic analysis of user-generated content may reveal narrative convergence, as language begins to reflect grievance-oriented discourse patterns.

Together, these signals provide a basis for operationalising algorithmic radicalisation processes. The indicators introduced in the following section translate these reinforcement dynamics into measurable variables that can be incorporated into quantitative models of radicalisation trajectories.

6. Operationalisation of Radicalisation Indicators

The mechanisms described in the previous sections can be translated into measurable indicators that capture different dimensions of algorithmically mediated radicalisation processes. These indicators function as empirical proxies for the theoretical dynamics discussed earlier and provide the foundation for the quantitative modelling framework developed in the following section.

Rather than attempting to measure radicalisation directly, the framework identifies behavioural and structural signals that may indicate escalating exposure trajectories within digital platforms.

6.1. Extremity Drift Index (EDI)

The **Extremity Drift Index (EDI)** reflects changes in the ideological intensity of content exposure over time. Instead of focusing on individual posts, the index captures the overall direction of a user's information environment.

Rising EDI values may that recommended content is gradually shifting towards more polarised or grievance-oriented narratives. Such shifts may reflect progressive exposure to increasingly extreme material within algorithmically curated feeds.

6.2. Engagement Volatility Score (EVS)

The **Engagement Volatility Score (EVS)** captures fluctuations in user interaction patterns across time. Rather than measuring engagement volume alone, the metric focuses on variability in interaction behaviour.

Sudden spikes in engagement may signal emotionally driven responses to highly salient narratives. In algorithmically mediated environments, these reactions can serve as feedback signals that strengthen recommendation patterns associated with emotionally charged content.

6.3. Homophily Density Metric (HDM)

While the EDI and EVS capture exposure and engagement dynamics, the **Homophily Density Metric (HDM)** examines the structure of the user's interaction network.

This indicator measures the degree of ideological clustering within a user's social environment. High HDM values suggest that users are embedded in dense networks of actors who share similar beliefs, narratives, or engagement patterns.

Network structures play an important role in reinforcing ideological alignment. Users often interact primarily with others who consume similar content, which may reduce exposure to

alternative viewpoints. Figure 3 illustrates the clustering dynamics associated with network homophily.

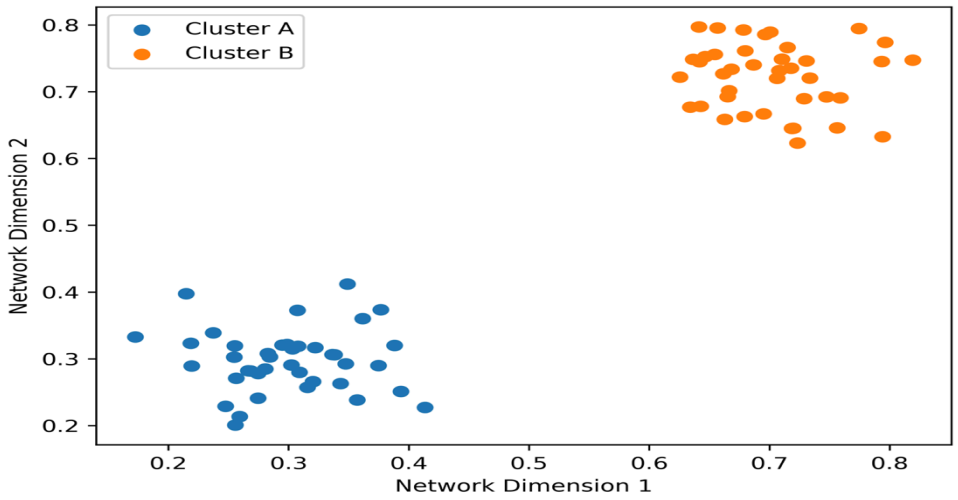

Figure 3. Network Homophily Density Clustering.
Illustrative representation of ideological clustering within digital interaction networks. High homophily density indicates the presence of echo-system structures in which users are primarily exposed to ideologically similar actors.

As shown in Figure 3, interaction networks may gradually evolve into densely connected clusters characterised by ideological similarity. Such homophily structures can limit exposure to divergent perspectives while reinforcing shared narratives, thereby amplifying the effects of algorithmic recommendation systems.

6.4. Narrative Convergence Rate (NCR)

The **Narrative Convergence Rate (NCR)** reflects the linguistic similarity between user-generated content and established grievance-based or extremist narratives.

This indicator captures the extent to which users begin adopting specific frames, terminology, or interpretive patterns associated with particular ideological communities. Rising NCR values therefore suggest increasing alignment between individual discourse and broader narrative structures circulating within digital networks.

Indicator–Model Variable Mapping

The operational indicators introduced above correspond to the theoretical variables used in the quantitative modelling framework presented in the following section.

Specifically:

• **Extremity Drift Index (EDI)** captures extremity exposure dynamics (E_t = extremity exposure at time *t*).
• **Homophily Density Metric (HDM)** represents network clustering structures (H_t = homophily density at time *t*).
• **Engagement Volatility Score (EVS)** approximates reinforcement intensity (R_t) and stress reactivity (S_t).
• **Narrative Convergence Rate (NCR)** captures linguistic alignment with grievance narratives and contributes to extremity exposure dynamics.

Through this mapping, the indicators translate theoretical mechanisms into conceptually measurable variables that can be incorporated into models of radicalisation trajectories.

6.5. Simulated Dataset Illustration

To illustrate the feasibility of empirical operationalisation, a simulated dataset representing user interaction trajectories within algorithmically curated environments is introduced.

The dataset models hypothetical users interacting with recommended content over a twelve-month observation period. Variables include indicators of exposure intensity, engagement frequency, emotional valence of consumed content, network clustering density, and linguistic convergence with grievance-based narratives.

Although the dataset is synthetic, it allows the conceptual framework proposed in this study to be illustratively translated into measurable indicators. The simulation illustrates how algorithmically mediated radicalisation trajectories could be analysed in future empirical platform research.

7. Quantitative Modelling of Algorithmic Radicalisation

To translate the theoretical framework into an analytically tractable form, I have developed a simplified modelling approach capturing how algorithmic exposure, neurocognitive reinforcement, and network structure interact over time. Rather than aiming for precise prediction, the objective is to illustrate how radicalisation trajectories can be formally conceptualised as dynamic and cumulative processes.

7.1. Radicalisation Probability Model

To represent these interactions, radicalisation probability can be conceptually modelled using a logistic specification. Such models are commonly used to estimate the likelihood of transitions between states—in this case, from non-radicalised to radicalised exposure trajectories.

Let the following variables describe the exposure trajectory of user *u* at time *t*:

- $E_{u,t}$ = extremity level of content exposure of user *u* at time *t*
- $H_{u,t}$ = network homophily density of user *u* at time *t*
- $R_{u,t}$ = reinforcement intensity measured through engagement frequency of user *u* at time *t*
- $V_{u,t}$ = emotional valence intensity of consumed content of user *u* at time *t*
- $S_{u,t}$ = stress reactivity proxy derived from engagement volatility of user *u* at time *t*

The probability of radicalisation (Rad) of user *u* at time *t* can be expressed using the logistic function (1):

$$P(Rad_{u,t}) = \sigma(\beta_0 + \beta_1 E_{u,t} + \beta_2 H_{u,t} + \beta_3 R_{u,t} + \beta_4 V_{u,t} + \beta_5 S_{u,t} + \gamma P(Rad_{u,t-1})) \quad (1)$$

where

$$\sigma(x) = 1 / (1 + e^{-x}) \quad (2)$$

represents the logistic transformation and β_i represent estimated model parameters indicating the influence of each variable, and γ the autoregressive effect of prior radicalisation probability. Our model suggests that the radicalisation is a cumulative process. The inclusion of the lagged term $P(Rad_{u,t-1})$ introduces temporal dependence, allowing the model to capture radicalisation as a path-dependent and cumulative process. This formulation shows that radicalisation is not an isolated event but a path-dependent process, in which prior exposure and reinforcement increase the likelihood of continued escalation.

7.2. Extremity Drift Dynamics

Radicalisation trajectories rarely evolve in a linear manner. Instead, they tend to display nonlinear escalation patterns shaped by feedback loops between user behaviour and algorithmic recommendation systems.

Let E(t) represent the ideological intensity of content exposure over time. Low values of E(t) correspond to exposure to mainstream or neutral content, while higher values indicate increasing interaction with polarised or extremist narratives.

The ***Extremity Drift Index*** measures the rate at which exposure intensity changes over time:

$$EDI(t) = \frac{dE(t)}{dt} \quad (3)$$

Acceleration of Extremity Drift represents the rate at which the drift itself intensifies (the "acceleration" towards the ideological fringe) can be expressed as (4):

$$EDIacc(t) = \frac{d^2E(t)}{dt^2} \quad (4)$$

Interpretation of Risk

- $EDIacc(t) = 0$ represents linear exposure: the user is consuming extreme content at a steady, predictable rate.
- $EDIacc(t) > 0$ represents accelerating exposure**:** this indicates a compounding "echo chamber" effect, where the user is being drawn into increasingly extreme material at an escalating pace, signifying a high radicalisation risk.
- $EDIacc(t) < 0$ represents de-escalation or moderation of content consumption.

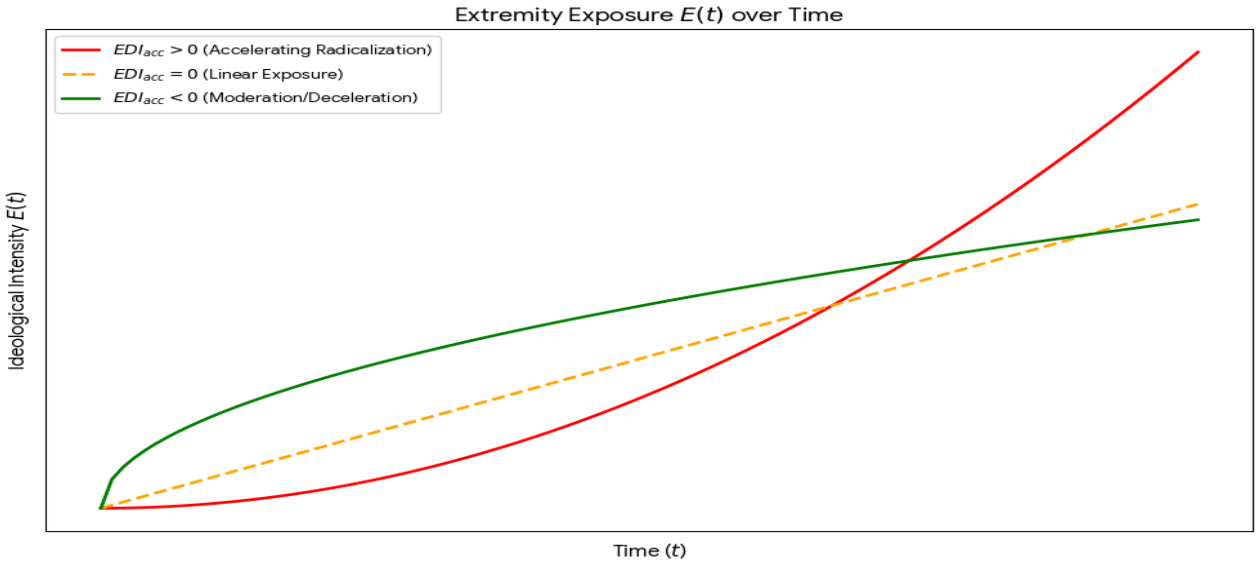

Figure 4. Extremity Exposure Trajectories

Figure 4 illustrates three possible trajectories of ideological exposure over time as defined by extremity drift acceleration.

1. The "Rabbit Hole" ($EDI_{acc} > 0$)

- **active radicalisation**. The user isn't just consuming extreme content; they are consuming it at an ever-increasing rate. Each interaction may lower the threshold for the next, more extreme piece of content, creating a self-reinforcing feedback loop typical of algorithmic amplification.

2. The Linear Path ($EDI_{acc} = 0$)

- **steady state exposure**. The intensity of the content increases at a constant, predictable rate. While the user is moving towards the fringe, the process lacks the explosive "spiral" effect of the first category.

3. The Moderation Path ($EDI_{acc} < 0$)

- **saturation or de-escalation**. Although the user may still be exposed to some extreme ideas, the momentum is decreasing. The rate of change is slowing down, which may indicate

reduced engagement with extreme content, potential disengagement from reinforcing dynamics , or the influence of moderating factors.

These trajectories suggest that radicalisation is not solely determined by the level of exposure, but by the rate at which exposure intensifies over time.

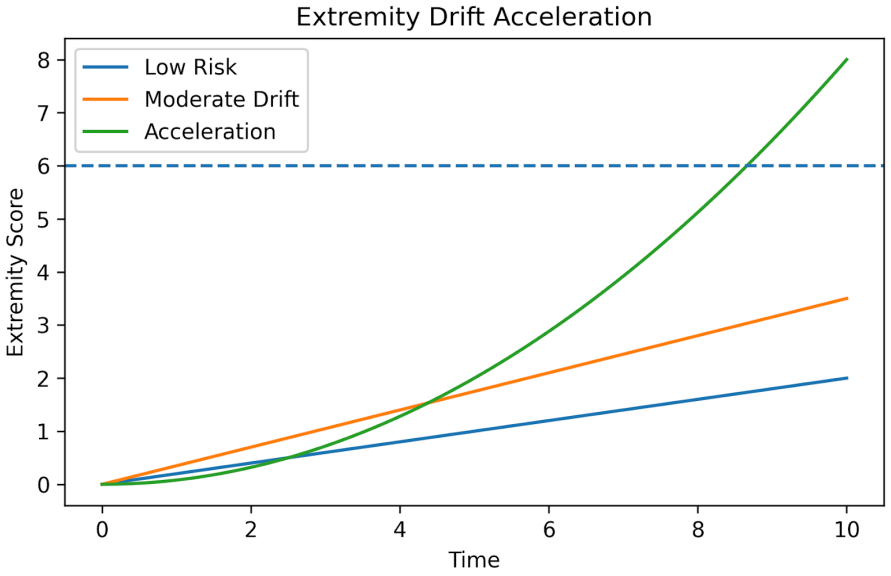

Figure 5. Extremity Drift Acceleration.
Illustrative representation of ideological extremity progression over time. Accelerating trajectories indicate increasing risk of radicalisation as exposure to extreme narratives intensifies.

Figure 5 further illustrates how different exposure trajectories correspond to varying levels of radicalisation risk. Users experiencing accelerating drift are more likely to encounter compounding reinforcement effects, particularly when high engagement intensity and dense network homophily are present. This distinction introduces a dynamic perspective in which early detection of positive extremity drift acceleration may serve as a predictive signal of future radicalisation trajectories.

7.3. From Modeling to Empirical Application

While the present framework remains conceptual, it outlines how radicalisation processes may be conceptually operationalised using analytically defined indicators of exposure dynamics, engagement patterns, and network structure. More advanced approaches—such as network centrality analysis, time-to-event modelling, or simulation-based methods—could extend this framework in future research. However, the primary contribution of this study lies in establishing a theoretical and operational foundation for identifying and monitoring algorithmically mediated radicalisation dynamics.

8. Governance and Regulatory Implications

The interaction between engagement-optimised algorithmic architectures and human neurocognitive vulnerabilities raises significant governance challenges for contemporary regulatory frameworks. If algorithmically mediated environments can amplify emotionally salient narratives and reinforce ideological drift, regulatory strategies should address not only harmful content but also the structural mechanisms that shape information exposure. The identification of extremity drift acceleration as an early warning signal suggests that regulatory frameworks should move beyond static content moderation towards dynamic monitoring of exposure trajectories, potentially enabling intervention before radicalisation processes fully consolidate.

8.1. European Union: AI Act and Systemic Risk Regulation

The European Union has adopted a risk-based regulatory approach through the **Artificial Intelligence Act** (European Parliament & Council, 2024) and the **Digital Services Act** (European Parliament & Council, 2022). The AI Act establishes obligations for high-risk AI systems and introduces governance requirements for general-purpose AI models. Although recommender systems are not automatically classified as high-risk technologies, large online platforms deploying such systems may still fall under systemic risk assessment obligations.

The Digital Services Act requires Very Large Online Platforms to evaluate and mitigate risks related to the dissemination of harmful content, manipulation of services, and impacts on democratic processes. Within this regulatory context, algorithmic amplification of grievance narratives may intersect with risks concerning psychological harm, polarisation, and the integrity of civic discourse.

Indicators such as the **Extremity Drift Index (EDI)**, **Engagement Volatility Score (EVS)**, **Homophily Density Metric (HDM)**, and **Narrative Convergence Rate (NCR)** could potentially serve as indicators-based signals within systemic risk assessments, enabling regulators and auditors to identify patterns of escalating exposure within algorithmically curated environments.

8.2. United Kingdom: Online Safety Act

The **UK Online Safety Act** introduces a platform duty of care requiring digital services to reduce the risk of harm to users. The framework focuses primarily on content moderation, illegal content removal, and user protection obligations.

However, the current regulatory emphasis remains largely centred on moderating individual pieces of content rather than addressing the underlying architecture of recommender systems. As a result, transparency requirements related to algorithmic ranking mechanisms and amplification dynamics remain comparatively limited.

8.3. United States: Fragmented Governance Landscape

In contrast to the European approach, the United States currently lacks comprehensive federal legislation specifically regulating AI-driven recommender systems. Governance is primarily shaped by **Section 230 of the Communications Decency Act**, oversight by the **Federal Trade Commission**, and emerging state-level initiatives addressing algorithmic accountability.

This regulatory landscape places strong emphasis on innovation and free-speech protections. However, it also provides fewer mechanisms for systematic monitoring of algorithmic amplification risks within large digital platforms.

8.4. Towards a Neuro-Algorithmic Governance Framework

To address the structural drivers of algorithmically mediated radicalisation, this study proposes a policy-integrated neuro-algorithmic governance model. The framework combines insights from behavioural science, platform regulation, and algorithmic auditing in order to support more effective systemic oversight.

The proposed governance structure includes four interrelated layers.

Layer 1: Algorithmic Design Constraints

Platform design requirements aimed at reducing the amplification of highly emotionally charged narratives.

Examples include:

- mechanisms for increasing exposure diversity within recommendation systems
- thresholds limiting excessive amplification of high-arousal content

Layer 2: Continuous Exposure Monitoring

Monitoring systems capable of detecting early indicators of escalating exposure dynamics.

Examples include:

- Extremity Drift Index tracking
- engagement volatility anomaly detection

Layer 3: Independent Audit Infrastructure
External auditing mechanisms designed to evaluate systemic platform risks.
These mechanisms may include:

- standardised transparency reporting
- independent algorithmic auditing procedures

Layer 4: Neurocognitive Risk Assessment
Analytical frameworks capable of evaluating the psychological and behavioural impacts of algorithmically mediated information environments.
Examples include:

- stress-amplification analysis
- emotional salience impact assessment

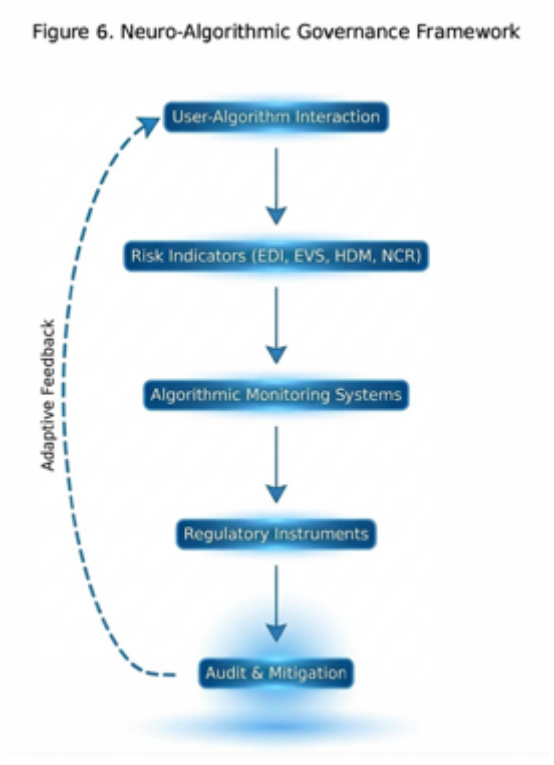

Figure 6. Neuro-Algorithmic Governance Framework
Conceptual framework illustrating a multi-layer governance architecture designed to mitigate algorithmic amplification risks in digital platforms. The model integrates algorithmic design constraints, continuous monitoring of exposure dynamics, independent auditing mechanisms, and neurocognitive risk assessment processes.

As illustrated in Figure 6, effective governance requires intervention across multiple layers of the digital platform ecosystem. Platform design constraints may help reduce the amplification of highly emotionally charged content by introducing diversity-enhancing exposure mechanisms and limiting excessive engagement-driven ranking dynamics. Audit and mitigation processes feed back into platform design and algorithmic systems, enabling adaptive intervention and continuous reduction of systemic amplification risks.

Continuous monitoring systems can identify early signals of radicalisation trajectories through indicators such as extremity drift and engagement volatility. Independent auditing structures provide external oversight capable of evaluating systemic platform risks, while neurocognitive risk modelling enables regulators to assess the broader psychological effects of algorithmically mediated information environments.

Taken together, these mechanisms form an integrated governance architecture designed to address the structural amplification processes that may contribute to algorithmically mediated radicalisation.

9. Discussion

The present study develops a neuro-algorithmic framework for explaining radicalisation in platform societies as the result of interactions between engagement-optimised algorithmic architectures and human neurocognitive reinforcement mechanisms. By integrating criminological theory, digital sociology, and cognitive neuroscience, the analysis illustrates how algorithmic amplification processes may transform episodic exposure to grievance narratives into progressively reinforced ideological trajectories.

A key implication of this framework concerns the changing nature of media influence in digital environments. Algorithmic recommendation systems differ significantly from traditional mass media structures. Earlier media systems could amplify deviant or extremist narratives, but such amplification typically occurred in discrete or episodic ways. In contrast, AI-driven recommendation systems continuously adapt information exposure through behavioural feedback loops. This adaptive structure allows exposure patterns to evolve dynamically in response to user engagement, potentially producing gradual escalation in ideological intensity.

Insights from predictive coding theory further help explain how such exposure dynamics may contribute to ideological consolidation. Cognitive models of belief formation suggest that individuals update beliefs when encountering disconfirming information that generates prediction errors. In digitally mediated echo systems, however, exposure to such disconfirming information may decrease substantially. When information environments repeatedly confirm existing beliefs, prediction errors are reduced and belief systems may become increasingly stable and resistant to revision.

The framework also highlights the distributed nature of responsibility for algorithmically mediated radicalisation. From the perspective of **Actor–Network Theory** (Latour, 2005), radicalisation cannot be attributed solely to individual psychological vulnerability or isolated pieces of content. Instead, it emerges from interactions among multiple actors, including technological systems, platform design choices, monetisation incentives, and user engagement behaviours. Algorithmic recommendation architectures therefore function as active components within networks that shape the visibility and circulation of particular narratives.

These findings carry important implications for both platform governance and regulatory oversight. If engagement-optimised recommendation systems systematically amplify emotionally salient and grievance-oriented narratives, regulatory responses that focus exclusively on removing individual pieces of harmful content may be insufficient. Governance frameworks may also need to address the structural amplification mechanisms embedded within algorithmic systems.

Regulatory instruments such as the **European Union Artificial Intelligence Act** (European Parliament & Council, 2024) and the **Digital Services Act** (European Parliament & Council, 2022) already require large digital platforms to evaluate systemic risks associated with their services. The indicators proposed in this study—Extremity Drift (EDI), Engagement Volatility (EVS), Homophily Density (HDM), and Narrative Convergence (NCR)—could potentially contribute to such assessments by providing analytical signals of emerging radicalisation trajectories. Incorporating these indicators into algorithmic auditing procedures may allow earlier detection of escalating exposure patterns and support the development of platform safeguards aimed at reducing the amplification of high-risk informational dynamics.

10. Limitations

Several limitations should be acknowledged when interpreting the findings of this study.

First, the proposed framework is primarily conceptual and modelling-oriented rather than based on proprietary platform datasets. Due to the limited accessibility of large-scale algorithmic interaction data from major digital platforms, the analysis relies on theoretical synthesis and

simulated modelling rather than direct observation of recommender system behaviour within commercial platforms.

Second, the neurocognitive mechanisms discussed in this study are inferred from existing research in cognitive neuroscience and behavioural psychology, rather than from direct neuroimaging observations of users interacting with digital platforms. While previous studies provide strong evidence regarding dopaminergic reward systems, emotional salience processing, and belief reinforcement dynamics, the specific neural responses associated with algorithmically curated information environments require further empirical investigation.

Third, the regulatory landscape addressed in this study continues to evolve. Legal frameworks such as the **European Union Artificial Intelligence Act** (European Parliament & Council, 2024) and the **Digital Services Act** (European Parliament & Council, 2022) are still in early stages of implementation, and the practical mechanisms through which systemic risk assessments and algorithmic audits will be conducted remain under development.

Future research should therefore seek to validate the indicators and modelling framework proposed in this study using empirical transparency datasets, platform audit data, and controlled exposure experiments. Such studies could provide more precise measurements of algorithmically mediated radicalisation trajectories and further refine the neuro-algorithmic governance framework introduced here.

11. Conclusion

This study develops a neuro-algorithmic framework for understanding radicalisation within contemporary platform societies. The analysis suggests that radicalisation in digitally mediated environments increasingly emerges from interactions between engagement-optimised AI recommendation systems and human neurocognitive reinforcement mechanisms.

Several key conclusions follow from this framework:

First, engagement-driven recommender systems can potentially accelerate extremity drift by repeatedly amplifying emotionally salient narratives through recursive feedback mechanisms. Because recommendation systems adapt to user behaviour, exposure patterns may gradually shift towards increasingly polarised content.

Second, neurocognitive processes associated with reward learning, emotional salience, and stress responses interact with algorithmic exposure dynamics. Repeated engagement with emotionally charged narratives may therefore reinforce ideological interpretations and contribute to the consolidation of identity-based belief systems.

Third, radicalisation trajectories within algorithmically mediated environments can be conceptually operationalised through indicators-based behavioural measures. Metrics such as the **Extremity Drift Index (EDI)**, **Engagement Volatility Score (EVS)**, **Homophily Density Metric (HDM)**, and **Narrative Convergence Rate (NCR)** may provide a foundation for identifying patterns of escalating exposure within digital platforms.

Finally, the findings suggest that governance strategies should move beyond isolated content moderation towards systemic assessment of algorithmic amplification processes. Regulatory frameworks addressing platform risks may benefit from incorporating cognitive impact metrics and independent algorithmic auditing procedures capable of detecting harmful exposure dynamics.

Taken together, the results indicate that radicalisation in digital platform environments should be understood not solely as a product of individual beliefs or social grievances, but as an emergent phenomenon shaped by the interaction between human cognitive vulnerabilities and large-scale algorithmic infrastructures. Addressing these dynamics may require greater transparency in recommendation systems, improved auditing mechanisms, and the development of standardised indicators capable of identifying harmful amplification patterns within algorithmically mediated information ecosystems.

Author Contributions: IVS was responsible for the conceptual and analytical development of the study and authored Sections 2, 3, 4, 5, 6, 7, 8 and 9 of the manuscript. BP and VN contributed to the development of the introductory and concluding components of the article, and co-authored the Abstract and Sections 1, 10, and 11. They also contributed to the refinement of the mathematical models presented in Sections 7.1 and 7.2.

References

Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. *Criminology, 30*(1), 47–88. https://doi.org/10.1111/j.1745-9125.1992.tb01093.x

Agnew, R. (2006). *Pressured into crime: An overview of general strain theory*. Oxford University Press.

Akers, R. L. (1973). *Deviant behavior: A social learning approach*. Wadsworth.

Akers, R. L., & Jensen, G. F. (2006). The empirical status of social learning theory of crime and deviance. *Advances in Criminological Theory, 15*, 37–76.

Arnsten A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature reviews. Neuroscience*, *10*(6), 410–422. https://doi.org/10.1038/nrn2648

Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. Kurtines & J. Gewirtz (Eds.), *Handbook of moral behavior and development* (pp. 45–103). Erlbaum.

Bauman, Z. (2000). *Liquid modernity*. Polity Press.

Becker, H. S. (1963). *Outsiders: Studies in the sociology of deviance*. Free Press.

Clark, A., & Hohwy, J. (2024). Predictive processing and belief rigidity in polarized environments. *Trends in Cognitive Sciences, 28*(3), 185–198. doi: 10.1007/s11023-017-9441-6

European Parliament and Council. (2022). *Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act).*

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 on Artificial Intelligence (Artificial Intelligence Act).*

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Hirschi, T. (2002). Causes of Delinquency (1st ed.). Routledge. https://doi.org/10.4324/9781315081649

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford University Press. https://doi.org/10.1093/oso/9780199256044.001.0001

Merton, R. K. (1938). Social structure and Anomie. *American Sociological Review, 3,* 672–682. https://doi.org/10.2307/2084686

Neumann, P. (2016). *Radicalized: New jihadists and the threat to the West*. I.B. Tauris.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in cognitive sciences*, *9*(5), 242–249. https://doi.org/10.1016/j.tics.2005.03.010

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Zuboff, S (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs.