White Paper

How Recommendation Systems Amplify Harm

*The AMPH Framework (**A**mplification **M**echanisms **of P**latform **H**arms) for understanding amplification mechanisms in recommendation systems*

Authors: Dr Akshi Kumar, *Goldsmiths, University of London, United Kingdom*
Dr Saurabh Raj Sangwan, *Maharishi Markandeshwar (Deemed to be University), India*

Published: May 2026

Type: Technical White Paper

Topics: Algorithmic Amplification, Recommendation Systems, Platform Risk, Trust & Safety, AI Governance

Executive Summary

Algorithmic amplification refers to the ways recommendation systems increase the visibility and repetition of content, creating risk through exposure patterns rather than individual items. Regulatory frameworks in the United Kingdom, European Union, Australia, and the United States now require platforms to assess these system behaviours, yet existing models describe harm categories without explaining how amplification occurs.

This white paper introduces the **A**mplification **M**echanism of **P**latform **H**arms (AMPH) taxonomy, which identifies six amplification mechanisms within recommendation systems: algorithmic looping, escalation, network driven boosting, artificial virality, targeted intensification, and cross modal reinforcement. AMPH links each mechanism to diagnostic signals that can be used in audits, risk assessments, and platform evaluations. The framework presents an end-to-end model showing how user actions translate into engagement signals, ranking decisions, and amplification patterns within recommendation systems. It also outlines detection indicators, mitigation strategies, and an operational audit checklist aligned with regulatory expectations under the Online Safety Act and the Digital Services Act.

Together, these components provide a structured approach for evaluating amplification mechanisms and supporting compliance, transparency reporting, and Trust & Safety operations.

What AMPH Introduces

- Six amplification mechanisms
- Diagnostic signals for audits
- Risk scoring framework
- Operational audit matrix
- Mechanism-based regulatory analysis
- Trust & Safety implementation guidance

Why This Matters

- Recommendation systems shape billions of online experiences every day, influencing what users watch, read, engage with, and believe.
- Online harm is often driven by patterns of amplification and repeated exposure rather than by individual pieces of content alone.

- Recommendation systems can unintentionally reinforce emotional distress, misinformation, polarisation, harassment, and harmful behavioural pathways.
- The framework supports platform audits, Trust & Safety operations, transparency reporting, risk assessments, and AI governance practices.
- AMPH provides a shared operational vocabulary for regulators, researchers, engineers, and policy teams working on platform safety and systemic risk.
- Current regulatory frameworks increasingly require platforms to assess systemic risks created by algorithmic recommendation and ranking systems.
- Existing models explain categories of harm but provide limited guidance on how amplification mechanisms operate inside recommendation systems.
- AMPH introduces a mechanism-based framework for identifying and analysing amplification behaviours within digital platforms.

Key Insight

Recommendation systems do not simply recommend content. They shape exposure trajectories through repetition, escalation, virality, and behavioural reinforcement.

Understanding these amplification mechanisms is essential for platform safety, AI governance, and systemic risk assessment.

Who This White Paper Is For

This white paper is intended for:

- Trust & Safety teams responsible for platform integrity and risk mitigation
- AI governance researchers examining systemic algorithmic harms
- Policymakers developing online safety and platform accountability frameworks
- Platform engineers and recommendation system designers
- Digital regulators and compliance professionals
- Audit teams conducting algorithmic risk and transparency assessments
- Platform risk analysts and online safety practitioners
- Researchers working on recommender systems, misinformation, and digital harms
- Organisations developing AI safety, transparency, and accountability strategies

Amplification **M**echanism of **P**latform **H**arms **(AMPH) at a Glance**

Mechanism	Core Behaviour	Primary Risk
Algorithmic Looping	Repetition of similar content	Emotional reinforcement
Escalation	Gradual movement toward extreme material	Radicalisation
Network-Driven Boosting	Coordinated amplification	Manipulation and harassment
Artificial Virality	Engagement-driven spread	Misinformation
Targeted Intensification	Reinforcement based on inferred traits	Vulnerability exploitation
Cross-Modal Reinforcement	Narrative repetition across formats	Persistent harmful narratives

1. The Amplification Problem

Recommendation systems are central to how information is organised and consumed online (Stohr & Viswanathan, 1998). They determine which videos appear next, which posts rise in a feed, and which topics sustain visibility. These systems influence news exposure, social interaction, and public understanding of major events (Raza and Chen, 2022). Their widespread adoption has created an information environment in which a small number of algorithmic decisions can shape the experiences of millions of users.

While recommendation systems increase relevance and convenience, research shows that they can also create significant risk. Studies document repetitive surfacing of self-harm material, amplification of emotionally charged misinformation, uplift of hostile communities, and rapid movement toward more extreme content (Goswami, 2024). On major video platforms, users often progress from mainstream material to more intense or harmful content within short sequences. Ribeiro et al. (2019) show transitions from centrist to far-right channels, while Papadamou et al. (2020) identify movement from general relationship videos to incel and male-supremacist content. Similar patterns occur in political contexts, where engagement with mainstream topics leads to increasingly polarised or misleading recommendations (Hosseinmardi et al., 2021; van der et al., 2025). These risks arise from amplification behaviours within recommender systems rather than from any single item (Ukkola, 2025). Research across self-harm communities, misogyny ecosystems, public health misinformation, and extremist mobilisation shows that the *sequence and frequency* of recommendations can generate strong reinforcement effects (Benkler et al., 2018; Regehr, 2025). As a result, the behaviour of the recommendation system itself becomes a primary driver of risk (Furini, 2024).

Regulatory frameworks now recognise these system-level risks. The United Kingdom Online Safety Act requires assessment of harms arising from algorithmic design choices (UK Parliament 2023). The European Union Digital Services Act mandates evaluation of systemic risks linked to misinformation, civic manipulation, and automated recommendation systems (European Union 2022). Comparable expectations appear in Australia (eSafety Commissioner 2021) and in U.S. advisory reports (White House OSTP 2022). Although these frameworks acknowledge algorithmic risk, they do not provide a consistent structure for describing the mechanisms through which amplification operates.

The Amplification Mechanism of Platform Harms (AMPH) taxonomy addresses this gap. AMPH is a mechanism-based framework that identifies six amplification behaviours within recommendation systems. Each mechanism reflects a distinct system behaviour and a corresponding pattern of user exposure. These include repetition of similar content, progression toward more intense material, uplift driven by network activity, rapid circulation resembling virality, concentrated reinforcement for specific users, and propagation of related narratives across formats.

The purpose of this white paper is to describe each mechanism, demonstrate how it operates using documented examples, and show how the AMPH taxonomy supports regulatory and audit practice. The framework provides an operational structure for risk assessments, transparency reporting, algorithmic audits, and Trust and Safety evaluations. It offers a practical way to understand how system-level behaviours influence exposure patterns and can be addressed through mechanism-level interventions. This paper makes the following contributions:

- Introduces AMPH as a mechanism-based taxonomy grounded in observable system behaviours.
- Identifies six amplification mechanisms and provides evidence-based examples of each.
- Presents diagnostic signals, operational indicators, and an assessment matrix for detecting amplification patterns.
- Provides a risk scoring model and audit tools that translate the taxonomy into practical applications for regulatory compliance and platform safety evaluations.

The remainder of this paper is organised as follows. Section 2 outlines key concepts and summarises the regulatory landscape that recognises amplification as a systemic risk. Section 3 introduces the AMPH taxonomy, explains each mechanism, and presents an end-to-end model linking user actions to exposure trajectories. Section 4 compares AMPH with existing regulatory and research frameworks. Section 5 sets out system-level mitigation strategies aligned with ranking adjustments and engineering practice. Section 6 provides a practical AMPH audit checklist and outlines ethical

considerations for researchers. Section 7 concludes by summarising the value of mechanism-based analysis and identifying directions for future work.

2. Understanding Amplification and Regulatory Risk

This section introduces the core terms used throughout the paper and summarises the regulatory landscape that recognises amplification as a source of online harm.

2.1 Key Concepts and Definitions

A small number of technical ideas recur throughout this paper. The explanations below provide non-technical readers with the foundation needed to understand amplification and system behaviour.

- *Algorithms:* Rules that tell a platform how to make decisions. They determine which items appear first, which items are recommended next, and which items receive visibility (Burrell, 2016).
- *Recommendation Systems:* The part of the platform that selects and ranks content for each user. It decides the sequence of videos or posts a person sees (Ricci et al., 2010*)*.
- *Ranking:* The ordering of content. Items placed earlier are treated as more relevant or more engaging (Karatzoglou et al., 2013).
- *Engagement Signals:* Indicators such as views, likes, comments, shares, and watch time. Platforms treat these signals as evidence of interest and often use them to promote content (Covington et al., 2016).
- *Amplification:* The process through which a system increases the visibility or repetition of certain items. Harm often arises from these patterns of exposure rather than from any single item (Huszár et al., 2022).
- *Exposure Trajectory:* The sequence of content that a user encounters over time. Trajectories matter because they can gradually shift mood, beliefs, or perceptions (Heiss et al., 2020).
- *Network Activity:* Collective patterns of interaction, such as coordinated posting or rapid bursts of engagement. These patterns can artificially boost visibility (Yang, 2024).
- *Cross Format Content:* Narratives that appear across text, images, audio, and video. When a theme spreads across formats, it becomes more persistent and harder to moderate (Zhang and Ruixue, 2025).
- *Systemic Risk:* Risk created by the design and behaviour of the platform itself rather than by individual items of content. Regulators require platforms to identify and manage these risks (Müller and Matthias, 2024).
- *Amplification Mechanism:* A specific system behaviour that drives visibility, repetition, or progression. The AMPH taxonomy identifies six such mechanisms, which form the analytical basis of this paper (Diberardino et al., 2024).

2.2 Regulatory Background

Regulators in several jurisdictions have increasingly recognised that online harms arise not only from the presence of harmful content but also from the design and operation of recommendation systems that determine how users encounter information. The United Kingdom's *Online Safety Act* introduces statutory duties of care that require platforms to assess and mitigate risks created by algorithmic ranking, personalisation, and automated recommendations (Ofcom, 2023). Under this framework, platforms must demonstrate that their systems do not inadvertently steer users toward harmful material through repeated exposure, ranking choices, or interaction-driven optimisation.

A similar emphasis appears in the European Union's *Digital Services Act*, which identifies algorithmic amplification as a core driver of systemic risk. The DSA requires Very Large Online Platforms to evaluate how automated recommendations contribute to misinformation, civic manipulation, public health misinformation, and the spread of illegal or harmful content (European Commission, 2022). Platforms must also document the logic of their recommendation systems and ensure that risk mitigation measures address design-level factors rather than only content-level moderation.

Beyond the UK and EU, comparable expectations have emerged internationally. The Australian eSafety Commissioner highlights algorithmic curation and engagement-driven recommendation as

sources of harm for vulnerable users and calls for transparency around ranking and amplification processes (eSafety Commissioner. 2021). In the United States, federal advisory reports and policy guidance similarly note that recommendation systems can magnify harmful narratives, influence mental health, and shape political discourse through automated patterns of exposure (White House Office of Science and Technology Policy. 2022). International organisations such as the OECD also identify algorithmic amplification as a mechanism that can intensify risks for children, minority groups, and information integrity across digital environments (OECD, 2021).

Across these frameworks, a shared recognition has emerged: harm can be created by system behaviour rather than by inherently harmful content. However, despite acknowledging algorithmic amplification, current regulatory models provide limited methodological guidance for analysing the specific system behaviours that produce harmful exposure patterns. They describe harms and require risk assessment, but they do not offer a consistent structure for identifying how recommendation systems generate repetition, escalation, network-driven visibility, or cross-format propagation.

This gap motivates the development of the AMPH taxonomy introduced in Section 3. Existing regulatory frameworks require platforms to assess systemic risks, yet they provide limited guidance on how to identify the specific system behaviours that create harmful exposure patterns. Current approaches typically focus on:

- content categories (e.g., misinformation, hate, self-harm),
- user outcomes (e.g., polarisation, emotional distress), or
- broad descriptions of algorithmic influence,

but they do not explain *how* recommendation systems produce these harms through their internal operations. As a result, regulators, auditors, and engineers lack a consistent method for diagnosing the mechanisms that underpin algorithmic amplification.

The AMPH taxonomy addresses this gap by providing a mechanism-based approach that identifies the specific behaviours within recommendation systems that drive harmful visibility and repetition. It differs from existing models in several important ways:

- It examines system behaviours rather than content themes, allowing analysis across domains and contexts.
- It identifies six amplification mechanisms, including repetition, escalation, and network-driven uplift.
- Each mechanism is tied to observable diagnostic signals, making amplification measurable in operational settings.
- It enables technical auditing by translating high-level risk duties into concrete system indicators that can be monitored.
- It complements, rather than replaces, regulatory harm categories by supplying the analytical layer that links regulatory expectations to platform engineering practice.

Through this mechanism-oriented structure, AMPH supports more precise risk evaluation, clearer accountability for design choices, and more targeted mitigation strategies than existing classification models.

3. The Six Amplification Mechanisms

The Amplification Mechanism of Platform Harms (AMPH) taxonomy is designed to analyse *mechanisms* rather than content categories or user outcomes. “Amplification Mechanism” refers to the distinct system behaviours within recommendation architectures that increase the visibility, repetition, or progression of content. “Platform Harms” signals that the risks arise from design-level processes embedded in ranking and recommendation logic, not from isolated content items. AMPH therefore shifts focus from topic-based harm classifications to a mechanism-oriented analysis that can be used in audits, transparency reporting, and regulatory compliance assessments.

The AMPH taxonomy identifies six amplification mechanisms, each representing a specific behavioural pattern in recommendation systems and its effect on user exposure. To support operational clarity, each mechanism is presented with a definition, real-world examples, and its associated risks. This structure enables consistent application across regulatory analysis, academic research, and platform auditing. Figure 1 situates the six mechanisms within the end-to-end operation

of a recommendation system, showing how user behaviour is translated into engagement signals, processed through ranking logic, and expressed as amplification patterns in user exposure.

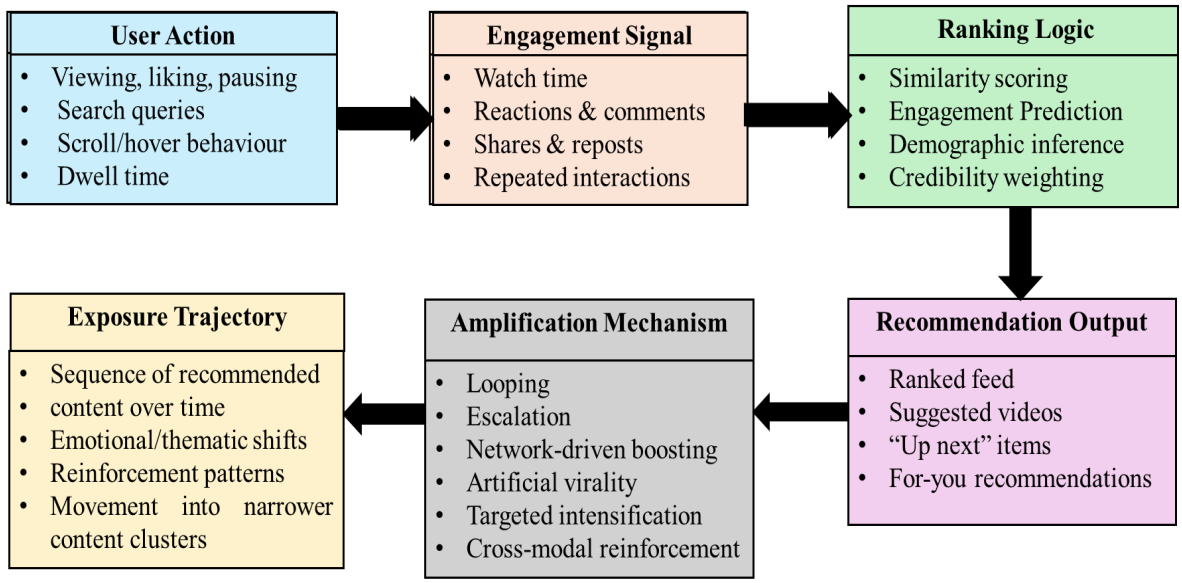

Fig. 1. End-to-End Amplification Framework: From User Actions to Amplification Mechanisms

3.1 Algorithmic Looping

Core Pattern: *Repetition that traps users in a narrow stream of similar content.*

Definition
Algorithmic looping occurs when a recommendation system repeatedly shows content with similar themes, tones, or emotional signals. Early engagement is treated as a stable preference, which leads the system to keep presenting similar content and narrowing what the user sees over time.
Examples

- *Emotional content loops:* A user watches one video about work stress. The system starts recommending material on burnout, anxiety, depression, and loneliness. US Senate hearings reported similar patterns, where teenagers received long sequences of depression related content after very limited interaction (US Senate, 2021).
- *Body image loops:* A teenage girl views one fitness or dieting video. The system then recommends thinspiration clips, calorie restriction advice, and comparison-based content. The Molly Russell inquest in the United Kingdom documented repeated exposure to appearance and self-harm material (Coroner's Court of London 2022; Cataldo et al., 2021).
- *Fear and threat loops:* Watching one home security video can lead to recommendations about burglary, violent crime, and home invasions. Studies show that repeated crime related exposure can increase fear and distort perceptions of personal risk, even in areas with low crime rates (Chadee et al., 2019; Buckley 2025).
- *Mood based music loops:* Listening to one sad song can prompt the system to offer more sad tracks, then heartbreak songs, and then content with heavier emotional themes. This gradually shifts the user's music feed toward a consistently negative mood (Büyüksağnak, 2021).

Risk
Algorithmic looping reduces the variety of content that users encounter. It reinforces particular emotional or thematic patterns and can increase exposure to material linked to distress, body image pressure, fear, or self-harm. Because similar items repeat, users have fewer chances to see balancing or corrective content.

3.2 Escalation

Core Pattern: *Step-by-step shifts toward stronger or higher-risk material.*

Definition
Escalation refers to a sequence of recommendations that moves a user from mainstream or mild content to material that is more intense, polarised, or harmful. The shift takes place in small steps, as the system prioritises content that is similar but gradually stronger in tone or theme.

Examples

- *Misogyny escalation pathway:* A user watches a general dating advice video. The system then recommends critiques of dating norms, followed by creators linked to men's rights narratives, and later content associated with male supremacist or incel communities. Investigations in the United Kingdom found that boys who viewed motivational or lifestyle videos were soon recommended Andrew Tate content and related gender hostile material (University College London 2024). Studies of YouTube recommendation patterns show that neutral men's issues or dating content can lead to incel aligned and misogynistic material (Papadamou et al. 2020).
- *Political escalation pathway:* Viewing mainstream political or election content can lead to increasingly partisan material, then polarising opinion clips, and in some cases, conspiratorial or extremist viewpoints. Research shows that algorithmic curation can reinforce ideological preferences and contribute to filter bubbles and polarisation (Park and Park 2024). Other analyses show that ranking systems can amplify extremism or highly divisive political content even when user interactions begin with neutral material (Whittaker et al. 2021).
- *Health escalation pathway:* Engaging with general wellness or fitness material can lead to unverified medical advice or anti vaccine narratives. Computational studies of YouTube trajectories show that videos on general health and COVID 19 vaccines can direct users toward misinformation and misleading interpretations (Ng et al., 2023). Further research shows that exposure to vaccine misinformation can shape later information seeking and increase uncertainty or distrust (Kessler and Humprecht 2023). Studies with young people also show that harmless dieting content can escalate into extreme calorie restriction and other unsafe health narratives (Sauerwein 2025).
- *Anger and outrage escalation:* Content that begins with reasonable criticism can be followed by items that emphasise anger, threat, or hostility toward groups. Internal platform documents show that material producing strong emotional reactions, especially anger, receives higher algorithmic weighting, making escalation more likely (Nieman Lab 2021). Empirical work also shows that engagement with grievance-oriented content can heighten feelings of threat and increase prejudice toward immigrant or minority communities (Ahmed et al. 2024).

Risk
Escalation increases the chance that users will encounter more extreme or misleading content than they intended. It can support radicalisation, strengthen adversarial or exclusionary attitudes, encourage unsafe behaviour, and increase hostility toward targeted groups.

3.3 Network Driven Boosting

Core Pattern: *Amplification created by coordinated or engineered engagement.*

Definition
Network driven boosting occurs when content gains visibility because of coordinated activity or structural patterns in a network, rather than genuine individual interest. Influencers, organised groups, automated accounts, or bot networks can raise engagement signals in ways that the system interprets as popularity.

Examples

- *Coordinated harassment campaigns:* Groups may target an individual or community by repeatedly commenting, sharing, or reacting to hostile posts. The system reads this clustered activity as engagement and increases the visibility of the content. Research shows that coordinated manipulation and harassment often rely on organised groups that amplify abusive material through structured or semi structured actions (Marwick and Lewis 2017; de Lima Santos and Ceron 2023). Analyses of influence networks show that coordinated patterns can significantly expand the reach of disinformation spreaders and hostile accounts (Yang and Williams 2024).
- *Influencer chains:* Influencers working in the same ideological or thematic area may repeatedly reference or promote each other, creating network loops that boost visibility. Research shows that such cross promotion helps conspiracy, ideological, or fringe content circulate more rapidly across follower groups (Heřmanová 2022). Studies of far-right ecosystems show that these networks often use strategic linking to extend reach and increase the prominence of their content in ranking systems (Rothut et al. 2024).
- *Bot assisted amplification during elections:* Automated accounts can inflate engagement around political content, increasing the chance that misleading or polarising material is promoted by ranking systems. Research shows that bot networks have been used for years to influence electoral discourse by amplifying partisan narratives (McKelvey and Dubois 2017). Further studies show that bots promote divisive or inaccurate content at key moments in election cycles (Boichak et al. 2018). Recent work highlights the risk these systems pose to democratic processes by shaping visibility and boosting emotionally charged content (Feldman and Nahmias 2024). Emerging analyses warn that more advanced AI driven agents may further increase this influence (Ash, Galletta, and Opocher 2025).
- *Raid communities:* Organised groups can coordinate high volumes of hostile posts, flooding comment sections with insults or threats. Platforms often interpret these rapid bursts of interaction as relevance, which increases content visibility. Research documents how creators on live streaming platforms experience coordinated hate raids (Meisner 2023). Computational studies show that these actions can be detected through patterns of synchronised behaviour and clustered interaction networks (Magelinski, Ng, and Carley 2022). Broader social network research shows that collective raids rely on network structures that allow rapid mobilisation and collective hostility (Glowacki et al. 2016).

Risk

Network driven boosting makes harmful content appear popular or widely supported when it is the result of coordinated activity. This can increase the scale of harassment, distort search and recommendation patterns, and expose users to harmful narratives they did not seek.

3.4 Artificial Virality

Core Pattern: *Rapid spread driven by engagement signals rather than accuracy.*

Definition

Artificial virality refers to the rapid spread of content driven by algorithms that prioritise engagement signals such as watch time, reactions, and emotional intensity. Content becomes widely visible because it triggers strong responses, not because users intentionally search for it or because it is accurate.

Examples

- *False health claims:* Videos promoting unproven cancer treatments, anti-vaccine narratives, or misleading medical advice often gain wide visibility when personal testimonies or sensational framing generate high engagement. Studies of health content on YouTube show that unreliable or misleading material can circulate broadly and create environments where users repeatedly encounter false medical claims (Di Marco, Cinelli, and Quattrociocchi 2021). Research on cancer misinformation also finds that social media hosts large volumes of

inaccurate treatment advice that spreads rapidly through user interaction (Loeb et al. 2024). Newer video platforms show similar patterns where pseudoscientific and unverified wellness claims are frequently amplified (Andrikopoulou, Talam, and Kanta 2025).

- *Provocative misinformation:* Conspiracy theories often rise in visibility when large numbers of comments, whether supportive or critical, are interpreted by algorithms as indicators of interest. Research shows that users are more likely to engage with conspiratorial content when it is emotionally charged or cognitively stimulating, creating a feedback loop that increases amplification (Stecula and Pickup 2021). Studies from the COVID 19 period show that sensational and conflict driven misinformation spreads rapidly across platforms (Kuzelewska and Tomaszuk 2022). Experimental work shows that crisis contexts make such narratives more appealing and more likely to escalate through algorithmic ranking (Pummerer and Sassenberg 2020).
- *Crisis misinformation:* During emergencies, unverified claims can spread quickly as uncertainty and anxiety increase the likelihood of resharing. Research shows that crisis events create conditions where rumours circulate before official information is available (Oh, Agrawal, and Rao 2013). Experimental studies show that misleading crisis content, such as false warnings or sensational explanations, attracts high engagement, making algorithmic amplification more likely (Lee et al. 2024). Sociotechnical analyses show that crises can produce amplification loops where user behaviour and platform design jointly increase the visibility of inaccurate claims (Eriksson Krutrok and Lindgren 2022).
- *Outrage loops:* Content that triggers anger, shock, or moral outrage often spreads rapidly because ranking systems prioritise emotionally intense reactions. Research shows that anger increases the speed and reach of online content (Chuai and Zhao 2022). Studies of emotional information flows show that high affect material, especially anger, produces more interaction and therefore receives more algorithmic amplification (Chen 2023). Recent computational work confirms that emotional valence strongly influences virality, with anger driven posts receiving disproportionate boosting (Yu et al. 2025).

Risk
Artificial virality speeds up the spread of false or misleading content and increases public confusion. It can weaken trust in institutions and influence behaviour at scale. Because visibility is based on engagement rather than accuracy, misleading material can reach many users quickly and continue circulating even after corrections appear.

3.5 Targeted Intensification

Core Pattern: *Repeated recommendations shaped by inferred traits or vulnerabilities.*

Definition
Targeted intensification occurs when a recommendation system repeatedly presents content to users based on inferred demographic or behavioural characteristics. The system strengthens patterns found in user activity even when those patterns involve sensitive topics or behaviours linked to risk.

Examples

- *Gender specific harm:* Young men may be shown repeated content promoting aggression in gender relations, dominance, or confrontational dating advice. Safeguarding reviews report that boys who watched motivational or lifestyle material were repeatedly exposed to videos from Andrew Tate and similar gender hostile creators (University College London 2024). Teen girls are often shown body image and dieting content, a pattern identified in reviews linked to the Molly Russell case, which documented sustained exposure to appearance focused and emotionally harmful material (Coroner's Court of London 2022; Cataldo et al., 2021).
- *Age based vulnerability:* Younger users may receive content that encourages risky or impulsive behaviour, such as dangerous challenges or confrontational pranks. Research on children's online experiences shows that seemingly playful content can be followed by

material involving more problematic activities (Haddon and Livingstone 2014). Studies of digital participation patterns indicate that age and gender shape the types of content young people are exposed to (Carcelen Garcia, Narros Gonzalez, and Galmes Cerezo 2023). In contrast, adults are more likely to be directed toward political or value-based content aligned with inferred identity traits, and research suggests that intergroup communication dynamics can increase susceptibility to escalatory material among older groups (Chen and Chen 2025).

- *Behavioural profiling:* Recommender systems infer user preferences from interaction patterns, and these inferred profiles influence what is shown next. Research on dynamic user modelling shows that even small shifts in user behaviour can lead systems to introduce more extreme or sensitive content (Chee et al. 2024). Studies on profiling and high-risk classification note that automated inferences about user traits can unintentionally expose users to risk relevant content when these traits are treated as stable interests (Nannini 2025). Earlier work on adaptive news systems shows that open profiles can create reinforcement loops that narrow information environments (Ahn et al. 2007). Experimental research indicates that exposure to conspiracist narratives alone is not persuasive, suggesting that profiling-based amplification rather than simple exposure is the key risk (Nera, Pantazi, and Klein 2018).
- *Lifestyle categorisation:* Brief engagement with content about financial stress or budgeting can lead to recommendations involving investment advice or aspirational finance material. Research shows that social media shapes financial decision making, particularly among younger or less experienced users, who may be exposed to higher risk content (Miettinen 2025). Studies on financial influencer ecosystems show that platforms can amplify speculative or unverified investment guidance through performance following and social transmission bias (Cao et al. 2025). Other work shows that social media can guide users toward unsuitable or risky financial choices (Singh, Mahajan, and Kaur 2025). Research on personalised advertising indicates that young and socioeconomically vulnerable groups may be disproportionately affected by targeted financial promotions (Saez Linero and Jimenez Morales 2025). Broader studies on AI driven content filtering show that recommendation systems can influence youth lifestyles by introducing promotional or aspirational financial content beyond the user's initial intent (Misnawati et al. 2025).

Risk

Targeted intensification produces unequal and sometimes discriminatory content patterns. It strengthens vulnerabilities in specific groups, increases exposure to harmful or misleading narratives, and hides from users the extent to which their feed reflects inferred traits rather than intentional choices.

3.6 Cross Modal Reinforcement

Core Pattern: *The same narrative pushed across multiple content formats.*

Definition

Cross modal reinforcement occurs when a platform presents related content across several formats, such as text posts, short videos, long videos, images, and audio. The system links themes across formats, creating a pattern of exposure that makes the narrative more persistent and more coherent for the user.

Examples

- *Misogynistic ecosystems:* Engagement with one form of gender hostile content can lead to related material in other formats. Research on transnational incel and misogynistic communities shows that users who view text posts about frustration with dating or gender roles often see short videos, memes, and other visual or audio content that repeat the same themes (Anastasi 2025). Multimodal corpus studies show that misogynistic narratives are sustained through coordinated use of text, images, and video (Anastasi et al. 2023). Studies of misogyny memes show that images and other multimodal content play a major role in reinforcing hostile gender narratives (Chen et al. 2024). Additional work on YouTube shows

that gender abusive language often appears within audiovisual and comment based patterns that support the same frames (Esposito and Zollo 2021).

- *Political misinformation:* Political falsehoods often appear across images, text posts, short videos, and long videos. Multimodal forensic linguistic work shows that political hoaxes combine visuals, captions, and extended commentary to support misleading interpretations of events or data (Astuti and Mulyadi 2025). Research on cross platform misinformation shows that images and text posts are frequently linked to video content that embeds the same narrative in a broader argument (Micallef et al. 2022). Multimodal disinformation studies find that political falsehoods become more coherent and more persistent when they are presented across multiple formats (Wilson et al. 2023).
- *Deepfake harm:* Synthetic or manipulated media often spread across multiple formats. An initial AI generated image, or short clip may be followed by recommended videos, commentary, or memes that extend the same false scenario. Research on deepfakes and political disinformation shows that synthetic video can increase deception, uncertainty, and mistrust, especially when supported by additional narrative content (Vaccari and Chadwick 2020). Analyses of generative AI memes during elections show that manipulated visuals often appear within larger multimodal ecosystems (Kuo 2025). Overviews of deepfake manipulation show that fabricated images and videos gain credibility when surrounded by supporting discourse that makes the false scenario appear more plausible (Gangavarapu 2025). Foundational work also shows that deepfakes and cheap fakes rely on repeated circulation across text, video, and sharing patterns to increase visibility and impact (Paris and Donovan 2019).
- *Lifestyle reinforcement:* Engagement with dieting or wellness content can lead to related images, videos, and other material that emphasise similar appearance focused messages. Research on fitspiration videos shows that this type of content often leads users to additional videos and images that promote restrictive behaviour and appearance ideals (Ratwatte and Mattacola 2021). Multimodal studies of images show that body comparison photos and aestheticised depictions of thinness or fitness influence how adolescents perceive body image (Imtiaz 2023). Critical work on sports and weight loss imagery finds that visual content often promotes unrealistic standards, reinforcing weight focused narratives (Li and Huang 2025). Experimental studies show that idealised or manipulated images can negatively affect body image in adolescent girls (Kleemans et al. 2018).

Risk

Cross modal reinforcement increases the strength and persistence of harmful narratives by repeating them across several formats. This makes it more likely that users accept or internalise misleading or harmful themes and makes moderation more difficult because the narrative appears in diverse media types.

3.7 AMPH Mechanism Identification and Audit Matrix

This section presents an operational matrix for applying the AMPH taxonomy in regulatory reviews, platform audits, and risk assessments. The matrix sets out the main diagnostic signals for each mechanism, the methods used to detect them, the evidence required, typical red flags, mitigation priority, and the level of detection complexity. It provides a system level view of how amplification patterns can be identified within real-world recommendation environments.

Table 1. AMPH Mechanism Assessment Matrix

AMPH Mechanism	Key Diagnostic Signal	Primary Detection Method	Evidence to Collect	Red Flags	Mitigation Priority	Detection Complexity
Algorithmic Looping	Repetition of closely related content across consecutive recommendatio	Semantic similarity analysis, diversity scoring	Recommendation sequences across time, topic clusters	Very low diversity despite varied user behaviour	Medium	Low

	ns					
Escalation	Gradual movement toward material with greater intensity or risk	Sentiment trajectory analysis, intensity gradient modelling	Ordered recommendation chains showing thematic or emotional shifts	Clear increase in intensity across a short sequence	High	Medium
Network Driven Boosting	Engagement spikes that do not match ordinary user activity	Network analysis, coordinated activity detection, bot identification	Timestamped engagement logs, network structure, probability of automated activity	Sudden clusters of identical comments or engagement bursts	High	High
Artificial Virality	Rapid growth in engagement from low credibility sources	Engagement velocity analysis, credibility scoring	Engagement curves, source metadata, timeline of spread	Fast circulation of sensational or low reliability content	Medium to High	Medium
Targeted Intensification	Recommendations aligned with inferred demographic or behavioural traits	Fairness audits, clustering analysis, inferred profile detection	Demographic proxies, behavioural logs, correlations between traits and recommendations	Repeated sensitive topic content tied to identity or vulnerability	High	High
Cross Modal Reinforcement	Recurrence of the same narrative across text, image, video, and audio formats	Cross modal similarity analysis, narrative linkage detection	Cross format content maps and metadata	The same theme appearing in several formats within a short period	Medium	Very High

How to Use This Matrix

Auditors and researchers can follow three steps:

1. Identify observable signals in recommendation sequences and match them to the diagnostic indicators in the matrix.
2. Collect the relevant evidence, including engagement logs, recommendation chains, or signs of coordinated behaviour, to confirm which mechanism is present.
3. Apply the mitigation priority to decide whether the mechanism requires immediate action, ongoing monitoring, or longer-term design changes.

This structured process supports consistent identification of amplification mechanisms and helps regulators, platforms, and independent auditors assess how specific mechanisms contribute to systemic risks under the Online Safety Act, the Digital Services Act, and similar regulatory frameworks.

3.8 Quantitative Risk Scoring Framework

This section introduces a quantitative risk scoring framework for comparing the six AMPH mechanisms. The framework assesses potential harm across five dimensions: exposure, amplification magnitude, vulnerability, intensity shift, and spread speed. Each dimension uses a three-point scale to represent low, moderate, or high risk. The scores are used to create a risk heatmap that highlights

which mechanisms require the most attention in regulatory assessments and platform level safety reviews.

Table 2. Risk Scoring Framework for AMPH Mechanisms

Dimension	Score 1 (Low)	Score 2 (Moderate)	Score 3 (High)
Exposure	Limited reach or confined to narrow feeds	Moderate reach across a wider set of users	Widespread visibility across the platform
Amplification Magnitude	Minimal amplification with limited ranking influence	Clear amplification where ranking shapes exposure	Strong amplification where ranking drives dominant visibility
Vulnerability Level	General adult population	Mixed groups including some sensitive users	Minors, vulnerable users, or individuals in crisis
Intensity Shift	No meaningful change in risk across recommendations	Moderate increase in tone, intensity, or harmfulness	Clear progression toward extreme, harmful, or deceptive content
Spread Speed	Slow diffusion	Moderate spread velocity	Rapid expansion with features of virality

This scoring approach helps identify system level patterns rather than focusing on isolated content items. It can be used in transparency reports, compliance assessments, and internal safety monitoring to determine which mechanisms present the greatest systemic risk. Figure 2 illustrates how this scoring approach highlights system-level amplification patterns rather than isolated content items, and supports use in transparency reporting, compliance assessments, and internal safety monitoring.

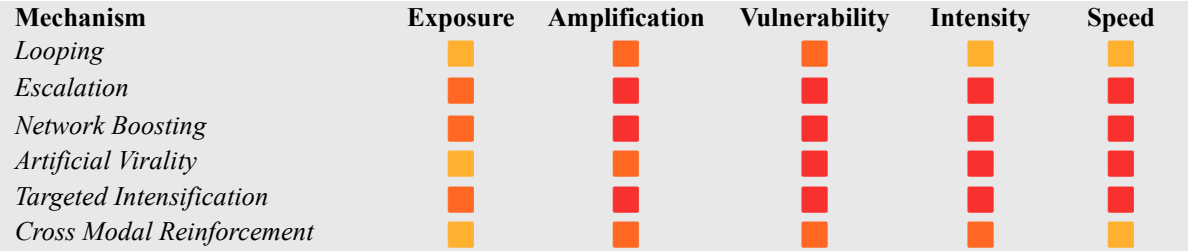

Fig.2. AMPH Risk Scoring Heatmap

The heatmap shows that escalation, network driven boosting, artificial virality, and targeted intensification score high across several dimensions. These mechanisms have the potential for rapid spread, strong intensity shifts, and significant impact on vulnerable users. Looping and cross modal reinforcement show moderate but sustained risk and can influence user experience over longer periods. The distinctions support targeted mitigation by indicating where amplification mechanisms create the most significant systemic effects.

Amplification mechanisms can affect several domains, including emotional wellbeing, body image, identity-based hostility, misinformation, radicalisation, coordinated harassment, democratic processes, public health, and risky behaviour. These outcomes reflect the cumulative exposure patterns created by recommendation systems rather than individual pieces of content.

4. How AMPH Differs from Existing Models

Current regulatory and research frameworks identify forms of online harm, but they rarely explain the system behaviour that creates harmful exposure patterns. Most approaches categorise harm by content, outcome, or impact, while AMPH focuses on the amplification processes that shape user exposure. The table 3 below summarises how AMPH aligns with, differs from, and enhances existing regulatory and research models.

Table 3. Comparison of AMPH With Existing Regulatory and Research Frameworks

Framework / Model	What It Focuses On	What It Does Not Capture	How AMPH Adds Value
Ofcom Risk of Harm Framework	Identifies harmful content categories and risks to users	Does not explain system behaviour that increases exposure to harmful content	AMPH identifies repetition, directional shifts, engagement clustering, and cross format reinforcement, enabling analysis of underlying design processes
European Union Digital Services Act	Requires assessment of systemic risks created by automated recommendations and platform design choices	Does not distinguish between types of amplification or provide indicators for identifying them	AMPH provides diagnostic signals such as velocity of spread, network uplift, and cross format propagation, making DSA assessments operational
Engagement Driven Harm Models	Examine echo chambers, emotional reinforcement, and algorithmic bias	Do not differentiate distinct amplification pathways that produce these effects	AMPH separates repeated exposure, directional escalation, coordinated uplift, and rapid circulation, enabling finer grained analysis of harmful trajectories
Content Based Taxonomies	Classify harm by topic (misinformation, harassment, hate, extremism)	Do not explain why certain content becomes more visible than others or how exposure sequences form	AMPH focuses on system behaviour rather than content theme, enabling cross domain and cross platform applicability
Safety Engineering Models	Identify hazards and states that increase the probability of harm	Lack a structured model for algorithmic amplification and its system level indicators	AMPH introduces six measurable amplification pathways that can be monitored through log analysis, ranking audits, and network mapping

4.1 Distinct Contribution of AMPH

The AMPH framework provides three contributions that are not present in existing models:

- It identifies the system behaviours that generate harmful exposure patterns, rather than categorising harm by content or outcome.
- It links these behaviours to measurable indicators, allowing auditors and regulators to detect amplification effects through empirical evidence.
- It applies consistently across topics, user groups, and platforms, creating a unified vocabulary for analysing amplification related risk.

By introducing a mechanism-based layer of analysis, AMPH bridges the gap between high level harm categories and the operational signals required for technical assessment under regulatory frameworks. To summarise the relationship between AMPH and existing regulatory and research approaches, Figure 3 presents a Venn diagram that shows how AMPH intersects with content-based frameworks, engagement harm models, and safety engineering approaches.

The Venn diagram illustrates that existing frameworks focus on different levels of analysis: content-based frameworks classify what material is harmful, engagement models examine the outcomes that occur for users, and safety engineering approaches assess the conditions that increase the probability of harm. None of these models explain the system behaviour that creates harmful exposure trajectories. AMPH fills this gap by identifying the amplification pathways that drive visibility, repetition, progression, and cross format reinforcement. This mechanism level position shows how AMPH complements both regulatory models and research frameworks by explaining the operational processes that shape user experience.

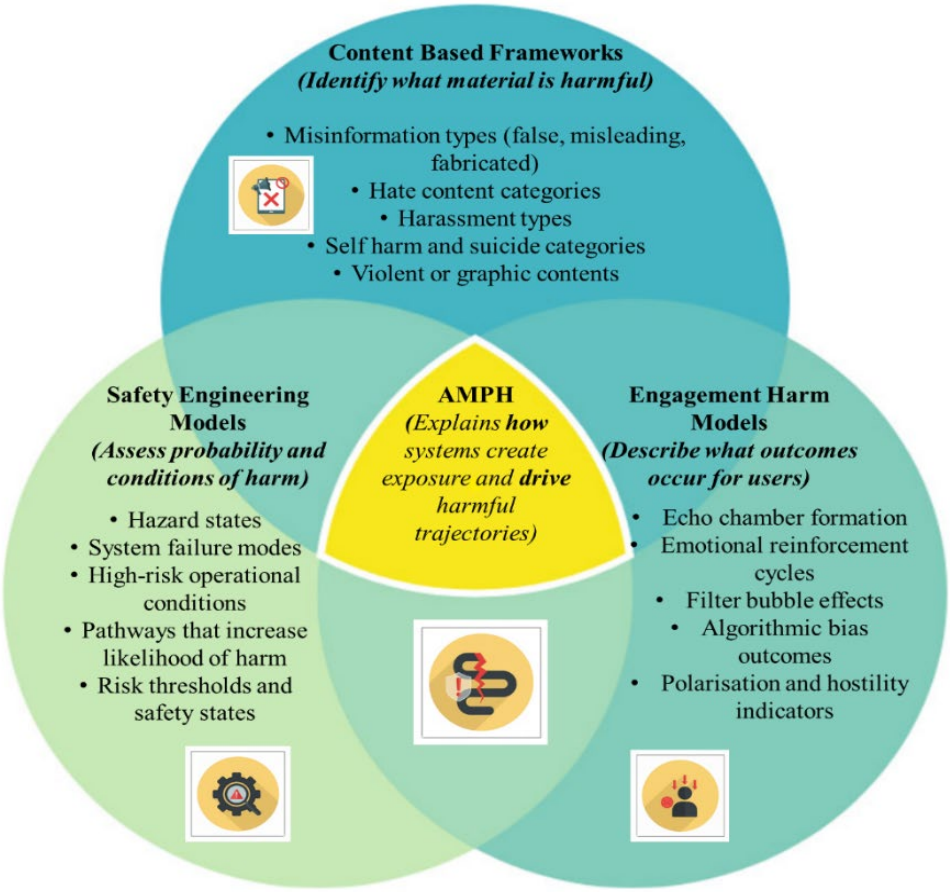

Fig. 3. Venn Diagram Positioning AMPH Within Existing Frameworks

5. Reducing Amplification Risk

The AMPH framework identifies the system behaviours that create harmful exposure patterns. Mitigation therefore needs to focus on the design and operation of recommendation systems rather than on the properties of individual items. The strategies outlined in this section are design oriented recommendations derived from the amplification mechanisms identified in AMPH and from regulatory expectations. They are not presented as empirically evaluated interventions but as system level actions that address identifiable amplification behaviours. Effective mitigation requires three complementary approaches: adjustments to ranking logic, alignment with regulatory duties, and integration of amplification checks into routine engineering practice.

5.1 Technical Adjustments to Ranking Logic

Platforms can reduce harmful amplification by modifying how recommender systems interpret and prioritise engagement signals. Several interventions are relevant:

- *Diversity constraints:* Ranking logic can incorporate diversity checks that prevent prolonged exposure to highly similar material and reduce the likelihood of narrow loops.
- *Tone and intensity monitoring:* Systems can track changes in emotional or adversarial tone across recommendation sequences and reduce the ranking weight of items that intensify a trajectory.
- *Credibility weighting:* Trust indicators and source verification signals can be incorporated into ranking so that rapid engagement does not automatically increase visibility when credibility is low or uncertain.
- *Velocity safeguards:* Content that receives an abrupt rise in interactions can be subject to slower distribution until verification checks are completed, particularly during periods of heightened public risk.
- *Cross format monitoring:* Platforms can observe how narratives appear across text, images, audio, and video, and apply additional scrutiny when a theme spreads rapidly across formats.

These measures target the mechanism level behaviours identified in AMPH and help prevent harmful exposure patterns from forming.

5.2 Alignment with Regulatory Expectations

Regulatory frameworks in the United Kingdom, the European Union, and other jurisdictions require platforms to demonstrate that systemic risks created by algorithmic design are identified and managed. AMPH supports these duties by clarifying which system signals must be monitored. Platforms are expected to:

- track exposure patterns and engagement velocity across user groups
- detect coordinated activity and unusual network behaviour
- apply additional safeguards for users who may be vulnerable or in crisis
- preserve logs and technical documentation that allow auditors to reconstruct system behaviour
- activate enhanced monitoring during elections, emergencies, and public health incidents

These requirements reflect the duties set out in the Online Safety Act, the Digital Services Act, and comparable regulatory frameworks.

5.3 Integration into Engineering and Operational Practice

AMPH can be embedded into engineering workflows and operational processes to ensure consistent oversight of amplification behaviour.

- *Pre-deployment testing:* New ranking features should be assessed to determine whether they increase repetition, intensification, or network driven uplift.
- *Real time dashboards:* Internal tools can display indicators associated with the AMPH mechanisms, allowing engineers to identify unusual exposure patterns early.
- *Shared vocabulary:* Product teams, ranking engineers, and trust and safety teams should use a common terminology for describing amplification, enabling consistent communication and aligned decision making.
- *Incident analysis:* When harmful exposure occurs, AMPH provides a structure for analysing which mechanism contributed to the trajectory and which signals failed to regulate it.

These practices integrate amplification awareness into the engineering lifecycle and support accountable system design.

Together, these technical, regulatory, and engineering measures show how amplification harms can be addressed through system level adjustments. They shift the focus from individual items to the behaviours of the recommendation architecture, ensuring that platforms reduce risk in a structured and transparent way.

5.3.1 Routine Monitoring Indicators

To support operational implementation, platforms can track a small set of system-level indicators linked to the AMPH mechanisms. These indicators can serve as early warning signals in Trust and Safety dashboards:

- Sequence similarity index (for looping)
- Intensity change score across recommendations (for escalation)
- Engagement concentration ratio (for network-driven boosting)
- Velocity-of-spread metric (for artificial virality)
- Trait-correlation drift metric (for targeted intensification)
- Cross-format narrative overlap score (for cross modal reinforcement)

Tracking these indicators enables continuous monitoring and provides Trust and Safety teams with actionable signals that align directly with the AMPH taxonomy. Routine monitoring should also account for ranking drift, where model updates or feature-weight changes alter amplification behaviour over time.

5.4 Operational Implications for Trust and Safety Teams

The AMPH taxonomy provides Trust and Safety teams with a practical structure for identifying how recommendation systems contribute to harmful exposure. It focuses on system behaviour rather than individual items, which makes it suitable for cases where no single piece of content violates policy, but the overall pattern creates risk.

- *Early detection and monitoring:* Each AMPH mechanism is associated with observable signals such as repeated recommendations, increases in content intensity, sudden engagement shifts or consistent narratives appearing across formats. These signals can be integrated into internal dashboards and alert systems to support early detection of emerging problems.
- *Policy alignment and safety categorisation:* Trust and Safety teams typically classify harm into categories such as hate, harassment, self-harm, extremism and misinformation. AMPH helps link exposure patterns to these categories. Escalation pathways can indicate movement toward extreme views, and looping patterns can indicate repeated exposure to sensitive themes. This supports more consistent policy enforcement.
- *Incident investigation and review:* When a harmful recommendation outcome occurs, AMPH provides a way to identify which system behaviour contributed to the outcome. Teams can determine whether looping, escalation, network activity or rapid spread played a role and use this information during incident reviews and postmortems.
- *Targeted mitigation within recommendation systems:* Many safety interventions can be connected to specific AMPH mechanisms. Diversity controls can reduce looping; credibility signals can reduce the impact of low-quality engagement and rate limits can slow rapid spread. The taxonomy helps map interventions to the mechanism that requires adjustment.
- *Coordination across platform teams:* Safety work often involves policy, engineering, integrity and data teams. AMPH gives these groups a shared vocabulary for describing system behaviour. This supports clearer communication and more consistent decision-making.

5.5 Transparency and Audit Limitations

The application of AMPH in external audits and research depends on the quality and availability of platform data. Although the taxonomy identifies clear system behaviours, access to detailed internal logs is uneven.

- *Limited access to ranking information:* External researchers generally do not have access to internal ranking scores, feature weights or prediction signals. This limits the ability to verify exactly how the system prioritised certain items and makes some mechanisms visible only through indirect observation.
- *Incomplete exposure reconstruction:* Mechanisms such as looping, escalation and cross modal reinforcement require analysis of recommendation sequences. Platforms may not store complete sequences or may produce incomplete logs for external audits. Personalisation, item removal and missing timestamps make reconstruction difficult.
- *Constraints across formats and languages:* Cross modal detection depends on understanding how text, images, audio and video are linked through embedding models. These models are proprietary and rarely disclosed. Similar issues occur in languages where safety tooling and transparency resources are limited.
- *Variation in audit practices:* Platforms differ in the type and granularity of data they provide for research under regulatory or voluntary access programmes. This makes it difficult to compare audits across platforms and reduces consistency in external evaluations.
- *Limited visibility of feedback loops:* Recommendation systems update based on user actions, but external auditors generally cannot observe model retraining, decay functions or real time adjustments. This makes it difficult to assess the full feedback cycle that supports amplification.

These limitations do not prevent AMPH from being used in audits, but they influence how precisely mechanisms can be identified. They also indicate a need for more standardised data access frameworks and clearer disclosure practices.

Effective auditing also depends on the retention of structured logs that capture key elements of the recommendation process. At minimum, platforms should maintain timestamped recommendation sequences, engagement signals, and model versioning records so that exposure pathways can be reconstructed during regulatory or independent audits.

5.6 Ethical Considerations and Researcher Safety

Research on algorithmic amplification often involves reviewing sensitive, distressing or harmful content. It may also require examining material related to self-harm, hostility, misinformation or

extremist narratives. For these reasons, audits and evaluations based on the AMPH taxonomy should follow clear ethical guidelines. Researchers and analysts should have access to appropriate support and should be trained in handling harmful material. Access to sensitive content should be limited to those who need it for the assessment and should be logged and reviewed. Where possible, content should be blurred or summarised rather than viewed directly. Studies involving recommendation systems should avoid identifying individual users. Only aggregated or de-identified data should be used. When working with user-generated content, care should be taken to avoid exposing personal information or reproducing harmful material in publications. Research teams should follow institutional review processes when analysing potentially harmful content or working with vulnerable populations. Clear protocols should be in place for storing, accessing and deleting material used in audits. When external researchers work with platform-provided data, the terms of access should be transparent and should minimise unnecessary exposure to harmful content. These considerations support researcher wellbeing, reduce unnecessary exposure to harmful material and ensure that assessments based on AMPH follow responsible research practices.

6. Operational Audit Checklist

This section summarises the key steps and data required to assess amplification patterns using the AMPH taxonomy. It is intended for use in platform audits and Trust and Safety evaluations.

6.1 Data required

Auditors should collect the following information where available:

i. sequences of recommended items with timestamps
ii. engagement logs including views, watch time, reactions and shares
iii. metadata for each item such as topic, format and creator
iv. network level interaction patterns and engagement distributions
v. information linking text, image, audio and video formats

6.2 Identifying the mechanism

Audit teams should review the data to determine which AMPH mechanism is present:

i. repetition of similar items indicates looping
ii. movement toward stronger or more intense themes indicates escalation
iii. sudden engagement spikes or coordinated activity indicate network driven boosting
iv. rapid spread or high engagement velocity indicates artificial virality
v. repeated exposure for specific groups indicates targeted intensification
vi. matching narratives across formats indicate cross modal reinforcement

6.3 Analysing exposure trajectories

After identifying the mechanism, auditors should examine:

i. how content changed across the recommendation sequence
ii. shifts in topic, tone or risk level
iii. the presence of reinforcement patterns or narrowing content clusters
iv. whether vulnerable users were more exposed

6.4 Assessing harm

Auditors should map the exposure trajectory to relevant harm categories, such as:

i. misinformation
ii. self-harm or mental health
iii. body image pressure
iv. hate or hostility
v. political or civic distortion

6.5 Reporting findings

Audit reports should include:

i. the mechanism identified
ii. the evidence supporting the assessment

iii. the affected user groups
iv. a summary of the exposure pathway
v. a severity assessment
vi. any data limitations or missing logs

This checklist provides a consistent process for applying AMPH in real recommendation environments without requiring full access to internal ranking systems.

7. Key Takeaways

This white paper presents AMPH, a mechanism-based framework that explains how recommendation systems create harmful exposure patterns. Existing regulatory and research models classify harm by content or user outcome but do not describe the underlying system behaviours that shape how users encounter information. AMPH addresses this by identifying six amplification mechanisms and linking each one to observable signals that can be used in audits, risk assessments, and system evaluations. The framework provides a structured vocabulary for analysing how repetition, progression, coordinated activity, spread velocity, targeted exposure, and cross-format propagation influence user trajectories. It offers regulators a practical way to interpret systemic risk obligations and gives platforms a method for monitoring the design-level processes that contribute to harmful exposure. AMPH is operationally feasible because it relies on signals that most large platforms already collect, including recommendation sequences, engagement logs, and content similarity measures. These signals are organised into mechanism-based indicators, allowing the framework to function as an audit layer without requiring changes to core ranking architectures. The mitigation strategies outlined in this paper show that amplification risks can be addressed through adjustments to ranking logic, alignment with regulatory expectations, and integration into routine engineering workflows. These interventions operate at the system level and do not depend on content categorisation or subjective judgments about individual items. By focusing on mechanism-level behaviour rather than topic-specific harm categories, AMPH supports consistent analysis across domains, user groups, and platform environments. Because AMPH focuses on system behaviour rather than platform-specific taxonomies, it enables cross-platform comparisons of amplification risks and supports more standardised evaluation across different recommender architectures. It provides regulators, researchers, and engineers with an adaptable tool for understanding how amplification contributes to online risk. AMPH does not assess user intent or evaluate content norms; its scope is limited to identifying system-level patterns that shape exposure.

As recommendation systems continue to shape digital experiences at global scale, mechanism-based frameworks such as AMPH will become increasingly important for supporting accountable, transparent, and safety-oriented platform design.

Suggested Citation

Kumar, A., & Sangwan, S. R. (2026). *AMPH: A Taxonomy of Amplification Mechanisms in Algorithmic Recommendation Systems.* Technical White Paper.

References

Ahmed, Saifuddin, Kokil Jaidka, Vivian Hsueh Hua Chen, Mengxuan Cai, Anfan Chen, Claire Stravato Emes, Valerie Yu, and Arul Chib. "Social media and anti-immigrant prejudice: a multi-method analysis of the role of social media use, threat perceptions, and cognitive ability." Frontiers in psychology 15 (2024): 1280366.

Ahn, Jae-wook, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. "Open user profiles for adaptive news systems: help or harm?." In Proceedings of the 16th international conference on World Wide Web, pp. 11-20. 2007

Anastasi, Selenia, Tim Fischer, Florian Schneider, and Chris Biemann. "IDA-Incel Data Archive: a multimodal comparable corpus for exploring extremist dynamics in online interaction." 14–15 September 2023, University of Mannheim, Germany: 23.

Anastasi, Selenia. "Misogyny beyond borders: A Cross-Linguistic Corpus Assisted Analysis of Transnational Incel Communities." (2025).

Andrikopoulou, Elisavet, Nicholas Talam, and Aikaterini Kanta. "MedTok or MythTok? Classifying health misinformation on TikTok with AI." In EFMI Special Topic Conference 2025, pp. 67-71. IOS Press, 2025

Ash, Elliott, Sergio Galletta, and Giacomo Opocher. "BallotBot: Can AI Strengthen Democracy?." Available at SSRN (2025).

Astuti, Wulan Tri, and Ari Mulyadi. "Selected Political Hoax Cases on Social Media: Multimodal in Forensic Linguistics." Policy & Governance Review 9, no. 1 (2025): 44-63.

Benkler, Yochai, Robert Faris, and Hal Roberts. Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press, 2018.

Boichak, Olga, Sam Jackson, Jeff Hemsley, and Sikana Tanupabrungsun. "Automated diffusion? Bots and their influence during the 2016 US presidential election." In International conference on information, pp. 17-26. Cham: Springer International Publishing, 2018.

Buckley, Sophie. "Guns, Policing, and Crime: How the Internet and Social Media Shape Public Opinions." Master's thesis, University of Arkansas, 2025.

Burrell, Jenna. "How the machine ‘thinks’: Understanding opacity in machine learning algorithms." Big data & society 3, no. 1 (2016): 2053951715622512.

Büyüksağnak, Mert. "The effect of Spotify playlists on the music consumption habits of music consumers." (2021).

Cao, Shixiang, Zhigang Qiu, Xinyi Shao, and Ke Song. "How Financial Influencers Rise: Performance Following and Social Transmission Bias." Available at SSRN 4852857 (2025).

Carcelén-García, Sonia, María José Narros-González, and María Galmes-Cerezo. "Digital vulnerability in young people: gender, age and online participation patterns." International Journal of Adolescence and Youth 28, no. 1 (2023): 2287115.

Cataldo, Ilaria, Ilaria De Luca, Valentina Giorgetti, Dorotea Cicconcelli, Francesco Saverio Bersani, Claudio Imperatori, Samira Abdi, Attilio Negri, Gianluca Esposito, and Ornella Corazza. "Fitspiration on social media: Body-image and other psychopathological risks among young adults. A narrative review." Emerging Trends in Drugs, Addictions, and Health 1 (2021): 100010.

Chadee, Derek, Sven Smith, and Christopher J. Ferguson. "Murder she watched: Does watching news or fictional media cultivate fear of crime?." Psychology of popular media culture 8, no. 2 (2019): 125.

Chee, Jerry, Shankar Kalyanaraman, Sindhu Kiranmai Ernala, Udi Weinsberg, Sarah Dean, and Stratis Ioannidis. "Harm mitigation in recommender systems under user preference dynamics." In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 255-265. 2024.

Chen, Mengyuan. "Angry Friends, Happy Crowd: How the Emotionally Charged Content Influences Online Behavior." (2023)

Chen, Shijing, Usman Naseem, Imran Razzak, and Flora Salim. "Unveiling misogyny memes: A multimodal analysis of modality effects on identification." In Companion Proceedings of the ACM Web Conference 2024, pp. 1864-1871. 2024.

Chen, Yueying, and Hongliang Chen. "Exploring the mechanism of adult users’ cyber-aggression against adolescents: The roles of online communication, age group identity, and online moral disengagement." Journal of Interpersonal Violence 40, no. 9-10 (2025): 1979-2005.

Chuai, Yuwei, and Jichang Zhao. "Anger can make fake news viral online." Frontiers in Physics 10 (2022): 970174.

Coroner’s Court of North London. 2022. Inquest into the Death of Molly Russell: Findings and Prevention of Future Deaths Report. London: Judiciary of England and Wales.

Covington, Paul, Jay Adams, and Emre Sargin. "Deep neural networks for youtube recommendations." In Proceedings of the 10th ACM conference on recommender systems, pp. 191-198. 2016.

Das, Shanti. "How TikTok bombards young men with misogynistic videos." The Guardian (2022).
de-Lima-Santos, Mathias-Felipe, and Wilson Ceron. "Coordinated amplification, coordinated inauthentic behaviour, orchestrated campaigns: A systematic literature review of coordinated inauthentic content on online social networks." Mapping lies in the global media sphere (2023): 165-184.

Di Marco, Niccolò, Matteo Cinelli, and Walter Quattrociocchi. "Infodemics on YouTube: Reliability of content and echo chambers on COVID-19." arXiv preprint arXiv:2106.08684 (2021).

Diberardino, Nathalie, Clair Baleshta, and Luke Stark. "Algorithmic harms and algorithmic wrongs." In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1725-1732. 2024.

Eriksson Krutrök, Moa, and Simon Lindgren. "Social media amplification loops and false alarms: Towards a Sociotechnical understanding of misinformation during emergencies." The Communication Review 25, no. 2 (2022): 81-95.

eSafety Commissioner. 2021.
Safety by Design: Overview and Framework. Canberra: Australian Government. https://www.esafety.gov.au/safety-by-design.

Esposito, Eleonora, and Sole Alba Zollo. "“How dare you call her a pig, I know several pigs who would be upset if they knew” A multimodal critical discursive approach to online misogyny against UK MPs on YouTube." Journal of language aggression and conflict 9, no. 1 (2021): 47-75.

European Commission. 2022. Regulation (EU) 2022/2065 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%3A32022R2065

European Union. 2022.
Regulation (EU) 2022/2065 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%3A32022R2065.

Feldman, Dalit Ken-Dror, and Yifat Nahmias. "From bots to ballots: Democratic integrity in the era of digital manipulation." Minn. JL Sci. & Tech. 26 (2024): 228

Furini, Marco. "Recommendation Systems: Issues, Challenges and Regulations." In 2024 IEEE Symposium on Computers and Communications (ISCC), pp. 1-6. IEEE, 2024.
Gangavarapu, Rajendra. "Unmasking Deepfakes: Navigating Challenges and Solutions in the Age of AI-Driven Manipulation." In Mastering AI Governance: A Guide to Building Trustworthy and Transparent AI Systems, pp. 57-61. Cham: Springer Nature Switzerland, 2025.

Glowacki, Luke, Alexander Isakov, Richard W. Wrangham, Rose McDermott, James H. Fowler, and

Nicholas A. Christakis. "Formation of raiding parties for intergroup violence is mediated by social network structure." Proceedings of the National Academy of Sciences 113, no. 43 (2016): 12114-12119.

Goswami, Arijit. "Recommendation system as a Social Determinant of Health." Digital Society 3, no. 2 (2024): 28.

Haddon, Leslie, and Sonia Livingstone. "The meaning of online problematic situations for children: The UK report." EU Kids Online (2014).

Heiss, Raffael, Johannes Knoll, and Jörg Matthes. "Pathways to political (dis-) engagement: Motivations behind social media use and the role of incidental and intentional exposure modes in adolescents' political engagement." Communications 45, no. s1 (2020): 671-693.

Heřmanová, Marie. "“We are in control”: Instagram influencers and the proliferation of conspiracy narratives in digital spaces." Slovenský národopis 70, no. 3 (2022): 349-368.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. "Examining the consumption of radical content on YouTube." Proceedings of the national academy of sciences 118, no. 32 (2021): e2101967118.

Huszár, Ferenc, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. "Algorithmic amplification of politics on Twitter." Proceedings of the national academy of sciences 119, no. 1 (2022): e2025334119.

Imtiaz, Nadra. "Impact of still advertisements in advancing the notion of body image among adolescents: a multimodal discourse analysis." (2023).

Karatzoglou, Alexandros, Linas Baltrunas, and Yue Shi. "Learning to rank for recommender systems." In Proceedings of the 7th ACM Conference on Recommender Systems, pp. 493-494. 2013.

Kessler, Sabrina Heike, and Edda Humprecht. "COVID-19 misinformation on YouTube: An analysis of its impact and subsequent online information searches for verification." Digital Health 9 (2023): 20552076231177131.

Kleemans, Mariska, Serena Daalmans, Ilana Carbaat, and Doeschka Anschütz. "Picture perfect: The direct effect of manipulated Instagram photos on body image in adolescent girls." Media Psychology 21, no. 1 (2018): 93-110

Kuo, Wen-Hung. "The Multimodality of Generative AI Internet Memes in the 2024 US Presidential Election." PhD diss., Robert Morris University, 2025.

Kużelewska, Elżbieta, and Mariusz Tomaszuk. "Rise of conspiracy theories in the pandemic times." International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique 35, no. 6 (2022): 2373-2389.

Lee, Yen-I., Di Mu, Ying-Chia Hsu, Bartosz W. Wojdynski, and Matt Binford. "Misinformation or hard to tell? An eye-tracking study to investigate the effects of food crisis misinformation on social media engagement." Public Relations Review 50, no. 4 (2024): 102483.
Li, Run, and Xiang Huang. "Visual (Mis) Representations on Sports and Weight Loss: A Multimodal Critical Discourse Study of Images in Official Health Posts on WeChat in China." Health Communication (2025): 1-13.

Loeb, Stacy, Aisha T. Langford, Marie A. Bragg, Robert Sherman, and June M. Chan. "Cancer misinformation on social media." CA: A Cancer Journal for Clinicians 74, no. 5 (2024): 453-464.

Magelinski, Thomas, Lynnette Ng, and Kathleen Carley. "A synchronized action framework for detection of coordination on social media." Journal of Online Trust and Safety 1, no. 2 (2022).

Marwick, Alice, and Rebecca Lewis. 2017. Media Manipulation and Disinformation Online. New York: Data & Society Research Institute.

McKelvey, Fenwick, and Elizabeth Dubois. "Computational propaganda in Canada: The use of political bots." (2017).

Meisner, Colten. "Networked responses to networked harassment? Creators’ coordinated management of “hate raids” on Twitch." Social Media+ Society 9, no. 2 (2023): 20563051231179696.

Micallef, Nicholas, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. "Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 651-662. 2022.

Miettinen, Marko. "The Impact of Social Media on Investors' Risk-Taking." (2025).Singh, Guruansh, Swayam Mahajan, and Rajneet Kaur. "The Dual-Edged Influence of Social Media in Financial Decisions." Journal of Finance and Accounting 5, no. 2 (2025): 37-48.

Misnawati, Desy, Bosya Perdana, Sunda Ariana, Novita Damayanti, Mohamad Rakhmansyah, and Ardirra Yolandita. "AI-Driven Content Filtering on Instagram and Its Impact on Youth Lifestyle and Interaction." In 2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT), pp. 1-6. IEEE, 2025.

Müller, Martin, and Matthias C. Kettemann. "European approaches to the regulation of digital technologies." Hannes Werthner· Carlo Ghezzi· Jeff Kramer· Julian Nida-Rümelin· Bashar Nuseibeh· Erich Prem· (2024): 623.

Nannini, Luca. "From Categorical to Contextual: Interpreting High-Risk Classification for Profiling-Based AI Recommender Systems in the EU AI Act." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 2, pp. 1836-1847. 2025.

Nera, Kenzo, Myrto Pantazi, and Olivier Klein. "“These are just stories, Mulder”: Exposure to conspiracist fiction does not produce narrative persuasion." Frontiers in Psychology 9 (2018): 684.

Ng, Yee Man Margaret, Katherine Hoffmann Pham, and Miguel Luengo-Oroz. "Exploring YouTube’s recommendation system in the context of COVID-19 vaccines: Computational and comparative analysis of video trajectories." Journal of medical Internet research 25 (2023): e49061.

Nieman Lab. 2021. “More Internal Documents Show How Facebook’s Algorithm Prioritized Anger — and the Posts That Triggered It.” Nieman Journalism Lab, October 2021. https://www.niemanlab.org/2021/10/more-internal-documents-show-how-facebooks-algorithm-prioritized-anger-and-posts-that-triggered-it/

OECD. 2021. OECD Digital Education Outlook 2021: Pushing the Frontiers with AI, Blockchain and Robots. Paris: OECD Publishing. https://doi.org/10.1787/589b283f-en

Ofcom. 2023. Online Safety Act: Duties, Risk Assessments and Regulatory Guidance. London: Ofcom. https://www.ofcom.org.uk/online-safety

Oh, Onook, Manish Agrawal, and H. Raghav Rao. "Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises." MIS quarterly (2013): 407-426.

Papadamou, Kostantinos, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. "" How over is it?" Understanding the Incel Community on YouTube." Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW2 (2021): 1-25.

Papadamou, Kostantinos, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. "Understanding the incel community on youtube." (2020).

Paris, Britt, and Joan Donovan. "Deepfakes and cheap fakes." United States of America: Data & Society 1 (2019).

Park, Han Woo, and Sejung Park. "The filter bubble generated by artificial intelligence algorithms and the network dynamics of collective polarization on YouTube: the case of South Korea." Asian Journal of Communication 34, no. 2 (2024): 195-212.

Pummerer, Lotte, and Kai Sassenberg. "Conspiracy theories in times of crisis and their societal effects: Case “corona”." Societal Effects Of Corona Conspiracy Theories (2020).

Ratwatte, Priyanjali, and Emily Mattacola. "An exploration of ‘fitspiration’content on YouTube and its impacts on consumers." Journal of health psychology 26, no. 6 (2021): 935-946

Raza, Shaina, and Chen Ding. "News recommender system: a review of recent progress, challenges, and opportunities." Artificial Intelligence Review 55, no. 1 (2022): 749-800.

Regehr, Kaitlyn, Caitlin Shaughnessy, Minzhu Zhao, Idil Cambazoglu, Alfie Turner, and Nicola Shaughnessy. "Normalizing toxicity: the role of recommender algorithms for young people's mental health and social wellbeing." Frontiers in Psychology 16 (2025): 1523649.

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. "Auditing radicalization pathways on YouTube." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 131-141. 2020.

Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Introduction to recommender systems handbook." In Recommender systems handbook, pp. 1-35. Boston, MA: springer US, 2010.

Rothut, Sophia, Heidi Schulze, Julian Hohner, and Diana Rieger. "Ambassadors of ideology: A conceptualization and computational investigation of far-right influencers, their networking structures, and communication practices." New Media & Society 26, no. 12 (2024): 7120-7147

Sáez-Linero, Carolina, and Mònika Jiménez-Morales. "Young, lower-class, and algorithmically persuaded: exploring personalized advertising and its impact on social inequality." Communication & Society (2025).

Sauerwein, Jessica. "Scrolling Through Recovery: Youth-Centred Priorities for Supporting Eating Disorder Recovery in the TikTok Era." (2025).
Stecula, Dominik A., and Mark Pickup. "Social media, cognitive reflection, and conspiracy beliefs." Frontiers in Political Science 3 (2021): 647957.

Stohr, Edward A., and Sivakumar Viswanathan. "Recommendation systems: Decision support for the information economy." (1998).

UK Parliament. 2023. Online Safety Act 2023. London: The Stationery Office. https://www.legislation.gov.uk/ukpga/2023/50/enacted

Ukkola, Anssi. "The influence of social media recommendation algorithms on opinion polarization." (2025).

United States Senate. 2021. Protecting Kids Online: Testimony from Facebook Whistleblower Frances Haugen. Hearing before the Subcommittee on Consumer Protection, Product Safety, and Data Security. Washington, DC: U.S. Senate.

University College London. 2024. "Social Media Algorithms Amplify Misogynistic Content to Teens." UCL News, February 5, 2024. https://www.ucl.ac.uk/news/2024/feb/social-media-algorithms-amplify-misogynistic-content-teens.

Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news." Social media+ society 6, no. 1 (2020): 2056305120903408.

van der Breggen, Max Matthias, João Gonçalves, and David Boeren. "Polarization by recommendation: analyzing YouTube's polarization dynamics around Dutch political parties." Journal of Information Technology & Politics (2025): 1-15.

White House Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. Washington, DC: Executive Office of the President of the United States. https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/.

Whittaker, Joe, Seán Looney, Alastair Reed, and Fabio Votta. "Recommender systems and the amplification of extremist content." Internet Policy Review 10, no. 2 (2021).

Wilson, Anna, Seb Wilkes, Yayoi Teramoto, and Scott Hale. "Multimodal analysis of disinformation and misinformation." Royal Society Open Science 10, no. 12 (2023): 230964.

Yang, Aimei, and Dmitri Williams. "Quantifying Networked Influence: How Much Do Disinformation Spreaders' Networks Drive Their Public Engagement Outcomes?." Social Media+ Society 10, no. 3 (2024): 20563051241265865.

Yang, Aimei, and Dmitri Williams. "Quantifying Networked Influence: How Much Do Disinformation Spreaders' Networks Drive Their Public Engagement Outcomes?." Social Media+ Society 10, no. 3 (2024): 20563051241265865.

Yu, Yifan, Shan Huang, Yuchen Liu, and Yong Tan. "Emotions in online content diffusion." Information Systems Research (2025).

Zhang, Yini, and Ruixue Lian. "How Right-Wing Media, Twitter, Facebook, and YouTube Circulated Antipathy and Threat Cues About Immigrants: A Cross-Media and Cross-Platform Approach." Mass Communication and Society (2025): 1-20.

Disclaimer

This white paper is intended for research, policy, educational, and public interest purposes. The AMPH framework is designed to support analysis of system-level amplification behaviours within recommendation systems and does not assess individual users, assign intent, or evaluate the legality of specific content or platforms. References to platforms, incidents, or public cases are included solely for illustrative and analytical purposes based on publicly available information.