Exploring the Formation Mechanisms and Preventive Strategies of Cognitive-Degrading Content on the Xiaohongshu Platform

Weixiang Gan[1], Mengfei Xiao[1, *], Naiqian Zhang[2]
[1]Graduate School of Business, SEGi University, Petaling Jaya, Selangor 47810, Malaysia
[2]Chongqing University of Arts and Sciences, Chongqing 402160, China
*Corresponding email: 282584787@qq.com

Abstract

Against the backdrop of social media becoming deeply embedded in everyday life and continuously reshaping cognitive structures and consumer decision-making pathways, Xiaohongshu has evolved into a major platform through which users obtain lifestyle advice and value judgments. However, beneath the appearance of a flourishing content ecosystem, a form of manipulative information characterized by high concealment and systematic harm has been spreading. This study conceptualizes such information as “cognitive-degrading content”, referring to content that imitates scientific discourse, logical reasoning, authoritative endorsement, and moral appeals as rational symbolic systems to construct an argumentation shell that appears rigorous but is in fact fallacious, thereby unconsciously weakening users’ critical thinking and evidence evaluation abilities. Based on a structured analysis of content forms, this study identifies five typical types of cognitive-degrading content: pseudo-scientific marketing and terminology accumulation, fallacious evaluations and the inducement of cognitive shortcuts, fabricated authority and selective quotation, extreme emotional manipulation accompanied by identity binding and conspiratorial attribution, and immersive fictional realities. To explain the mechanisms through which such content is continuously produced and algorithmically amplified, this study constructs a dual-level analytical framework spanning the micro and macro levels. At the micro level, dual-process theory is introduced to reveal how cognitive-degrading content systematically activates intuitive judgment associated with System 1 while suppressing rational scrutiny associated with System 2. At the macro level, drawing on perspectives from the attention economy and the political economy of algorithms, the study demonstrates that recommendation logics centered on completion rates and interaction volume structurally reward emotionalized content with low cognitive cost and high dramatic intensity, thereby forming a self-reinforcing diffusion cycle. Based on this diagnostic analysis, the study further proposes a multi-actor collaborative governance pathway, including algorithmic restructuring and the introduction of quality weighting at the platform level, benefit constraints and professional boundary norms at the creator level, and the enhancement of media literacy and participatory co-governance mechanisms at the user level. The theoretical contribution of this study lies in disentangling cognitive-degrading content from explanations grounded in individual moral deviance and reconstructing it as a structural information ecology problem jointly shaped by platform institutions, creators’ rational choices, and users’ cognitive constraints, thereby offering systematic policy implications for information governance on lifestyle-oriented platforms.

Keywords

Cognitive-degrading content, Xiaohongshu, Dual-process theory, Attention economy, Algorithmic governance, Information ecology

Introduction

Social media has become deeply embedded in everyday life and continues to reshape cognitive structures. Within this context, the Xiaohongshu platform has evolved into an important public space for hundreds of millions of users. Built around the core concept of sharing authentic experiences, it provides references for lifestyle advice, consumer decision-making, and value judgments [1]. However, beneath the apparent prosperity of the platform's content ecosystem, a form of information characterized by high concealment and systematic harm has been rapidly spreading, namely what can be described as cognitive-degrading content. Unlike traditional forms of misinformation or crudely produced low-quality content, cognitive-degrading content does not simply represent inferior expression but constitutes a manipulative information form in which an irrational core is wrapped in a rational façade [2]. Its defining feature lies in the deliberate imitation of cognitive symbolic systems, such as scientific discourse, logical reasoning, authoritative endorsement, and moral appeals. Through these means, it constructs seemingly rigorous yet fundamentally fallacious argumentative structures, subtly weakening users' critical thinking abilities. This erosion ultimately serves purposes like traffic harvesting, commercial monetization, or even value manipulation [3]. What is particularly concerning is that such content rarely appears in overtly anti-intellectual or vulgar forms. Instead, it is highly disguised as legitimate formats such as popular science explanations, in-depth evaluations, industry insider information, expert advice, or even expressions of social justice, leading users to gradually internalize its preset positions and conclusions while experiencing a sense of usefulness, resonance, and participation, thus forming cognitive judgments that appear autonomous but are in fact guided [4].

From the perspective of actual operational mechanisms, the diffusion of cognitive-degrading content on the Xiaohongshu platform has exhibited clear structural characteristics. On the one hand, under algorithm-driven recommendation logics, content featuring counter-intuitive claims, intense conflict, and high emotional arousal is more likely to gain exposure, objectively incentivizing creators to continuously generate cognitive stimulation and oppositional topics [5]. On the other hand, driven by fierce competition for traffic and pressures for commercial monetization, some creators actively transgress professional boundaries and ethical norms by exaggerating risks, manufacturing anxiety, fabricating authority, and selectively presenting facts in order to capture attention resources [6]. At the same time, insufficient media literacy and scientific rationality among some users make them more susceptible to carefully packaged pseudo-professional content, while existing platform review and regulatory mechanisms still lag in identifying emerging forms of manipulative content [7,8]. The superposition of these factors has gradually transformed cognitive-degrading content into a structural ecological problem jointly shaped by platform mechanisms, creator strategies, and user cognitive structures, rather than a sporadic phenomenon attributable to individual creators.

The adverse consequences triggered by this phenomenon have continued to manifest across multiple levels. At the micro level, misleading consumption advice and pseudo-scientific health information directly harm users' economic interests and physical and mental well-being. At the meso level, spaces for professional knowledge production and rational discussion are increasingly compressed, placing genuinely competent creators at a disadvantage in traffic competition. At the macro level, this phenomenon continuously erodes the foundations of public rationality, weakens trust in science, expertise, and institutional authority, amplifies group-based emotional polarization, and ultimately undermines the trust capital and consensus foundations upon which social cooperation depends. It is therefore evident that cognitive-degrading content constitutes a systematic and implicit erosion mechanism spanning cognitive styles, emotional judgment, and social values. Against this background, moral critique based solely on individual responsibility or content quality is clearly insufficient. There is an urgent need to systematically examine the generative logic, diffusion pathways, and efficacy mechanisms of cognitive-degrading content from multiple dimensions, including platform algorithmic mechanisms, creator incentive structures, and user psychological cognition, and to further propose operational governance pathways and institutional response strategies. Consequently, conducting an in-depth theoretical analysis of this phenomenon not only carries significant academic value but also holds pressing practical relevance for optimizing platform governance, enhancing public media literacy, and safeguarding a healthy information ecology. This study is undertaken precisely in response to this problem awareness.

Categories and potential harms of cognitive-degrading content

So-called cognitive-degrading content is by no means a vague or purely rhetorical label of criticism, but rather a category with clear and identifiable forms. Based on the core characteristics of its manipulative techniques, such content can be systematically classified into the following five types.

Pseudo-scientific marketing and professional discourse stacking

Among various forms of cognitive-degrading content, pseudo-scientific marketing content is particularly prevalent and poses significant risks. It typically targets high-anxiety domains closely related to individual well-being, such as health, beauty, and parenting, and completes a closed loop from anxiety construction to commercial monetization through the systematic appropriation and oversimplification of scientific discourse. Producers of such content often adopt an unequivocal tone to claim that certain products can “fade pigmentation within seven days”, “eliminate ten years of intestinal toxins” or “unlock children’s potential”, while consistently revolving around three core absences: the absence of authoritative sources, manifested in the deliberate avoidance of peer-reviewed literature. The absence of empirical evidence, reflected in the lack of verifiable clinical trials or reports. And the absence of mechanism-based explanations, indicated by the inability to clearly articulate how the claimed effects are produced. For example, a product promoted as “graphene energy wellness socks” may be accompanied by short videos filled with dazzling terms such as “quantum microcirculation” and “far-infrared resonance”, with influencers repeatedly demonstrating sensations of warmth after wearing them yet failing to provide even the most basic medical device registration number or third-party material testing report. The essence of this strategy lies in reducing complex scientific knowledge systems into a handful of technology-sounding labels, exploiting public reverence for expertise and anxiety about personal health, and packaging purely commercial promotion as a supposedly benevolent form of knowledge provision, thereby accelerating the diffusion of pseudo-scientific narratives under the guise of health information and generating tangible real-world risks [9].

Closely intertwined with this and often operating in tandem is professional discourse stacking content. This type does not necessarily involve direct product promotion, yet its core objective similarly lies in constructing an unquestionable aura of authority through language. This paves the way for subsequent persuasion, whether in terms of viewpoints or consumption. Its defining feature is the inflation of terminology, whereby creators freely extract sophisticated terms from disparate fields such as medicine, psychology, quantum physics, and traditional philosophy, and assemble them in a disordered manner that disregards logic and context. For instance, in a short video on anti-aging, a creator may rapidly introduce concepts such as telomerase activation, mitochondrial energy repair, quantum-level penetration, and five-element balance of qi and blood, without offering any definitional clarification or logical linkage. This linguistic strategy produces a state of cognitive vertigo, in which audiences instinctively equate incomprehension with profundity and mistake expressive confusion for knowledge density, thereby becoming more inclined to judge the information as highly credible. Experimental studies have shown that technical language and high terminology density can alter audiences’ assessments of information reliability by influencing processing fluency and perceived credibility and may even make unreliable information appear more authentic on the surface [10]. The deeper harm lies in its distortion of the very definition of professionalism. Genuine expertise should be grounded in clear concepts, rigorous methods, and verifiable evidence, whereas such content substitutes professionalism with flamboyant linguistic style and performative posture. Research on the role of terminology in information credibility further indicates that under certain conditions, technical terms can directly elevate perceived credibility, thereby amplifying the cognitive bias of finding information more authoritative precisely because it is difficult to understand [11]. Over time, the public's ability to discern genuine expertise becomes blunted, potentially giving rise to a paradoxical cognition whereby the less one understands, the more convincing the content appears, further marginalizing serious and cautious knowledge production and dissemination in the competition for attention.

Fallacious testing and cognitive shortcut inducement

Fallacious testing content skillfully exploits the short-video medium's visual immediacy, staging simple, intuitive, and visually striking demonstrations to transform complex scientific or quality judgments into seemingly definitive on-site proof [12]. Its core trap lies in substituting scientific rigor with theatrical performance. A widely circulated example is the so-called antioxidant beverage test, in which creators pour different drinks into containers, add colorful reagents, stir them with the same utensil, and then declare the drink that fades fastest to possess the strongest antioxidant capacity. In this process, essential elements of scientific experimentation, such as variable control, reagent specificity, pH interference, and the mechanistic relationship between in vitro reactions and in vivo cellular processes, are deliberately or inadvertently ignored. The implicit message conveyed is that what you can visually comprehend constitutes the entirety of truth [13]. Such content essentially functions as a form of cognitive conditioning, encouraging audiences to settle for surface-level causal associations, gradually relinquish inquiries into underlying mechanisms, and equate trusting one's eyes with completing the act of thinking.

Closely related yet more insidious is cognitive shortcut inducement content. Rather than inviting viewers to observe experiments, it directly provides ready-made and ostensibly efficient universal judgment rules, aiming to compress all complex decisions into one-step operations. Common expressions include claims such as "when buying food, just check the first three ingredients", "three sentences to see through a person", or "this formula applies to all workplace problems". While superficially framed as practical methodologies, such content systematically encourages cognitive laziness [14]. It forces a multidimensional, dynamic, and context-dependent reality into a single static evaluative framework, completely disregarding individual differences and situational complexity. Prolonged exposure leads users to develop a reliance on template-based thinking, whereby problems are approached not through contextual analysis but through the mechanical application of memorized formulas. This not only simplifies thinking but fundamentally erodes individuals' willingness and capacity to perceive complexity, exercise independent judgment, and engage in deep reflection, immersing them in a false sense of mastery while distancing them from the richness and variability of the real world [15].

Fabricated authority and selective quotation

Fabricated authority content directly targets the trust foundations upon which modern societies operate, with its core mechanism involving the systematic theft of professional credibility. To rapidly establish an appearance of unquestionable trustworthiness, some creators resort to forging or impersonating authoritative identities, settings, and credentials as shortcuts [16]. The most common manifestation is symbolic deception through scenarios and attire, such as actors wearing white lab coats, standing in front of laboratory-like backdrops filled with microscopes and test tubes, and delivering explanations of the so-called cellular activation principles of a hair loss treatment serum in a calm and professional tone, despite their actual identity being that of a commercial influencer. More advanced and deceptive forms involve technical forgery, such as merchants selling weight loss products who use image-editing software to fabricate top-tier hospital examination reports, handwritten product names in the physician recommendation section, or manipulated screenshots of academic papers that appear to endorse the efficacy of certain supplements [17]. Such practices extend far beyond exaggerated advertising and constitute direct appropriation of the public credibility associated with medicine and science. Their deepest harm resembles the circulation of counterfeit currency within a monetary system, eroding trust mechanisms at a systemic level. As audiences repeatedly discover that white coats cannot be trusted and reports can be falsified, pervasive skepticism emerges, ultimately leading individuals to question genuine medical advice or authoritative scientific findings as well [18]. The depletion of societal trust caused by this process far exceeds the damage of isolated consumer fraud.

Closely related and more frequently encountered in everyday information streams is selective quotation content. Rather than fabricating authority outright, it selectively trims and decontextualizes authentic authoritative information to create the illusion that authority endorses a particular position, amounting to a legally permissible distortion. Typical practices include extracting a single conclusion sentence from a rigorous academic paper stating that a certain compound exhibits anti-inflammatory potential, while deliberately omitting crucial premises such as the study being limited to cell experiments and employing specific high-concentration dosages, or clipping expert statements made in contexts to support entirely unrelated arguments. For example, academic discussions by economists on how moderate debt may motivate individual effort are edited into viral videos suggesting that experts encourage young people to engage in debt-driven consumption. The subtlety and danger of this approach lie in the fact that each fragment presented is factually accurate, yet the composite narrative formed through selective assembly is highly misleading. Long-term exposure trains audiences to adopt a fragmented authority cognition, focusing solely on conclusions while neglecting background, methodology, and applicability boundaries. As a result, public understanding of knowledge becomes increasingly flattened, with diminishing concern for how research is designed or under what conditions conclusions hold, ultimately undermining the contextual awareness and critical inquiry upon which scientific thinking depends.

Extreme emotional manipulation, identity binding, and conspiratorial attribution

Extreme emotional manipulation represents the category of cognitive-degrading content with the highest moral risk. Its primary objective extends beyond product sales to constructing oppositional identity structures through the arousal of intense emotions, thereby achieving follower aggregation and sustained control [19]. Its operational pathway is highly standardized. It first identifies and amplifies social anxieties or moral indignation, such as educational pressure, class stratification, the plight of vulnerable groups, or patriotic sentiment. It then constructs simplistic attribution frameworks and immediately actionable solutions, for instance by staging scenes of exhausted low-income workers laboring late at night to promote narratives of the futility of education or the inherent injustice of academic credentials, only to ultimately channel audiences toward expensive self-improvement courses or educational planning services, thus distorting complex structural social problems into individual predicaments supposedly solvable through consumption [20]. Similarly, staged narratives involving struggling veterans or mistreated animals are used to evoke sympathy while directing viewers toward specific product links or unverified donation channels [21]. Such content never offers evidence-based policy analysis or constructive discussion, but instead instrumentalizes suffering and social contradictions, extracting emotional and financial resources while continuously corroding rational, moderate, and cooperative public discourse [22].

Closely intertwined with this is identity binding and group antagonism content, which constructs clear camp boundaries and transforms differences in opinion into moral opposition. Expressions such as “not supporting this means you are unpatriotic” or “not buying means you do not love your children” frame dissenting views not as debatable positions but as markers of stupidity, coldness, or hostility, thereby triggering emotional polarization, sidelining rational deliberation, and compressing public discourse into an arena of antagonistic emotions [23]. Further compounded is conspiratorial and black-box attribution content, which reduces complex social phenomena to claims of hidden manipulation by certain individuals, capital, or institutions, positioning creators as revealers of truth through rhetoric such as "secrets the platform does not want you to know" or "you have been deceived for too long". All counterevidence is reinterpreted as proof of deeper conspiracies, forming a self-sealing explanatory system [24]. In essence, this replaces evidence with speculation and analysis with emotion, encouraging audiences to abandon rational verification and indulge in the psychological gratification of perceived awakening, thereby further entrenching group antagonism and cognitive closure [25].

Immersive fictional reality

This type represents a dramatized mutation of content that has flourished on short-video platforms in recent years. Its defining feature lies in deliberately packaging entirely fictional plots as documentary-style narratives, leading audiences to mistakenly perceive them as real social events. Unlike traditional audiovisual works that clearly label dramatization, such content intentionally conceals its fictional nature, adopting techniques such as handheld filming, bystander perspectives, surveillance camera aesthetics, and non-professional actors to create a sense of immediacy and authenticity. For instance, videos depicting food delivery workers rescuing people at night only to be fired, female students bullied by landlords, or elderly individuals scavenging to support grandchildren circulate widely on platforms such as Xiaohongshu and Douyin. These clips often trigger waves of outrage and vigilantism in the comment sections, only to later be exposed as scripted performances [26]. In early 2024, the Douyin platform conducted a large-scale crackdown on accounts producing misery-themed staged content, including fabricated narratives of migrant workers being assaulted while demanding wages or single mothers selling blood to raise children, all officially confirmed to involve actors and fictional plots. The cognitive-degrading nature of such content does not lie in fiction itself, but in deliberately blurring the boundary between reality and performance, prompting audiences to emotionally immerse themselves and make moral judgments without verification. More concerning is the deep integration of such narratives with commercial conversion, such as suddenly inserting shopping links under the pretext of helping protagonists, guiding viewers to private messages for supposed sponsorship channels, or promoting so-called positive energy courses and life transformation camps, thereby completing a closed loop from emotional mobilization to commercial extraction [27]. This model essentially constitutes a form of immersive emotional fraud, in which intense psychological impact is first generated through narrative and monetization is achieved before audiences regain composure. Its deeper harm lies in the continuous depletion of public compassion, the entertainment-driven treatment of genuine social issues, and the emergence of emotional fatigue and compassion inflation, which ultimately weakens societal trust and support for genuinely vulnerable groups [28].

Antecedents of cognitive-degrading content on the xiaohongshu platform

To understand why cognitive-degrading content can be continuously generated within the platform ecosystem and attain wide diffusion, it is clearly insufficient to remain at the level of descriptive observation alone. Rather, it is necessary to return to the underlying mechanisms that sustain and propel this phenomenon. From a theoretical standpoint, the emergence of cognitive-degrading content involves both individual-level modes of cognitive processing and platform-level institutional and algorithmic structures. Accordingly, this study advances a dual-layer analysis across micro and macro levels. At the micro level, it introduces dual-process theory to examine how individuals' information processing is guided toward judgment pathways that require minimal cognitive expenditure. At the macro level, it draws on the attention economy and the political economy of algorithms to reveal how platform recommendation mechanisms structurally amplify emotionalised and simplified content. Through an integrated analysis along these two theoretical pathways, this study seeks to systematically elucidate the internal logic underpinning the production and diffusion of cognitive-degrading content.

A dual-process theory perspective

At the level of theoretical explanation, dual-process theory provides a central cognitive framework for understanding why cognitive-degrading content can achieve extensive diffusion in practice. The theory posits that human information processing is not a unitary rational process but instead relies on two distinct mechanisms operating in parallel. One is a fast, automatic system that depends on intuition and affective responses, commonly referred to as System 1. The other is a slow, deliberative system that requires cognitive effort and is commonly referred to as System 2. System 1 excels at rapid intuitive judgments, yet it is highly susceptible to heuristic cues and emotional influences. By contrast, System 2 plays the dominant role when individuals engage in complex logical reasoning and deep analysis, but it demands greater cognitive resources and time investment. Dual-process theory has become an influential direction within cognitive psychology and has received extensive empirical support in explaining judgment and decision-making. For instance, both theoretical and empirical syntheses examining the differentiated roles of System 1 and System 2 have clearly articulated the basic properties of these two processing routes and their interactive dynamics [29]. In everyday contexts, individuals are inclined to prioritise System 1 in order to conserve cognitive resources and are more likely to shift to System 2 only when they perceive that a problem is highly complex or entails substantial risk [30]. Dual-process theory therefore constitutes a key theoretical basis for explaining why audiences are readily influenced by information that is simplified, intuitive, and emotionally triggering. This cognitive architecture aligns closely with the diffusion patterns of cognitive-degrading content observed earlier.

Illustratively, promotional videos for products such as “graphene energy socks” or “quantum skincare sprays” that have repeatedly appeared on Douyin and Xiaohongshu since 2023 typically employ direct demonstrations such as “warmth in the soles immediately after wearing” or “instant firming after spraying”, prompting users to make judgments based on immediate bodily sensations, while few viewers further question whether the perceived heat is merely attributable to insulation or whether any medical evidence supports the claimed effects. Similarly, in so-called antioxidant beverage experiment videos, creators use colour changes to produce a visual conclusion regarding which drink is healthier, and comment sections often contain responses such as “so this is how you can tell” or “it is so intuitive”, with almost no questioning of variable control, reflecting System 1’s strong reliance on visual cues. Likewise, a group of misery-themed staged accounts that were banned in early 2024 produced content such as “a delivery rider saves someone but is fired” or “an elderly person scavenges to fund a grandchild’s education”. Without any explicit indication of fictionalisation, such videos rapidly triggered online outrage and even spontaneous doxxing, only later to be confirmed as scripted performances, yet during the early diffusion stage few users engaged in calm verification. Collectively, these cases suggest that whether in pseudo-scientific marketing, terminology stacking, fallacious demonstrations, fabricated authority, emotional manipulation, or fictionalised documentary-style narratives, the dominant expressive strategy revolves around rapid comprehension, immediate judgment, and emotional triggering, thereby systematically compressing audiences’ cognitive space. This implies that cognitive-degrading content does not merely happen to cater to users, but rather precisely embeds itself into humans’ default information processing routes, constructing a naturally friendly environment for diffusion at the level of cognitive mechanisms.

Building on this, the dual-process perspective further helps explain why cognitive-degrading content is continuously produced and replicated. From this vantage point, the creation of such content can be understood as a form of structural design targeting System 1. First, it generates high-arousal affect through the juxtaposition of fear and hope, for example when parenting influencers repeatedly claim that “a deficiency in this element will impair brain development”, and then immediately recommend expensive supplements, prompting parents to purchase rapidly under anxiety. Second, it constructs intuitive evidence through visualised demonstrations, such as the common “paper towel oil absorption test” or “flame comparison” used to judge food safety, while deliberately evading the logical discontinuity between such performances and the actual mechanisms of human metabolism. Third, it triggers authority heuristics through symbolic cues, as seen in multiple exposed cases of “lab-coat livestream selling” since 2023, where hosts recommend skincare devices or weight-loss products in clinic-like settings, despite lacking any medical background, yet audiences still comment that “what a doctor says must be reliable”. In addition, it reduces cognitive costs by offering cognitive shortcuts, such as workplace accounts repeatedly disseminating claims like “remember these three points and you will always win” or “this applies to all leaders”, compressing highly contextual social interactions into mechanical formulas. These operations converge on a single objective: to establish trust and trigger behaviour as quickly as possible, without allowing users to enter the more deliberative evaluation phase of System 2. Under these mechanisms, creators do not need to provide complete argumentation or authentic evidence. Instead, they merely need to construct emotional cues and perceptual stimuli sufficient to activate System 1, thereby inducing immediate trust and conversion behaviours.

Consequently, the production of cognitive-degrading content is not rooted in a simplistic disdain for audiences’ intelligence, but rather in the sophisticated exploitation of humans’ cognitive energy-saving tendencies. By systematically bypassing the logical reasoning and evidence scrutiny required by System 2, such content effectively outsources judgment to emotion, intuition, and symbolic authority, ultimately consolidating a cognitive inertia in which “if it is understandable, it must be true” and “if it feels right, it is right”. This mechanism also explains why even users with relatively high educational backgrounds may repeatedly be influenced by such content. The underlying issue is not a lack of capacity, but the predictable vulnerabilities embedded in human cognitive structures that are continuously and systematically activated and exploited.

Perspectives from the attention economy and the political economy of algorithms

At the macro-structural level, the attention economy provides a key explanatory framework for understanding why cognitive-degrading content is continuously amplified. The attention economy posits that in digital platform environments; user attention has itself become a scarce resource and a tradable commodity. The platform’s core profit logic is not anchored in content quality or social value, but in user dwell time, interaction frequency, and commercial conversion efficiency [31]. In other words, what platforms truly operate is not content per se, but users’ attention and time. This attention-conversion logic renders all interaction mechanisms quantifiable, from scrolling and liking to commenting and sharing, thereby feeding into algorithmic optimisation and forming a value assessment system centred on time and interactive behaviour [32]. Under this logic, recommendation algorithms do not optimise primarily for truthfulness or rigour, but instead dynamically adjust around behavioural indicators such as completion rates, likes, comment volumes, and shares. While such algorithms may increase participation, they simultaneously maximise attention capture by continuously strengthening the ranking of highly interactive content. This process constitutes an attention capture, feedback, and amplification cycle and represents a core driver of profitability within social platform business models [33].

In platform practice, this mechanism is highly visible. For example, videos themed around “quantum wellness” or “anti-ageing black technologies” that have remained popular on Douyin and Xiaohongshu since 2023 often construct health anxiety through sensational titles such as “if you do not supplement now, it will be too late” or “not supplementing is equivalent to chronic self-harm”, thereby forcing users to pause and watch. Pseudo-experiment videos such as “antioxidant beverage tests” or “paper towel oil absorption as a measure of food health” rely on colour changes and flame effects to create strong visual conflict and increase completion rates. Meanwhile, a batch of misery-themed staged accounts that were banned in early 2024 frequently reached tens of thousands of shares per single video prior to removal, with comment sections filled with emotionally charged interactions such as “it is heartbreaking” and “this must be shared.” These cases indicate that algorithms do not inherently care whether content is truthful, but rather whether it can capture attention. Cognitive-degrading content, by virtue of its high emotional intensity, low comprehension costs, and strong dramatic tension, aligns closely with algorithmic preferences, thereby repeatedly outperforming competitors within recommendation contests.

A deeper issue is that this structure does not require any subjective tolerance of fraud on the part of the platform. Simply through attention-centred metric design, it can naturally generate an ecological outcome in which inferior content crowds out superior content. In practice, many medical science communicators publicly note that a rigorous explanation of disease mechanisms often requires several minutes and yields low completion rates, whereas a single sentence such as “just eat this and you will be fine” can easily generate ten times the views. Similarly, within workplace content, systematic analyses of career pathways often attract little attention, whereas claims like “remember these three points and you will always win” repeatedly go viral. The algorithm does not necessarily reject rational content, yet under a selection logic that treats interaction data as the sole evaluative standard, complex expression is structurally disadvantaged and ultimately marginalised. This reflects the paradox highlighted by the political economy of algorithms. Platforms may appear value-neutral, yet their technical architectures implicitly shape the content ecosystem, and the rhetoric of neutrality can in fact conceal structural bias.

Within such an environment, creators quickly learn and replicate high-conversion templates, users develop conditioned clicking behaviours after repeatedly encountering similar content, and algorithms further amplify recommendations in response to interaction data. Over time, a positive feedback loop emerges in which algorithmic rewards incentivise creator imitation, imitation increases user click propensity, improved metrics further reinforce algorithmic distribution, and the cycle intensifies. Under this loop, cognitive-degrading content ceases to be episodic and becomes institutionally embedded within the platform ecology, evolving into a mainstream content form. As a result, public information spaces gradually tilt toward emotionalisation, simplification, and dramatization, constituting a deep structural problem of information degradation.

Preventive pathways and governance strategies for cognitive-degrading content on the Xiaohong Shu platform

Building on the preceding systematic analysis of the typological structure, generative mechanisms, and cognitive–institutional roots of cognitive-degrading content, it becomes evident that this phenomenon is not merely a matter of moral deviance on the part of individual creators. Rather, it constitutes a structural ecological problem jointly shaped by the platform’s algorithmic incentive architecture, creators’ rational choice logics, and users’ cognitive preferences. Consequently, governance cannot rely on a single actor or on simplistic technical blockage. Instead, it should be advanced in a coordinated manner across three dimensions: platform institutional design, creator incentive mechanisms, and the enhancement of users’ cognitive capacities, thereby establishing a durable multi-stakeholder co-governance framework.

Platform-level recommendations: From attention maximisation to information quality governance

Xiaohongshu’s recommendation logic, which is centred on completion rates, interaction volume, and conversion rates, is structurally predisposed to favour content that is emotionalised, simplified, and dramatised, thereby providing institutional soil in which cognitive-degrading content can consistently outperform alternatives. The key to platform governance thus lies not in after-the-fact account bans as an end-of-pipeline disposal measure, but in an auditable and operational reconstruction of the algorithmic incentive structure. First, the platform should introduce weights for information credibility and professional compliance into recommendation models and traffic distribution rules and implement graded management and mandatory disclosure regimes for high-risk domains such as health, parenting, finance, and psychology. For example, posts could be required to select a domain label at the point of publication and simultaneously provide verifiable sources, such as registration numbers filed with the National Medical Products Administration, links to authoritative guidelines, journal article DOIs, or the provenance of official statistics. The system could then automatically generate a source completeness score, which would serve as a prerequisite for entry into recommendation pools, search ranking privileges, and commercial conversion permissions. In addition, credential verification should be shifted from the account level to the content level by establishing an evidence card for each individual post. For claims involving efficacy, risk judgments, or causal inference, structured fields should be required, including evidence type, scope of applicability, sample origin, and whether the statement is based on personal experience. On the algorithmic side, the combination of missing evidence and high conversion signals should be treated as a risk indicator that triggers human spot checks and distribution throttling.

Second, the platform should develop a delayed recommendation and risk buffering mechanism. Content characterised by high emotional arousal and high controversy, but insufficient evidence should first enter an observation pool and be distributed on a trial basis to a limited user cohort within the first X hours. Decisions on whether to scale diffusion should then be made in conjunction with indicators such as reporting density, credibility prompts derived from comment interactions, for example votes on whether sources have been provided, and the outcomes of fact-checking. For staged performances, exaggerations, and pseudo-evaluations that clearly rely on explosive early-stage traffic, diffusion capacity can be weakened by restricting forwarding chains, reducing cross-community recommendations, and limiting exposure through trending entrances. In this way, the content diffusion dynamics can be transformed from explosive diffusion to controllable diffusion.

Beyond these constraint-oriented modifications, an incentive-based repair package is essential. Otherwise, rational content will remain structurally disadvantaged in attention competition. The platform can introduce a set of operational positive mechanisms. First, it can establish a rational content support package by granting a professional creation label to creators who provide authoritative sources, complete argumentation, and boundary conditions, and by offering stable exposure compensation within recommendation streams. In particular, long-form explanatory and methodological content should receive quality weighting that is not tied to completion rates, so that it is not systematically eliminated due to higher comprehension costs. Second, a mechanism to internalise the costs of misinformation should be established. For content verified as involving pseudo-scientific claims, fabricated authoritative endorsement, or staged narratives designed to induce conversion, removal should be accompanied by immediate restrictions on commercial rights, such as prohibiting link placement for a fixed period, restricting advertising, or downgrading commerce privileges. Violations should also be bound to an account credit score, ensuring that the profits of deception cannot be easily converted into long-term commercial capacity. Third, user-facing prompts and co-governance tools should be strengthened. For high-risk domain posts, a default source and boundary module can be displayed beneath the content, enabling one-click evidence viewing, scoring of source adequacy, and rapid reporting via granular categories such as pseudo-authority, pseudo-evaluation, or staged performance. This design allows users' rational queries to be recognised by algorithms as positive interaction signals, rather than allowing emotional comments alone to serve as the primary popularity indicator. Through an integrated institutional bundle comprising upstream admission, diffusion buffering, positive incentives, and violation costs, the platform can shift from an attention maximisation distribution logic toward an ecosystem repair pathway oriented to information quality governance.

Creator-level recommendations: Breaking the institutional lock-in of traffic template dependence

From the perspective of content generation mechanisms, the continuous replication of cognitive-degrading content and the formation of stable production chains essentially reflect creators' rational responses to platform incentive signals. Under current conditions of intense traffic competition, once creators discover that content forms such as anxiety manufacture, staged misery narratives, terminology stacking, and extreme positional expression are more likely to obtain recommendation exposure and commercial conversion, imitation and template adoption become the lowest-risk and most stable-return strategy. Within this structure, creators are not simply lacking in ethics. Rather, they are embedded within a payoff function centred on click-through and conversion rates, gradually developing path dependence and creative inertia. Governance that remains at the level of moral condemnation or isolated account bans is therefore unlikely to reach the root of the problem, and may even intensify a guerrilla-style strategy of switching accounts after penalties. To meaningfully change creators' behavioural logics, the cost-benefit calculation must be institutionally reconstructed, such that the expected returns of cognitive-degrading production decline, the risks rise substantially, and the long-term rewards of rational creation become increasingly salient, thereby breaking the institutional lock-in of traffic template dependence.

In concrete terms, the first step is to establish an enforceable content accountability and traceability mechanism. For accounts verified as engaging in severe misinformation, fabricated authority, staged fraud, or commercial deception, platform responses should extend beyond bans to include freezing historical commercial proceeds, reclaiming commerce revenue shares, and restricting the ability to re-register monetised accounts within a defined period. Serious violation records can further be integrated into the platform credit system for cross-account identification in order to prevent re-emergence under new identities. This combined mechanism of revenue traceability and credit-based punishment can fundamentally raise violation costs and shift the low-risk game in which creators treat account bans as inconsequential. Second, professional boundary institutionalisation should be advanced. Clear legal and platform responsibility boundaries should be established between personal experience sharing and professional advice. For content involving high-risk domains such as medicine, psychology, nutrition, and investment, credential verification, risk disclaimers, and source labelling should be mandatory. Otherwise, commercial conversion functions should be restricted. The survey demonstrates that when platforms visibly label information sources, users' trust in fabricated authority declines significantly, indicating that institutional design alone can reshape trust structures. At the same time, the platform should systematically introduce a slow content support programme through creating subsidies, traffic guarantees, and themed recommendation slots that specifically support in-depth science communication, long-form explanatory content, and methodological content, ensuring that rational expression is no longer structurally disadvantaged in algorithmic competition [34]. Only when explaining clearly is no longer institutionally penalised relative to explaining quickly will creators have genuine incentives to disengage from cognitive-degrading templates and return to the public value of content itself.

Conclusion

This study has systematically explored the formation mechanisms and preventive strategies concerning cognitive-degrading content on the Xiaohongshu platform. Through structured analysis, five dominant types of such content were identified: pseudo-scientific marketing and professional discourse stacking, fallacious testing and cognitive shortcut inducement, fabricated authority and selective quotation, extreme emotional manipulation coupled with identity binding and conspiratorial attribution, and immersive fictional reality. Each type employs distinct yet systematic techniques to mimic rational discourse while undermining users' critical thinking and evidence evaluation capabilities.

The persistence and amplification of cognitive-degrading content are explained through a dual-level analytical framework. At the micro level, dual-process theory reveals how such content strategically activates intuitive, heuristic-based System 1 thinking while suppressing effortful, analytical System 2 processing. At the macro level, perspectives from the attention economy and the political economy of algorithms illustrate how platform recommendation logics - prioritizing engagement metrics such as completion rates and interaction volume - structurally favor emotionally charged, low-cognitive-cost content, thereby creating a self-reinforcing diffusion cycle.

Importantly, this study moves beyond individual-level moral explanations and reconceptualizes cognitive-degrading content as a structural ecological problem co-shaped by platform architectures, creator incentives, and user cognitive tendencies. Consequently, effective governance requires coordinated multi-actor interventions. Platform-level measures should include algorithmic restructuring with quality-weighted recommendation models, delayed diffusion mechanisms for high-risk content, and enhanced source verification requirements. Creator-level strategies involve enforceable accountability mechanisms, professional boundary institutionalization, and incentives for rational content production. User-level approaches focus on improving media literacy and providing participatory co-governance tools.

In summary, this research contributes to the understanding of manipulative information ecosystems in lifestyle-oriented social media and offers a systematic, diagnostically grounded pathway for platform governance. Future studies may further examine the longitudinal effects of algorithmic interventions, cross-platform comparisons of content degradation patterns, and the efficacy of digital literacy programs in mitigating susceptibility to cognitive-degrading content.

Funding

This work was not supported by any funds.

Acknowledgements

The authors would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Amazeen, M. A. (2022) The psychological drivers of misinformation belief and spread. *Nature Reviews Psychology*, 1, 13-29.

[2] Tomassi, A., Falegnami, A., Romano, E. (2024) Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society. *PLOS One*, 19(5), e0303183.

[3] Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. H. (2024) Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47), e2409329121.

[4] Sun, Y., Xie, J. (2024) Do Heuristic cues affect misinformation sharing? Evidence from a meta-analysis. *Journalism & Mass Communication Quarterly*, 10776990241284597.

[5] Shin, D., Shin, E. Y. (2025) Cascading falsehoods: mapping the diffusion of misinformation in algorithmic environments. *AI & SOCIETY*, 1-18.

[6] Zenone, M., Kenworthy, N., Maani, N. (2023) The social media industry as a commercial determinant of health. *International Journal of Health Policy and Management*, 12, 6840.

[7] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., Rand, D. G. (2021) Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, 590-595.

[8] Denniss, E., Freeman, M., colleagues. (2025) Social media and the spread of misinformation. *Health Promotion International*, 40(2), daaf023.

[9] Melchior, C., Oliveira, M. (2022) Health-related fake news on social media platforms: a systematic literature review. *New Media & Society*, 24(6), 1500-1522.

[10] Fick, J., Rudolph, L., Hendriks, F. (2025) Jargon avoidance in the public communication of science: Single- or double-edged sword for information evaluation? *Learning and Instruction*, 98, 102121.

[11] French, A. M., Storey, V. C., Wallace, L. (2025) The impact of cognitive biases on the believability of fake news. European Journal of Information Systems, 34(1), 72-93.

[12] Elnaggar, O., Arelhi, R., Coenen, F., Hopkinson, A., Mason, L., Paoletti, P. (2023) An interpretable framework for sleep posture change detection and postural inactivity segmentation using wrist kinematics. *Scientific Reports*, 13(1), 18027.

[13] Dan, V., Arendt, F. (2021) Visual misinformation: the persuasive power of staged reality. *Communication Research*, 48(7), 1002-1024.

[14] Cho, Y. Y., Woo, H. (2025) Heuristic and Systematic processing on social media: pathways from literacy to fact-checking behavior. *Journalism and Media*, 6(4), 198.

[15] Wang, R., Yang, H., Wang, Y., Zhai, X. (2025) Understanding how users identify health misinformation in short videos: an integrated analysis using PLS-SEM and fsQCA. *Frontiers in Public Health*, 13, 1713794.

[16] Geels, J., Graßl, P., Schraffenberger, H., Tanis, M., Kleemans, M. (2024) Virtual lab coats: the effects

of verified source information on social media post credibility. *PLOS One*, 19(5), e0302323.

[17] Inwood, O., Zappavigna, M. (2024) The legitimation of screenshots as visual evidence in social media: YouTube videos spreading misinformation and disinformation. *Visual Communication*, 14703572241255664.

[18] Shahbazi, M., Bunker, D. (2024) Social media trust: Fighting misinformation in the time of crisis. *International Journal of Information Management*, 77, 102780.

[19] Van Knippenberg, D., Van Kleef, G. A. (2016) Leadership and affect: Moving the hearts and minds of followers. *Academy of Management Annals*, 10(1), 799-840.

[20] Guess, A., Nagler, J., Tucker, J. (2020) Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586.

[21] Swire-Thompson, B., Lazer, D. (2022) Public health and online misinformation: challenges and recommendations. *Annual Review of Public Health*, 43, 49-69.

[22] Bakshy, E., Messing, S., Adamic, L. A. (2021) Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132.

[23] Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., Starnini, M. (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America*, 118(9), e2023301118.

[24] Mancosu, M., Marchi, L., Pellegrini, G. (2020) The emotional underpinnings of fake news acceptance and sharing: Evidence from Italy. *Journal of Trust Research*, 10(2), 97-123.

[25] Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Quattrociocchi, W. (2021) Users polarization on Facebook and YouTube. *PLOS One*, 11(8), e0159641.

[26] Shu, K., Wang, S., Liu, H. (2022) Disinformation, misinformation, and fake news in social media. *IEEE Data Engineering Bulletin*, 45(1), 3-15.

[27] Mialon, M., Swinburn, B., Sacks, G. (2021) A proposed approach to systematically identify and monitor the commercial determinants of health. *Globalization and Health*, 17, 16.

[28] Slater, M. D., Long, M., Ford, V. (2023) Narrative persuasion, emotion, and entertainment: a meta-analysis. *Communication Research*, 50(1), 3-27.

[29] Da Silva, S. (2023) Dual-process theory: a review of System 1 and System 2 thinking. *Psych*, 5(4), 611-628.

[30] Ku, Y. (2025) Dual-process models of social media information processing: heuristic and systematic pathways. *Computers in Human Behavior*, 153, 108190.

[31] Subramanian, H., Mitra, S., Ransbotham, S. (2021) Capturing value in platform business models that rely on user-generated content. *Organization Science*, 32(3), 804-823.

[32] Sharma, V., Bray, K. E., Kumar, N., Grinter, R. E. (2022) Romancing the algorithm: Navigating constantly, frequently, and silently changing algorithms for digital work. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1-29.

[33] Fehrer, J. A., Woratschek, H., Brodie, R. J. (2018) A systemic logic for platform business models. *Journal of Service Management*, 29(4), 546-568.

[34] Wu, S., Cheng, H., Qin, Q. (2024) Physical delivery network optimization based on ant colony optimization neural network algorithm. *International Journal of Information Systems and Supply Chain Management (IJISSCM)*, 17(1), 1-18.