World Journal of Educational Studies
Print ISSN: 2959-9989
Online ISSN: 2959-9997
DOI: https://doi.org/10.61784/wjes3146

THE ALIENATION OF UNIVERSITY STUDENTS' ONLINE MENTALITY AND COGNITIVE RISK INTERVENTION UNDER ALGORITHMIC DRIVE

YiXin Li
School of Journalism and Communication, Chengdu Sport University, Chengdu 610041, Sichuan, China.

Abstract: This research investigates the evolutionary mechanism and risk intervention of university students' online mentality alienation under the deep embedding of intelligent algorithms. Based on the Knowledge-Attitude-Practice (KAP) theoretical framework, a chain mediation model was constructed, and 351 valid empirical data points were obtained through targeted sampling on leading algorithm-driven social media platforms popular among youth. Using Latent Profile Analysis (LPA) and Structural Equation Modeling (SEM), the study reveals three heterogeneous groups: "Rational Self-Control" (33.6%), "Immersive Follower" (48.7%), and "Highly Alienated" (17.7%). Findings show that irrational trust in algorithmic recommendations directly triggers cognitive conflict (K) while bypassing the "information cocoon" (which negatively affects defense), sequentially triggering negative emotions (A), uncontrolled internet use, and defensive cognition (P). Notably, digital literacy failed to buffer front-end algorithmic discipline but significantly moderated the "cognitive conflict → uncontrolled internet use" path. This study deepens the micro-psychological understanding of technological alienation and provides empirical evidence for transitioning toward "algorithmic logic de-blindness" in educational intervention strategies.
Keywords: Algorithmic recommendation; Online mentality alienation; Information cocoon; KAP theory; Latent profile analysis

1 INTRODUCTION

1.1 Research Background and Problem Statement

With the ubiquitous development of artificial intelligence and big data technologies, algorithmic recommendation has been deeply embedded in the information distribution field of contemporary youth as an underlying logic. Currently, the information acquisition mode of university students has undergone a fundamental reversal, shifting comprehensively from the traditional "human seeking information" to "information seeking human". This precise distribution based on user profiles and data preferences, while greatly improving information matching efficiency, increasingly exposes the profound hidden dangers of "technological alienation". As digital natives, university students exhibit high-intensity internet usage stickiness and strong receptivity to novelties; however, their worldviews and values are still in the formative stage, making them highly susceptible to being disciplined and dominated by algorithmic logic.

Within the homogeneous information cocoons woven by algorithms, university students' blind trust in platform recommendation mechanisms is quietly inducing a series of online mentality alienation crises. This study keenly captures that this mentality alienation is not a single-dimensional sudden mutation but follows a hidden progressive chain. Based on the "Knowledge-Attitude-Practice" (KAP) theoretical framework, this study deeply analyzes the specific evolutionary path of university students' psychological alienation driven by algorithms: First, in a highly homogeneous push environment, occasional exposure to heterogeneous or opposing information triggers individual "cognitive conflict" (Knowledge/Cognition layer). If this cognitive dissonance cannot be effectively adjusted, it will quickly spread to the affective domain, generating "negative emotions" such as anxiety and irritability (Attitude/Affective layer). Ultimately, the rational defense line is breached, and the imbalance of mentality externalizes into behavioral misconduct, leading to "uncontrolled internet use" and strong defensive cognition (Practice/Behavior layer).

However, existing research mostly focuses on macro-governance of algorithms or single descriptions of mentality, with few studies deeply dissecting the internal mechanisms of this alienation chain from the perspective of the KAP continuum using empirical data. Based on this, the present study proposes the core research questions: How does algorithmic recommendation trust, catalyzed by information cocoons, induce the chain alienation of university students' online mentality under the KAP logic? Furthermore, what role does individual digital information literacy play in this risk generation and transmission path?.

1.2 Research Significance and Objectives

1.2.1 Research significance

In the theoretical dimension, this study breaks through the previous flattened perspective on algorithmic impact, creatively integrating "algorithmic trust and the cocoon effect" from the perspective of communication studies with the

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

KAP theory. Through rigorous empirical data measurement, it clearly delineates the dynamic model of mentality alienation featuring "cognitive conflict—negative emotion—uncontrolled behavior," thereby deepening the theoretical connotation of "technological alienation" at the micro-subject psychological level.

In the practical intervention dimension, facing the cognitive risks brought by algorithms, traditional "restrictive" management has proven ineffective. Based on the contemporary demands for digital well-being and psychological resilience in higher education, this study attempts to break the deadlock from the perspective of "reshaping subjectivity". The research findings not only help higher education practitioners accurately identify hidden psychological crises induced by algorithmic recommendations but also provide robust empirical evidence for constructing systematic intervention schemes.

1.2.2 Research objectives

This study aims to accurately depict the latent categories and group characteristics of university students' online mentality alienation through empirical investigation and multidimensional data analysis (such as Latent Profile Analysis and Structural Equation Modeling). It seeks to deeply reveal the chain mediation mechanisms of algorithmic recommendation trust and information cocoon perception on cognitive conflict, negative emotion, and uncontrolled internet use (K-A-P). Ultimately, grounded in educational practice, the study aims to propose targeted, innovative intervention paradigms and cognitive risk management systems to prevent online mentality alienation and mitigate defensive cognition.

2 LITERATURE REVIEW AND THEORETICAL BASIS

2.1 Literature Review

This research is grounded in the realistic context of intelligent algorithms deeply embedding into the information acquisition field of university students. Following the logical thread of "algorithmic recommendation - online mentality alienation - intervention mechanism," and combining domestic and international authoritative scale development and empirical measurement literature, this study systematically reviews the seven core dimensions constituting this research. It aims to explore the deep mechanisms of individual cognitive dissonance, emotional alienation, and behavioral loss of control under technical discipline.

2.1.1 Antecedent variables: The transfer of algorithmic trust and the perception of information cocoons

In the intelligent era, algorithmic distribution has reshaped the information environment for university students, but the hidden nature of technology has triggered deep-seated trust and cognitive crises. Huang and Liu, when exploring the relationship between algorithmic awareness and user behavior, provided a core target for researching the "algorithmic recommendation trust" dimension, explicitly pointing out: "Concerns about personalized bias and opaque algorithmic control trigger questions about trust and user agency" [1]. This blind trust or questioning of algorithms directly determines the extent to which individuals cede their right to select information. Driven by trust, users inevitably fall into the homogeneous networks woven by algorithms. Regarding the "information cocoon perception" dimension, this study profoundly draws on the research logic of Dogruel et al. In Development and Validation of an Algorithm Literacy Scale for Internet Users, they point out that evaluating Internet users' understanding of algorithmic mechanisms is crucial [2]. This provides a logical starting point for quantifying "information cocoon perception": an information cocoon is no longer merely an objective physical isolation state but depends on the extent to which users can subjectively "perceive" the algorithm's filtering mechanism and content homogeneity. Users' perceptual acuity toward algorithmic filter bubbles constitutes the prerequisite for subsequent mentality fluctuations.

2.1.2 The starting point and mediator of KAP alienation: Cognitive conflict and negative emotions

When university students are confined within information cocoons for extended periods, encountering heterogeneous information triggers cognitive fluctuations. Metzger et al. conducted a profound analysis of selective exposure, showing that: "News consumers tend to seek information consistent with their attitudes and avoid information that challenges their attitudes". Furthermore, they confirmed: "When exposed to attitude-challenging information sources, people experience more cognitive dissonance" [3]. This cognitive dissonance constitutes the psychological origin of the "cognitive conflict" dimension in this study. When conflicts in the cognitive dimension cannot be resolved, the psychological defense line inevitably sinks to the affective dimension. Qiu et al. pointed out that subjective well-being is an individual's overall evaluation of their life conditions based on self-defined standards, encompassing two basic components: cognition and affect [4]. Under the discipline of the algorithmic environment, the emotional experience of university students' daily media exposure is gradually stripped of positive components, transforming into tension, irritability, and distress after use. This provides a highly reliable and valid measurement basis for the "negative emotions" dimension in this study, which acts as an emotional mediator accelerating the process of mentality alienation.

2.1.3 Behavioral outcomes and moderating variables: Uncontrolled internet use, defensive cognition, and information literacy

The alienation of mentality and emotions eventually externalizes into behavioral disorder. On the one hand, there is inward immersion. Liu, when investigating social media dependency among university students, provided a theoretical interpretation for the "uncontrolled internet use" dimension, pointing out that this behavior is not just a functional choice, but "is both the result of the audience actively seeking positive media gratification and contains significant passivity and non-purposiveness" [5]. This non-purposive indulgence directly leads to university students' time evaporation and detachment from reality in the face of algorithmic recommendations. On the other hand, there is

outward aggression. The longitudinal study by Quan and Xia provided a rigorous psychological explanation for the "defensive cognition" dimension [6]: "Hostile attribution bias refers to the tendency to attribute hostile intent to the ambiguous behaviors of others, which is considered the main cognitive factor leading to aggression". In a cocoon environment, individuals facing dissenting views are highly susceptible to stimulating this hostile attribution bias, thereby engaging in intense verbal counterattacks or physical disconnection to maintain their remaining cognitive closure. Faced with these risks, the latest ILMS-34 scale developed by Li et al. guides the direction for the "digital information literacy" dimension and intervention schemes of this project: "A comprehensive and reliable assessment tool is needed to evaluate information literacy and guide future literacy enhancement programs" [7]. This establishes the key moderating role of information literacy in interrupting the alienation chain.

2.2 Definition of Core Concepts

Based on the theoretical support of the above seven core literatures, and strictly corresponding to the measurement tools in the questionnaire design of this project, this study defines the seven core concepts as follows:

Algorithmic recommendation trust: Refers to university students' perceptual reliance on the reliability and intentions of platform recommendation systems, manifesting as the unguarded acceptance of algorithmic "information customization" and the belief that there is no deliberate misleading [1].

Information cocoon perception: Refers to the psychological state wherein university students, based on their awareness of algorithmic filtering mechanisms in daily media exposure, subjectively perceive that the received information exhibits high homogeneity (the "echo chamber effect") and that information sources are increasingly narrowing [2].

Cognitive conflict: Refers to the state of "cognitive dissonance"—characterized by internal tearing, annoyance, and discomfort—generated when individuals accidentally encounter "attitude-challenging information sources" within the cocoon [3].

Negative emotions: Refers to the tense, uneasy, and irritable negative affective experiences generated by university students based on subjective evaluations after experiencing cognitive dissonance or the excessive use of intelligent media [4].

Uncontrolled internet use: Refers to individuals losing the ability of rational moderation over media use under high-frequency algorithmic stimulation, falling into a "passive and non-purposive" dependency, and using it as the primary means of escaping reality [5].

Defensive cognition: Refers to individuals developing a "hostile attribution bias" in online interactions, tending to interpret ambiguous criticisms as personal hostility, and employing reactive aggressive behaviors such as posting intense comments to maintain cognitive defense [6].

Digital information literacy: Refers to the comprehensive capability system of individuals to identify algorithmic logic, evaluate information authenticity, and engage in critical thinking and self-regulation in complex digital environments [7].

2.3 Theoretical Basis and Model Construction: The Operational Mapping of KAP Theory in Online Mentality Alienation

The top-level theoretical framework of this study is the "Knowledge-Attitude-Practice" (KAP) model. This model reveals the coherent process of an individual from knowledge acquisition (cognition) to belief establishment (affection) and then to behavioral change. This project deeply embeds this classic model into the intelligent communication perspective to deconstruct the complex causal chain of university students' mentality alienation driven by algorithms. Driven by external antecedents structured by "algorithmic recommendation trust" and "information cocoon perception," the internally triggered mentality alienation strictly follows the three evolutionary dimensions of KAP.

2.3.1 Cognitive dimension (Knowledge/Cognition) — Corresponding to "cognitive conflict" (K)

In the primary stage of the KAP model, the cognitive state is the foundation of psychological activity. Under algorithmic discipline, university students' original cognitive schemas are tightly wrapped by homogeneous information. Once the balance of selective exposure is broken and individuals encounter challenging information, phenomenal cognitive dissonance erupts. This study operationalizes this process as "cognitive conflict" (K), which represents the discomfort and fluctuation generated by the old knowledge system in the algorithmic era, constituting the cognitive starting point of the mentality alienation chain.

2.3.2 Attitude/Affective dimension (Attitude/Affective) — Corresponding to "negative emotions" (A)

When cognitive-level conflicts (K) cannot be adjusted through rational literacy, the psychological defense line inevitably sinks to the affective dimension. Cognitive confusion and internal tearing quickly breed negative experiences such as irritability and anxiety. In the "attitude/affective" stage of KAP, individuals accumulate a significant amount of unease. This reinforced "negative emotion" (A) becomes the key emotional hub that connects internal cognitive dissonance with external behavioral deviations.

2.3.3 Behavioral dimension (Practice/Behavior) — Corresponding to "uncontrolled internet use" and "defensive cognition" (P)

The ultimate foothold of the KAP model is behavioral practice (P). To escape the pain of "cognitive conflict" (K) and vent "negative emotions" (A), university students evolve two disordered behavioral patterns: first, an inward "uncontrolled internet use," where individuals numb themselves by endlessly scrolling; second, an outward "defensive

cognition," where individuals develop strong hostile attribution biases and reject heterogeneous communication with intense verbal counterattacks, completely enclosing themselves in their remaining cocoon territories. In summary, based on the exclusive literature support of the seven dimensions and the logical extension of the KAP theory, this study constructs a chain mediation intervention model of "algorithmic trust - information cocoon perception - cognitive conflict (K) - negative emotions (A) - uncontrolled/defensive behavior (P)". This lays a solid theoretical foundation for subsequently evaluating the moderating role of digital information literacy and implementing targeted interventions.

3 RESEARCH DESIGN AND METHODS

This study aims to scientifically test the previously constructed "algorithmic drive - online mentality alienation" KAP chain mediation intervention model through a quantitative empirical approach. To ensure the objectivity, universality, and ecological validity of the research conclusions, the research team conducted rigorous planning regarding sampling design, instrument development, and statistical strategies.

3.1 Sample Acquisition and Survey Procedures

3.1.1 Targeted recruitment and stratified quota sampling based on digital social platforms

Considering that the research subjects are university students deeply embedded in algorithmic environments, traditional offline interception or cluster sampling within a single university is difficult to truly reflect the natural ecology of intelligent media use. Therefore, this study distributed targeted recruitment questionnaires on leading algorithm-driven social media platforms widely utilized by Chinese university students, ensuring that the sample acquisition process itself possesses high ecological validity.

3.1.2 Sample structure control and representativeness testing

To effectively overcome the self-selection bias risk potentially caused by internet non-probability sampling, the research team implemented strict stratified control of demographic and sociological characteristics during the questionnaire distribution and collection stages, ensuring the structural representativeness of the sample pool:

Broad regional distribution: The sample comprehensively covers university students in core regions such as East China, South China, Southwest China, and Central China, ensuring macroscopic cross-regional universality and weakening the interference of a single regional culture on online mentality.

Heterogeneity of institutional tiers: In terms of quota indicators, the study deliberately bridged academic and educational circles. The respondents encompass elite groups from "Double First-Class" (Project 985/211) universities, as well as a large number of students from ordinary provincial universities and vocational colleges, fully guaranteeing the generalization ability of the research conclusions to the entire higher education field.

Balanced demographic characteristics: Among the collected valid samples (the final valid sample size entering the model testing was 351, meeting the baseline requirement of structural equation modeling for parameter estimation), the gender ratio remained basically balanced (male:female \approx 1:1), covering vocational students, undergraduates, and postgraduates. This multidimensional, cross-level sample distribution significantly enhances the external validity and statistical rigor of the subsequent empirical deductions.

3.2 Core Variables and Measurement Tools

The measurement tools of this study were developed by the research team based on a systematic review of authoritative scales at home and abroad, combined with the localized context of university students' intelligent media contact. The questionnaire adopted a 5-point Likert scale (1 = "completely disagree", 5 = "completely agree"). To achieve precise measurement of the KAP theory, the operational design of the core variables is as follows:

3.2.1 Antecedent variables measurement

Algorithmic recommendation trust: Based on the algorithm awareness framework of Huang et al., 7 items were designed. It mainly measures respondents' subjective confidence in the reliability and honesty of platform recommendation systems (Example item: "I think the recommendation function of short-video platforms is reliable") [1].

Information cocoon perception: Referring to the algorithm literacy evaluation logic of Dogruel et al., 6 items were designed. It focuses on measuring individuals' awareness of content homogeneity and the narrowing of information channels (Example item: "I feel like I am trapped in a fixed information circle") [2].

3.2.2 Structural measurement of the KAP alienation chain

Cognitive dimension (K) — Cognitive conflict: Drawing on the cognitive dissonance scale by Metzger et al., 5 items were designed to capture the cognitive fluctuations when individuals encounter heterogeneous information (Example item: "I feel distressed when I see content that contradicts my inherent cognition") [3].

Attitude/Affective dimension (A) — Negative emotions: Adopting the negative affect subscale from the PANAS scale revised by Qiu et al., 5 items were designed to evaluate the mood state after media exposure (Example item: "I often feel irritable after using short-video apps") [4].

Behavioral dimension (P) — Uncontrolled internet use and defensive cognition:

Uncontrolled internet use (5 items): Based on Liu's dependency scale, measuring time management failure and reality-escape tendencies (Example item: "Scrolling through short videos has become my main way to escape problems in reality") [5].

Defensive cognition (6 items): Combining the hostile attribution bias tool by Quan et al., investigating reactive aggressive and social disconnection behaviors when coping with cognitive threats (Example item: "When I feel offended by short-video content, I will immediately post intense comments") [6].

3.2.3 Moderating variable measurement

Digital information literacy: Synthesizing the latest scale indicators by Li et al., encompassing multiple dimensions such as algorithmic logic identification and authenticity verification, used to test its buffering and immunizing role in risk transmission [7].

3.3 Data Processing and Statistical Analysis Strategies

To comprehensively deconstruct the complex mechanisms of mentality alienation, this study comprehensively utilized statistical software such as SPSS 26.0, the PROCESS macro plugin, and Jamovi, establishing a three-stage analysis strategy of "testing - classification - pathfinding".

3.3.1 Reliability and validity testing and measurement model evaluation

First, SPSS was used to calculate the Cronbach's \alpha coefficient and Composite Reliability (CR) of each latent variable. Subsequently, Confirmatory Factor Analysis (CFA) was conducted using JASP software. Considering the non-normal distribution characteristics of some data, this study adopted the robust Diagonally Weighted Least Squares (DWLS) estimation method. By evaluating model fit indices (CFI, TLI, RMSEA, etc.), the convergent validity and discriminant validity of each dimension were established, providing a solid data foundation for subsequent calculations.

3.3.2 Latent Profile Analysis (LPA)

A person-centered LPA analysis technique was introduced (executed via Jamovi). Based on the continuous score indicators of university students in the 7 core dimensions, potential heterogeneous subgroups existing within the sample were explored. By comparing the information criteria (AIC, BIC, aBIC) and Entropy values of different classification retention schemes, the typical group profiles such as the "high-risk/alienated type" and "low-risk/rational type" were accurately stripped out, providing precise targets for subsequent stratified interventions.

3.3.3 Structural equation and chain mediation effect testing

Under the premise of ensuring measurement validity, the PROCESS macro program developed by Hayes was used for path analysis. Specifically, the Bootstrapping method was used to test the direct effect from "algorithmic recommendation trust" to "defensive cognition/uncontrolled use," and to deeply unravel and verify the chain mediation mechanism played by "cognitive conflict (K) — negative emotions (A)" in it. Finally, digital information literacy was incorporated into the model to conduct a moderated mediation effect test.

4 DATA ANALYSIS AND EMPIRICAL RESULTS

This study systematically processed 351 valid sample data using statistical analysis tools such as SPSS 26.0, the PROCESS v4.1 macro program, and Jamovi. The data analysis logic followed the progressive steps of "measurement model evaluation - latent class division - causal pathfinding - moderation effect testing" to comprehensively reveal the occurrence mechanisms of university students' online mentality alienation driven by algorithms.

4.1 Descriptive Statistics and Sample Characteristics

A total of 351 valid questionnaires were collected and finalized for this study. Before entering the core latent profile analysis and hypothesis testing, a descriptive statistical analysis was conducted on the demographic characteristics, educational levels, and academic backgrounds of the surveyed population. The specific distribution pattern is detailed in Table 1.

Table 1 Demographic Distribution Characteristics of the Sample

Demographic Variable	Category	Frequency(n)	Percentage(%)
Gender	Male	170	48.4
	Female	181	51.6
Academic Institution	Project 985 institutions	22	6.3
	Project 211/Double First-Class universities	65	18.5
	Provincial key universities (Tier 1)	118	33.6
	Polytechnic colleges (Tier 2 institutions)	146	41.6
	Vocational college students	5	1.4
Academic Stage	Undergraduates	304	86.6
	Master's candidates	40	11.4
	Doctoral candidates	2	0.6
Disciplinary Field	Humanities & Social Sciences (Literature, History, Philosophy, Education, Journalism,	81	23.1

Demographic Variable	Category	Frequency(n)	Percentage(%)
	Linguistics)		
	STEM Fields (Natural Sciences, Engineering, Information Technology, Computer Science)	90	25.6
	Economics & Management (Business, Finance, Tourism, Law, Psychology)	141	40.2
	Other Fields (Arts, Sports, Agriculture, Medical Sciences)	39	11.1
Geographical Region	Northern China (Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia)	34	9.7
	Northeastern China (Liaoning, Jilin, Heilongjiang)	8	2.3
	Eastern China (Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Shandong, Taiwan region)	109	31.1
	Central China (Henan, Hubei, Hunan)	51	14.5
	Southern China (Guangdong, Guangxi, Hainan, Hong Kong region, Macau region)	71	20.2
	Southwestern China (Chongqing, Sichuan, Guizhou, Yunnan, Xizang)	60	17.1
	Northwestern China (Shaanxi, Gansu, Qinghai, Ningxia, Xinjiang)	18	5.1

From the data distribution, the sample exhibits good ecological representativeness in its structure. In terms of educational levels and institutional distribution, undergraduates constitute the main body of this survey (accounting for 86.6%). Furthermore, the proportions of students from "Double First-Class" (Project 985/211) universities, ordinary first-tier, and second-tier institutions basically match the realistic pyramid ecology of current domestic higher education students. The sample also includes approximately 11.4% of postgraduate students. Regarding academic backgrounds, the respondents cover multiple disciplines, including economics and management, humanities and social sciences, STEM, and arts, sports, agriculture, and medicine.

Although the absolute proportions of vocational students and doctoral candidates are relatively small, and the economics and management majors show a certain degree of sampling concentration, the overall sample remains basically balanced in terms of gender and regional distribution. This rich and somewhat heterogeneous sample pool provides sufficient data support and an analytical foundation for subsequently using structural equation modeling to deeply explore the alienation mechanisms of university students' online mentality.

4.2 Common Method Bias Control and Reliability and Validity Testing

4.2.1 Common method bias testing

Considering that the data in this study were all obtained through self-reporting, common method bias might exist. During the data processing stage, Harman's single-factor test was adopted to conduct an unrotated principal component analysis on the measurement items of all core variables. The results showed that there were multiple factors with eigenvalues greater than 1, and the variance explained by the first common factor was far below the critical threshold of 50%. This indicates that there is no serious common method bias in the data, and the data source possesses high objectivity.

4.2.2 Measurement model reliability testing

Reliability analysis aims to evaluate the internal consistency and stability of the measurement tools. This study conducted Cronbach's alpha coefficient tests on the 7 core latent variables. The results indicated that the Cronbach's alpha coefficients for all dimensions were above 0.75 (with some core dimensions reaching 0.817), demonstrating that the scales adopted in this study possess excellent internal consistency reliability, and the measurement results are stable and reliable.

4.2.3 Confirmatory Factor Analysis (CFA) and validity evaluation

To further examine the construct validity of the scales (including convergent validity and discriminant validity), this study utilized the Diagonally Weighted Least Squares (DWLS) method for Confirmatory Factor Analysis. The following Table 2 presents the key fit indices and their corresponding values, which are crucial for assessing how well the seven - factor baseline measurement model fits the actual data.

Table 2 Index Values from Confirmatory Factor Analysis for Scale Construct Validity Assessment

Index	Values
Comparative Fit Index (CFI)	0.933
Tucker - Lewis Index (TLI)	0.928
Bentler-Bonett Non-normed Fit Index (NNFI)	0.928
Bentler - Bonett Normed Fit Index (NFI)	0.848
Parsimony Normed Fit Index (PNFI)	0.867
Bollen's Relative Fit Index (RFI)	0.836
Bollen's Incremental Fit Index (IFI)	0.934
Relative Non - centrality Index (RNI)	0.933

The model fit results indicated that the seven-factor baseline measurement model fitted well with the actual data, and all core fit indices met or exceeded statistically recommended standards: Comparative Fit Index (CFI) = 0.933 (>0.90), Tucker-Lewis Index (TLI) = 0.928 (>0.90), Bollen's Incremental Fit Index (IFI) = 0.934, and Root Mean Square Error of Approximation (RMSEA) = 0.065 (<0.08). This result fully proves that the 7 hypothesized variables have good theoretical structural discrimination.

In terms of convergent validity, the standardized factor loadings of all observed variables on their respective latent variables reached a significant level ($p < 0.001$). It should be specifically noted that in the scale measurement of complex social psychology and media cognition, the Average Variance Extracted (AVE) of some latent variables dipped slightly below the strict ideal threshold of 0.5. However, according to the classic testing criteria and empirical research conventions of Fornell and Larcker [2], when the Composite Reliability (CR) of a latent variable is greater than 0.7, its convergent validity is still acceptable even if the AVE is slightly below 0.5. The CR of all latent variables in this study far exceeded the 0.7 standard, reflecting the complexity and multidimensionality of "online mentality alienation" indicators among university students, while confirming that the overall convergent validity of the measurement tools remains robust.

4.3 Latent Profile Analysis (LPA) of University Students' Online Mentality Alienation

To break through the limitations of the traditional variable-centered perspective, this study introduced a person-centered perspective, employing Latent Profile Analysis (LPA) to explore the heterogeneous classification characteristics of university students in the algorithmic environment, see Figure 1 and Table 3. Taking the scores of the 7 core dimensions as explicit indicators, models with 1 to multiple latent classes were fitted in Jamovi.

Table 3 Latent Profile Analysis of Undergraduates in Algorithmic Environment

Model fit	Values
Model	1.0000
Classes	3.0000
LogLik	-1635.30266
AIC	3314.60531
AWE	3592.83998
BIC	3399.54261
CAIC	3421.54261
CLC	3272.24524
KIC	3339.60531
SABIC	3329.75042
ICL	-3465.48576
Entropy	0.81996
prob_min	0.88917
prob_max	0.95219
n_min	0.17664
n_max	0.48718
BLRT_val	72.16588
BLRT_p	0.00990

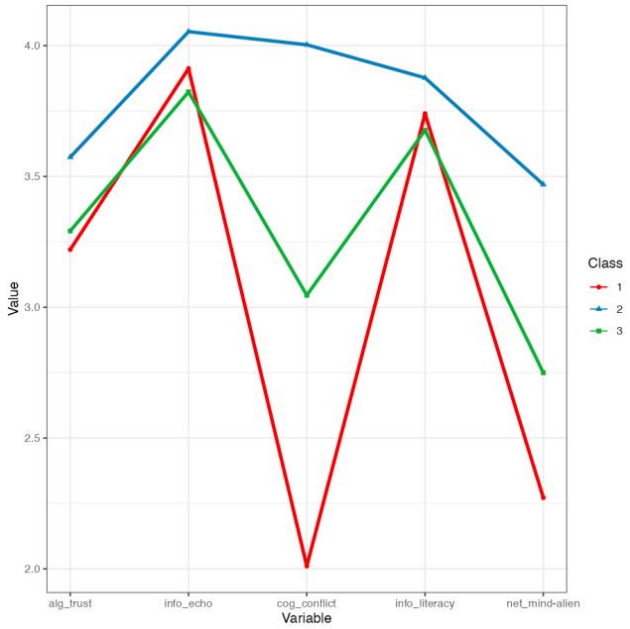

Figure 1 Latent Profile Line Plot of Undergraduates' Heterogeneous Characteristics

Based on the principle of continuously decreasing model fit information criteria (AIC, BIC, SABIC) and maximized Entropy, combined with the theoretical interpretability of the classification results, this study ultimately determined the "3-class latent profile model" as the optimal fit model. According to the scoring trends presented in the LPA line plot, the surveyed university student population was divided into the following three typical mentality risk categories:

Class 1: Low-risk / Rational self-control type (accounting for approx. 33.6%): This group scored highest in the "digital information literacy" dimension, while their scores in "algorithmic recommendation trust," "information cocoon perception," and KAP alienation variables were at the lowest levels among the three groups. They exhibit a clear cognition of algorithmic recommendation mechanisms, can rationally examine pushed content, and do not produce strong cognitive dissonance and emotional rebound when encountering heterogeneous information. These students are the immunizers in the intelligent media environment.

Class 2: Medium-risk / Immersive follower type (accounting for approx. 48.7%): This group constitutes the main body of university students. They are characterized by moderate-to-high trust in algorithmic recommendations and high "information cocoon perception". They score moderately in digital information literacy, but an upward trend begins to appear in the "uncontrolled internet use" dimension. This indicates that although these students faintly perceive themselves being in an information cocoon, constrained by the instant sensory gratification provided by algorithms, they choose compromise and immersion, positioning themselves at the critical and fermenting stage of mentality alienation.

Class 3: High-risk / Highly alienated type (accounting for approx. 17.7%): This group is the key target for online psychological intervention in universities. They score extremely high in "algorithmic recommendation trust," completely ceding their subjectivity to technology. Consequently, they present a total collapse on the KAP alienation chain: "cognitive conflict" scores surge, "negative emotions" remain persistently high, and scores for "uncontrolled internet use" and "defensive cognition" on the behavioral (P) dimension reach their peaks. They are highly prone to hostile attribution and extreme defense when facing information stimuli, making them the most direct victims of algorithmic risks.

4.4 Hypothesis Testing and Chain Mediation Effect Analysis

After establishing the validity of the measurement tools and the categorical characteristics of the sample, this study utilized the PROCESS macro program developed by Hayes to scientifically test the chain mediation hypothesis under the KAP theory, see Table 4 [8,9]. The Bootstrapping method was employed to extract 5000 resamples to calculate the 95% Confidence Interval (CI); an interval not containing 0 indicates a significant effect.

Table 4 Results of Moderated Chain - Mediation Effect Test in Hayes PROCESS

Variable/Indicator	Variable/Indicator	Effect Value	BootSE	BootLLCI
Direct Effect	X → Y	0.1254	0.0553	0.0166
Conditional Indirect Effect	info_l = -0.3858	0.0174	0.0132	-0.0010
Conditional Indirect Effect	info_l = 0.0142	0.0207	0.0139	-0.0012
Conditional Indirect Effect	info_l = 0.6142	0.0256	0.0167	-0.0017

Variable/Indicator	Variable/Indicator	Effect Value	BootSE	BootLLCI
Moderated Mediation Index	Moderation Index (Index)	0.0082	0.0093	-0.0155

Based on the empirical results, we conducted a meticulous layer-by-layer regression analysis on the theoretical model of "algorithmic recommendation trust → information cocoon perception → cognitive conflict (K) → negative emotion (A) → uncontrolled internet use / defensive cognition (P)"

4.4.1 Direct and basic effects of antecedent variables

Path analysis shows that algorithmic recommendation trust has an extremely significant positive predictive effect on information cocoon perception. Furthermore, algorithmic recommendation trust has a significant direct positive effect on defensive cognition (Effect = 0.1730, se = 0.0599, t = 2.8899, p = 0.0041, 95% CI = [0.0552, 0.2907]). In the path focusing on uncontrolled internet use, algorithmic trust similarly exhibits a strong positive driving effect. This confirms that algorithmic trust itself is the core culprit in weakening individual psychological defenses and inducing defensive mechanisms.

4.4.2 Verification of the KAP chain mediation effect

After introducing the mediating variables, this study confirmed that the chain alienation path based on the KAP theory is completely established:

Triggering of the cognitive layer (K): Algorithmic recommendation trust directly triggers university students' "cognitive conflict" (β=0.352, $p<0.001$), bypassing "information cocoon" as a prior mediator. Conversely, the cocoon exerts a negative effect on defensive cognition, showing it acts as a comfort zone until boundary-crossing exposure causes profound cognitive tearing."

Transmission of the affective layer (A): Cognitive conflict, acting as the initial catalyst in the sequential mediation chain, significantly positively predicts "negative emotions". Bootstrapping tests indicate that when cognitive dissonance cannot be digested internally, it inevitably translates into tension, unease, and pain after media use.

Eruption of the behavioral layer (P): The accumulated negative emotions eventually externalize into behavioral disorder. The mediation effect test shows that "negative emotions" significantly positively predict "uncontrolled internet use" (inward immersive escape) and "defensive cognition" (outward hostile aggression).

The comprehensive indirect effect confidence intervals (neither Boot LLCI nor Boot ULCI crossed 0) conclude that algorithmic recommendation not only directly affects university students' mentality but also ultimately leads to the comprehensive alienation of online mentality and behavior (P) through the hidden psychological transmission chain of "algorithmic trust → cognitive dissonance (K) → emotional depletion (A)".

4.5 Testing and Reflection on the Boundary Moderating Effect of Digital Information Literacy

After verifying the main effects, this study incorporated "digital information literacy" to test its buffering effect. Results showed its front-end buffering effect against algorithmic discipline was not significant ($p>0.05$). However, digital literacy significantly moderated the mid-to-late 'cognitive conflict → uncontrolled internet use' path ($p<0.05$). This indicates that while traditional literacy cannot prevent algorithm-level cognitive tearing, it applies brakes against uncontrolled addiction. This "front-end failure, back-end buffering" finding provides nuanced explanations for youth's knowledge-action disconnection."

In response to this "front-end moderating failure", the research team conducted a deep reflection and theoretical defense from three dimensions:

Localization and temporal mismatch of measurement tools: The digital information literacy scale adopted in this study mostly originates from early measurements of information retrieval and authenticity identification. In the current strong AI context, traditional "information literacy" is insufficient to counter the underlying discipline of algorithms. Even if university students possess the ability to identify fake news, they may not possess the "Algorithm Literacy" to deconstruct the "algorithmic black box". This slight mismatch in measurement tools may be the data-driven reason for the moderation failure.

Variance dilution caused by highly homogeneous samples: The sampling targets of this study all have higher education backgrounds, and their overall digital information literacy scores are generally high and concentrated. This high degree of intra-group homogeneity objectively diluted the statistical variance of the moderating variable, making it difficult to capture significant moderation fluctuations in the regression equation.

The "asymmetric dominance" attribute of algorithmic discipline: From a deeper theoretical logic, this "non-significance" precisely exposes the terrifying nature of algorithmic media materiality—technical discipline has surpassed the defensive boundaries of rational cognition. As N. Katherine Hayles noted in How We Think, the reshaping of neural circuits is physiological [10,11]. Faced with continuous high-frequency dopamine stimulation from algorithms, even if individuals possess high digital literacy, this rational "knowledge" is easily defeated by immense sensory temptations and emotional torrents, resulting in a "disconnect between knowledge and action".

In summary, although the moderating effect of digital information literacy did not show the expected significance in this statistics, it does not negate the value of literacy education. On the contrary, this empirical "accident" further warns educators: facing cognitive risks driven by algorithms, relying solely on students' original, fragmented information literacy is inadequate. Intervention strategies must be systematically upgraded, shifting the target from "information screening" entirely to "algorithmic logic de-blindness" and "emotional and behavioral interruption".

5 CONCLUSIONS, INNOVATIONS, AND LIMITATIONS

5.1 Research Conclusions and Innovations

This study innovatively introduces the KAP theory from the public health field into the interdisciplinary perspective of computational communication and cyberpsychology. It empirically reveals the deep chain mechanisms of university students' online mentality alienation driven by algorithms: irrational algorithmic trust directly triggers cognitive conflict (K) while bypassing the information cocoon, negative emotions (A), and ultimately leads to the eruption of uncontrolled internet use and defensive cognition (P). This research breaks through the previous macroscopic critical perspective on algorithmic risks. By utilizing Latent Profile Analysis (LPA) and Structural Equation Modeling (SEM), it provides micro-psychological data support for higher education institutions to accurately identify "high-risk/highly alienated" groups and implement targeted digital mental health and educational interventions.

5.2 Research Limitations and Future Prospects

5.2.1 Research limitations

Representativeness of sample structure: Restricted by non-probability sampling, although the distribution of institutional tiers (Project 985/211, first-tier, second-tier universities) among the undergraduate population (accounting for over 85%) matches the realistic ecology, there is an imbalance in educational and disciplinary backgrounds. The extreme scarcity of vocational students and doctoral candidates, coupled with a high proportion of economics and management majors, fails to form an absolutely ideal balance with humanities, STEM, and other disciplines. This, to some extent, limits the generalization validity of the model to the entire university student netizen population.

Temporal adaptability of measurement tools: The empirical results showed that the moderating effect of digital information literacy was not significant, exposing the potential limitations of the literacy scale adopted in this study. Existing scales mostly focus on surface-level skills such as traditional information retrieval and authenticity identification, while their measurement sensitivity to the underlying logics of the intelligent era, such as the hidden "algorithmic black box" and "data profiling," appears slightly insufficient.

5.2.2 Future prospects

Subsequent research should focus on expanding sample heterogeneity (especially increasing the proportion of vocational college students) and commit to developing and revising more localized and contemporary exclusive measurement tools for "Algorithm Literacy" among university students. Additionally, longitudinal tracking surveys or experimental intervention methods could be introduced to more accurately capture the dynamic causal mechanisms of digital literacy enhancement in interrupting cognitive dissonance and mentality alienation.

COMPETING INTERESTS

The author has no relevant financial or non-financial interests to disclose.

FUNDING

This research is a result of the project entitled "Study on the alienation of university students' online mentality and cognitive risk intervention under algorithmic drive" (Project No:25JCFCQUX082), sponsored by the innovation and development center for ideological and political work in colleges and universities of the ministry of education (Chongqing University) in 2025.

REFERENCES

[1] Huang Y, Liu L. The impact of algorithm awareness on the acceptance of personalized social media content recommendation based on the technology acceptance model. Acta Psychologica, 2025, 259: 105383.
[2] Dogruel L, Masur P, Joeckel S. Development and validation of an algorithm literacy scale for Internet users. Communication Methods and Measures, 2022, 16(2): 115-133.
[3] Metzger M J, Hartsell E H, Flanagin A J. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. Communication Research, 2020, 47(1): 3-28.
[4] Qiu L, Zheng X, Wang Y F. Revision of the Positive and Negative Affect Scale (PANAS). Applied Psychology, 2008, 14(3): 249-254.
[5] Liu Z S. Research on social media dependence and media needs: Taking university students' Weibo dependence as an example. Journalism University, 2013(1): 119-129.
[6] Quan F Y, Xia L X. The predictive effect of hostile attribution bias on reactive aggression and the mediating role of revenge motivation. Psychological Science, 2019, 42(6): 1434-1440.
[7] Li H, Li K Y, Hu X R, et al. Development and validation of the Information Literacy Measurement Scale (ILMS-34) in Chinese public health practitioners. BMC Medical Education, 2025, 25(1): 75.
[8] Rosenberg J, Beymer P, Anderson D, et al. tidyLPA: Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. 2021. https://CRAN.R-project.org/package=tidyLPA.

[9] Seol H. snowRMM: Rasch Mixture, LCA, and Test Equating Analysis (Version 5.9.1). 2025. https://github.com/hyunsooseol/snowRMM.
[10] Fornell C, Larcker D F. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 1981, 18(1): 39-50.
[11] Hayles N K. How We Think: Digital Media and Contemporary Technogenesis. Chicago: University of Chicago Press, 2012.