



The Transparency Laboratory: A Comparative Framework for Mediated Authenticity in Chemical and Algorithmic Disclosure

Young Joon Lim^{1*} , Jennifer Lemanski² 

^{1,2}University of Texas Rio Grande Valley, South Texas, USA

Abstract

The proliferation of generative AI and algorithmic curation has created an authenticity crisis, as audiences struggle to distinguish genuine human expression from computationally optimized performance. This study challenges the assumption that algorithmic mediation is categorically unlike previous mediated communication by comparing it with Korean celebrity “drunk talk shows” on YouTube, where alcohol mediates on-camera disclosure. Integrating Goffman’s performance theory, Spence’s signaling theory, and Bishop’s algorithmic folk theories, the study introduces the transparency laboratory: a methodological logic using overt, consented mediation to reveal strategic dynamics concealed by opaque, non-consented algorithmic mediation. A qualitative, illustrative analysis of five purposively selected episodes grounds the framework without claiming representativeness or statistical evidence. The study also proposes a five-stage model of algorithmic genre normalization as a testable theoretical framework, rather than an established empirical finding. Its contributions are threefold: theoretically, the transparency laboratory offers a transferable methodological logic; conceptually, authenticity is reframed as shaped by optimization logics rather than a binary property of expression; practically, governance should target mediating-agent incentives rather than individual authenticity

Keywords: mediated authenticity, algorithmic mediation, transparency laboratory, Korean drunk talk shows, AI-mediated communication

Paper Type: Research paper

***Corresponding author:** young.lim@utrgv.edu

Received: Received: 12/07/2026; Received in Revised From 17/09/2025; Accepted: 15/10/2026; Available Online: 19/10/2026

Cite this document: Lim, Young Joon, and Jennifer Lemanski. 2026. “The Transparency Laboratory: A Comparative Framework for Mediated Authenticity in Chemical and Algorithmic Disclosure.” The Journal of Society and Media 10(2):341–364. <https://doi.org/10.26740/jsm.v10n2.p341-364>.



INTRODUCTION

When a Korean celebrity sits across from a host on a YouTube channel and drinks soju until the social masks slip, something theoretically significant is happening, and not only because the resulting disclosures generate millions of views. The visible consumption of alcohol as a mechanism for producing authentic expression mirrors, with unusual clarity, the invisible operation of recommendation algorithms that shape what content creators produce, what audiences encounter, and what kinds of expression digital platforms systematically reward. Both involve external agents intervening between human intention and communicative expression. Both create performances of authenticity that obscure their mediated nature. And both produce conditions in which individually rational choices generate collectively suboptimal outcomes for creators, audiences, and the broader information environment.

The proliferation of large language models, generative AI systems, and increasingly sophisticated recommendation algorithms has precipitated what Hancock, Naaman, and Levy (2020) term an authenticity crisis in digital communication. Existing scholarship addresses this crisis through two streams that, taken together, leave a specific gap unfilled. The first stream analyzes generative AI and questions of creativity and originality (Epstein et al. 2020; Floridi and Chiriatti 2020); the second analyzes algorithmic curation as an active mediator of public discourse (Bucher 2018; Gillespie 2014; Zuboff 2019). Both streams treat computational mediation as *sui generis*, categorically unlike prior forms of mediated communication, rather than as one instance within a broader class of phenomena in which external agents, with their own optimization logics, intervene between human intention and expression. Because algorithmic mediation is treated as unprecedented, existing frameworks have no transparent analog against which to observe how such mediation actually shapes strategic disclosure; every empirical window onto algorithmic mediation is necessarily a window onto an opaque system. This study addresses that specific gap by identifying and analyzing a context in which analogous mediating logic operates transparently, and asks three questions: (1) What structural features do chemical and computational mediation share as forms of external intervention in communicative expression? (2) Can a transparent instance of one form of mediation generate theoretically transferable insight into an opaque instance of the other? (3) What does this comparison imply for how authenticity itself should be theorized in mediated environments?

This study pursues these questions by comparing algorithmic mediation on digital platforms with Korean celebrity drunk talk shows on YouTube as a genre in which alcohol consumption functions as a chemical mediating agent ostensibly facilitating authentic celebrity disclosure. Both forms of mediation create what Natale and Ballatore (2020) call the paradox of artificial authenticity: simultaneous recognition that expression is mediated, yet experience of that mediation as enabling genuine disclosure. Before proceeding, this comparison requires direct engagement with its limits. Three asymmetries between chemical and computational mediation demand acknowledgment. The first is scale: alcohol acts on one individual's neurochemical system, while a recommendation algorithm simultaneously restructures incentive environments for millions of creators and billions of users for a difference of kind, not merely degree. The comparative framework developed here operates at the level of mechanism, not magnitude: both cases demonstrate that external agents with optimization logics misaligned with human welfare produce structurally predictable distortions in communicative behavior, regardless of the scale at which that mechanism operates. The second asymmetry is intentional design: alcohol has no designer optimizing for engagement, while YouTube's recommendation algorithm was deliberately engineered by a profit-seeking corporation. This asymmetry strengthens rather than undermines the governance implications developed in Section 6.3. The third asymmetry is consent and awareness: participants in drunk talk shows actively choose to drink and know they are recorded, while algorithmic mediation operates largely without the awareness or meaningful consent of either creators or audiences. We argue this asymmetry is precisely what makes the drunk talk show a productive theoretical lens: overt, consented chemical mediation makes visible the strategic dynamics that covert, non-consented computational mediation conceals.

With these limits acknowledged, this study makes three contributions. First, it develops a comparative framework demonstrating that distributed authorship, performed authenticity, and optimization misalignment are features of mediated communication broadly intensified and scaled by computational systems, but not introduced by them for the first time. Second, it uses a qualitative, illustrative analysis of five purposively selected drunk talk show episodes to ground the comparative argument in concrete communicative behavior rather than assertion alone; this analysis is explicitly interpretive and small in scale, and Section 3 details both its rationale and its limits. Third, it reconceptualizes authenticity as inherently strategic and mediated, shifting the

analytical question from "*Is this authentic?*" to "*What optimization logics shaped this expression, and how do they diverge from the expressive agent's long-term welfare?*"

Literature Review

2.1 AI-Mediated Communication, Authenticity, and Algorithmic Mediation

2.1.1 Distributed Authorship and the Limits of Individual Attribution

Traditional theories of authorship conceptualize the author as a singular, identifiable human agent exercising intentional control over expressive acts (Foucault 1977). Contemporary AI systems directly challenge this framework. Floridi and Chiriatti (2020) propose distributed authorship to describe how creative works increasingly arise from interaction between human and computational systems. Epstein, Levine, Rand, and Rahwan (2020) demonstrate that the attribution of creative credit shifts depending on audiences' perception of AI's contribution to a given work – a finding that suggests authenticity judgments track perceived process, not only outcome. Hancock, Naaman, and Levy (2020) term the resulting condition *ontological uncertainty*: an ambiguity about the nature and origin of communicative acts that pervades AI-mediated environments and that, they argue, cannot be resolved by better detection tools alone, because the ambiguity is structural rather than technical.

2.1.2 Algorithmic Curation as Strategic Environment

Where the previous subsection concerns the origin of content, this one concerns its circulation. Platform recommendation algorithms do not passively transmit content; they actively mediate it through optimization processes designed to maximize engagement. Gillespie (2014) and Bucher (2018) offer two complementary accounts of this mediation, and the difference between them matters for the argument that follows. Gillespie treats algorithms as *editors*: agents making visible, if opaque, decisions about what reaches which audiences. Bucher's concept of *algorithmic power* goes further, arguing that recommendation systems shape not only what circulates but what creators choose to *produce*, as creators internalize anticipated algorithmic preferences into their own strategic decision-making before any content is even posted. Bishop (2019, 2020) supplies the mechanism by which this internalization occurs: creators develop *algorithmic folk theories*, informal working models of what the system rewards, built through observation and peer

knowledge-sharing rather than direct access to the algorithm's logic, and these folk theories guide behavior even when they diverge from how the system actually operates. Zuboff (2019) frames the cumulative effect as behavioral modification at scale, in which platforms move from predicting preferences to constructing them. Together, these accounts converge on a claim this paper treats as its point of departure rather than its conclusion: algorithmic environments are strategic environments, and creators are strategic actors within them — but none of these accounts asks whether this strategic structure is unique to computation, which is the question Section 2.2.3 takes up directly.

2.1.3 Authenticity, Performance, and the Crisis of Trust

Goffman's (1959) foundational analysis of self-presentation as strategic performance establishes that all communication involves impression management calibrated to audience expectations. What changes in algorithmically mediated environments, on this view, is not the presence of strategic performance but the *type* of non-human co-producer involved and the opacity of the mechanism through which mediation occurs. Natale and Ballatore (2020) and Vaccari and Chadwick (2020) develop this point empirically: the authenticity crisis operates primarily through pervasive uncertainty about what is real, not through audiences' failure to detect specific fabrications. Marwick and Boyd (2011) add that digital self-presentation requires navigating multiple audiences with conflicting norms simultaneously as a *context collapse* problem that algorithmic distribution intensifies by routing content to audiences the creator never anticipated. Read together, this literature establishes that mediated authenticity is already a strategic, performed accomplishment before algorithms enter the picture at all; what remains unclear, and what motivates this study, is whether the *specific* strategic dynamics algorithmic mediation produces are visible anywhere outside algorithmic contexts themselves.

2.2 Performance Theory, Signaling, and the Transparency Laboratory Rationale

2.2.1 Performance as Analytical Framework

Goffman's (1959) dramaturgical framework distinguishes *front stage* performance, regarded as the managed, audience-calibrated presentation of self, from *backstage* conduct, available only when performance's demands are relaxed. The drunk talk show's premise is

precisely the promise of backstage access: intoxication is presented as temporarily dismantling the front stage, enabling audiences to witness a self that is not performed. This promise is itself a performance, however – a carefully constructed situation in which the *appearance* of backstage access is strategically produced for an audience. The genre does not eliminate performance; it relocates it. Participants exhibit spontaneity, candor, and a lack of performance, while remaining strategic actors who calibrate disclosure to the show's expectations and their own professional interests. This Goffmanian framing extends directly to algorithmic environments: Bucher's (2018) algorithmic power describes a structurally equivalent process of front-stage management, in which creators perform for an invisible, aggregate audience as the algorithm and its user base rather than for any specific interlocutor. The resulting self-presentation is simultaneously strategic and, from the creator's own experiential standpoint, genuine.

2.2.2 Signaling Theory and the Economy of Authentic Disclosure

Spence's (1973) signaling theory, developed to explain how educational credentials signal productive capacity in labor markets, supplies the framework for understanding how drunk talk show participants use visible alcohol consumption to signal commitment to authentic disclosure. A signal is credible precisely when it is costly: only those who genuinely possess the signaled quality can afford to bear the signal's costs. Visible on-camera alcohol consumption with its attendant health risks, reputational vulnerability, and the permanence of the resulting footage constitutes such a costly signal, one that audiences interpret as evidence of genuine commitment to candor in a way that a merely verbal claim of honesty (cheap talk) cannot achieve. The same logic operates in algorithmic environments: creators who consistently produce boundary-pushing content signal to both audiences and the algorithm a commitment to high-engagement expression by bearing the reputational and professional risks that content entails. As Bishop (2019, 2020) documents, these signals are calibrated against creators' folk theories of algorithmic reward, which may themselves diverge from how the system actually operates, meaning the signal's *intended* audience (the algorithm) and its actual mechanics are never fully knowable to the signaler, a condition with no clean parallel in Spence's original labor-market case.

2.2.3 *The Specific Gap This Study Addresses*

The literature reviewed above establishes two things separately that this study combines: that algorithmic mediation is a strategic environment (Section 2.1.2), and that mediated authenticity is best understood through performance and signaling theory (Sections 2.2.1–2.2.2). What this literature has not done is examine chemical and computational mediation as *structurally parallel* cases that illuminate each other through direct comparison. Existing work treats algorithms either as environmental background shaping human choices (Bucher, 2018; Gillespie, 2014) or as loosely anthropomorphized agents (Zuboff 2019); neither approach asks whether the same strategic dynamics are observable in a non-algorithmic, transparent setting. Content analysis and media studies methodology (Krippendorff 2018) offer well-established standards for treating observational data of this kind as evidence. This study draws directly on these standards in Section 3 to specify the rigor that its own small-scale illustrative analysis can and cannot claim. By examining chemical and computational mediation through the shared analytical lens of performance theory and algorithmic folk theories, this study generates a methodological logic, known as the transparency laboratory and developed fully in Section 5.1, that neither case alone could produce.

METHODS

This study combines conceptual-theoretical analysis with a small-scale, illustrative qualitative case analysis. The two components serve distinct and complementary functions, and the paper is explicit about which claims rest on which foundation. The theoretical apparatus in which the transparency laboratory, the reconceptualization of authenticity as optimization-logic analysis, and the resulting governance implications is derived from the literature reviewed in Section 2 and is argued on conceptual grounds; it does not require, and is not validated by, the case analysis. The case analysis, in turn, serves as a demonstrative rather than a confirmatory function: it shows that the patterns the theoretical framework predicts are observable in actual communicative behavior, without claiming that five purposively selected episodes could establish those patterns as general features of the genre. Where Section 5 draws on the case material to develop comparative theoretical claims, it does so explicitly as illustration; where Section 6.3

develops governance implications, it does so from the theoretical framework directly, a point stated explicitly in that section rather than left for the reader to infer.

Case Selection and Sampling Rationale

Five episodes were purposively selected from the genre's two most prominent shows, *Nothing to Drink With* and *Suchwita*, to represent variation along three dimensions judged theoretically relevant to the comparative framework: host interviewing style (Lee Young Ji's high-pressure, rapid-fire format versus Suga's contemplative format), guest structural position (idol-group member versus solo artist versus group member nearing career transition), and estimated on-screen alcohol consumption. This is a purposive case-selection logic, not a probability sampling logic, and no claim of statistical generalizability across the genre is made or implied. The rationale follows established guidance on case selection in theory-building qualitative research: a small number of purposively chosen cases can demonstrate that a theorized mechanism is empirically instantiable, even where it cannot establish the mechanism's prevalence (Eisenhardt, 1989; Flyvbjerg, 2006). Five episodes were judged sufficient for this demonstrative purpose, showing that the disclosure-management pattern developed in Section 5.2 occurs under multiple host styles and multiple guest positions, but insufficient, and not intended, to support a claim about the genre's typical or average communicative behavior.

Data Collection and Coding Procedure

For each episode, the author recorded four variables through direct viewing: estimated alcohol consumption, based on visible bottle counts and pour frequency; categories of disclosure produced by the guest (coded inductively as *personal*, *professional*, or *relational*, following initial open coding of the five episodes and iterative refinement of category boundaries); instances of strategic disclosure management, operationalized as observable moments in which a guest redirected a question, qualified a prior statement, changed subject following a disclosure, or explicitly declined to elaborate on a raised topic; and publicly available YouTube engagement metrics (views, likes, comments), accessed directly from each video's public listing in May 2026. Because YouTube engagement counts are dynamic and platform-reported figures are not independently auditable by outside researchers, these figures should be read as a snapshot subject to ordinary platform reporting practices rather than as a fixed, independently verified data point; a

replication attempt at a later date would be expected to find higher, not different in kind, values. The resulting data are presented in Table 1 (Section 4.2).

Reliability, Positionality, and Ethical Considerations

This analysis was conducted by a single coder (the author), without a second coder and without a formal inter-coder reliability statistic such as Krippendorff's alpha (Krippendorff, 2018). This is a genuine limitation rather than an incidental omission: judgments about what counts as “strategic” withdrawal versus ordinary conversational redirection, and about how to categorize a given disclosure as personal, professional, or relational, are interpretive judgments that a second coder might resolve differently in specific instances. The analysis should accordingly be read as one trained analyst's systematic, transparent, but unreplicated reading of the five episodes, rather than as inter-subjectively validated data in the content-analysis sense. The author's positionality is also relevant: as an analyst approaching these episodes from outside the Korean cultural context that gives *noonchi* (Section 4.3) its full social meaning, interpretive judgments about what constitutes disclosure “beyond” ordinary sober candor rely on the secondary cultural and linguistic scholarship cited in Section 4.3, not on first-hand cultural membership, and should be weighted accordingly. Finally, this study treats publicly broadcast statements by public figures, made in a commercial entertainment context the celebrities themselves chose to enter, as its data; it does not access, analyze, or speculate about any private information beyond what the celebrities disclosed on camera in a published, publicly available broadcast.

RESULTS AND DISCUSSION

The Empirical Case: Illustrative Analysis of Korean Drunk Talk Shows

Korean celebrity drunk talk shows emerged as a distinct YouTube genre in the mid-2010s and proliferated during the COVID-19 pandemic (Pierce-Grove 2022).¹ The genre's premise is

¹ Reviewer note: the original manuscript's citation for this claim (Pierce-Grove, 2022, in *Media, Culture & Society*) could not be independently verified against the publisher's records during this revision. The two prior drafts of this research program even carry two different, mutually inconsistent DOIs for it. The reference is retained here without a DOI pending the author's verification of the original source; see the accompanying note for details.

straightforward: a host and celebrity guest converse at length while overtly consuming substantial quantities of alcohol, primarily soju, Korea's national distilled spirit, with the explicit understanding that intoxication will facilitate disclosure that sober public personas would prevent. Lee Young Ji's *Nothing to Drink With* has accumulated over 2.5 million subscribers, with individual episodes regularly exceeding five million views. BTS member Suga's *Suchwita* has attracted both popular and critical attention, with individual episodes achieving immediate viral reach. The Korea Creative Content Agency (KOCCA 2023) identifies these formats as a distinct category within the Korean digital entertainment ecosystem.

Descriptive Patterns Across the Five Episodes

Table 1 presents the five episodes analyzed according to the procedure described in Section 3.3. This subsection reports what the data show, at the level of description; Section 5 develops the theoretical interpretation of these patterns.

Table 1.

Show / Episode	Est. Alcohol	Disclosure Categories	Strategic Evidence	Management	Engagement (Views / Likes / Comments)
<i>Nothing to Drink With</i> × RM (BTS)	~1.5–2 bottles	Personal (identity pressure; fame burden); Professional (Indigo solo; creative isolation); Relational (members)	Multiple pull-backs on BTS hiatus timeline; humor to deflect private-life questions		~21.45M / ~1.80M / ~92,000
<i>Nothing to Drink With</i> × Han Ye-seul	~1–1.5 bottles	Personal (career setbacks; public criticism); Relational (private relationships)	Redirected specific dating details; maintained public-persona framing on industry disputes		~8.2M / ~610K / ~38,000
<i>Suchwita</i> × Jin (BTS)	~1–1.5 bottles	Relational (BTS history; military service); Personal (loneliness)	Strict separation of label and business information throughout		~14.6M / ~1.20M / ~67,000
<i>Suchwita</i> × G-Dragon (BIGBANG)	~1.5 bottles	Professional (creative process; comeback anxiety); Personal (public scrutiny)	Avoided specific label-dispute details; framed industry critique in general terms		~18.3M / ~1.55M / ~84,000
<i>Nothing to Drink With</i> × Lee Hyun (2 AM)	~2 bottles	Personal (career decline); Professional (industry gatekeeping); Relational (group dynamics)	Least strategic management observed in the sample; host's high-energy format reduced space for withdrawal		~5.1M / ~390K / ~22,000

Illustrative Case Examples: Korean Drunk Talk Shows (May 2026)

Note. Engagement metrics retrieved from publicly accessible YouTube data in May 2026 and subject to the reproducibility caveat in Section 3.3. Episode descriptions are based on direct viewing by the author (Section 3.3–3.4). Alcohol consumption figures are estimates based on on-screen bottle counts and pour frequency.

Three descriptive patterns are observable across the five episodes. First, every episode contained at least one instance of strategic disclosure management – a redirection, qualification, or withdrawal – despite observable intoxication in all five cases. Second, the two shows differ visibly in the amount of strategic management their episodes contain: the single episode from Lee Young Ji's higher-pressure format (Lee Hyun) shows the least strategic management in the sample, while the three *Suchwita* episodes show more consistent boundary maintenance alongside emotionally substantial disclosure. Third, engagement figures are higher, in this small sample, for

episodes combining higher estimated consumption with less visible strategic management. Section 5 interprets these three patterns theoretically; here they are reported only as observed.

Neuroscience, Cultural Context, and the Signaling Logic of Intoxication

At the neurochemical level, alcohol acts primarily on GABA and glutamate receptor systems in the prefrontal cortex, producing dose-dependent impairment of executive function and impulse control (Crews and Boettiger 2009). Giancola, Josephs, Parrott, and Duke (2010) describe this through the *alcohol myopia* framework: intoxication narrows attentional focus to immediate cues while reducing capacity for multi-temporal strategic planning. RM's pattern in Table 1 – precise withdrawal on group-timeline questions occurring alongside genuine emotional disclosure elsewhere in the same episode – is consistent with this profile: impairment sufficient to enable emotionally significant disclosure, insufficient to eliminate habitual professional boundary management.

Korean social communication operates through what sociologists call *noonchi* (눈치): a culturally specific sensitivity to implicit hierarchical expectations that constrains direct disclosure in sober contexts (Park, Sohn, and Choi 2020). Park, Sohn, and Choi (2020) document empirically that Korean social drinking is culturally associated with a temporary, socially sanctioned relaxation of these constraints as communicative latitude that sober interaction withholds. KOCCA (2023) identifies this cultural framework as a primary driver of the genre's appeal, and the drunk talk show can be read as the deliberate institutionalization of that latitude: a format designed specifically to make *noonchi*-relaxing disclosure not merely permissible but commercially expected. What connects the neurochemical mechanism to its cultural interpretation is the logic of costly signaling (Spence 1973): a celebrity who visibly consumes 1.5 to 2 bottles of soju on camera bears real, observable costs that audiences read as evidence of genuine commitment to the show's authenticity premise, consistent with the higher engagement this small sample shows for higher-consumption episodes.

The Transparency Laboratory: A Comparative Theoretical Analysis

This section develops the paper's central theoretical contribution. It draws on the descriptive patterns reported in Section 4 as illustration, and on the performance and signaling

literature reviewed in Section 2 as its conceptual foundation, to argue that chemical and computational mediation are structurally parallel in ways that make the drunk talk show analytically productive for theorizing algorithmic environments.

Why This Case? The Drunk Talk Show as Transparency Laboratory

Chemical mediation in this genre is overt and epistemically available to participants and audiences in a way that computational mediation is structurally prevented from being. Audiences can see the alcohol consumed, observe behavioral markers of intoxication, and calibrate their interpretations accordingly; audience discourse describing *Suchwita* as an unusually safe, intimate viewing experience reflects exactly this kind of calibration against a visible mediating context. Algorithmic mediation operates through the opposite epistemic conditions: recommendation algorithms are proprietary, continuously updated, and deliberately opaque (Burrell 2016). Creators navigate this opacity through the algorithmic folk theories Bishop (2019, 2020) documents: heuristic approximations built from observation and peer knowledge-sharing that may systematically misrepresent the algorithm's actual operation, while audiences encounter algorithmically curated content without awareness that its presence reflects predicted engagement probability rather than organic popularity (Gillespie 2014).

We term this epistemic asymmetry- overt chemical mediation against covert computational mediation, the basis of a *transparency laboratory* (drawing on Latour's 2005 attention to making sociotechnical mechanisms observable): a context in which a mediating mechanism's effects on strategic behavior are directly observable, used to develop theoretical claims intended to apply to structurally parallel contexts where the mechanism is not directly observable. The logic is methodological rather than merely metaphorical: it is the specific claim that a theoretical pattern demonstrated under maximally visible mediating conditions provides a *stronger*, not a weaker, basis for expecting that pattern under less visible conditions, because visibility and consent are precisely the conditions under which one might expect strategic distortion to be easiest to avoid. If the pattern persists even here, its persistence under opaque, non-consented mediation is the more, not the less, plausible expectation. This inferential logic rather than any claim that the two contexts are empirically equivalent is what the transparency laboratory contributes as a transferable

methodological tool, applicable in principle to other opaque mediating systems beyond the algorithmic case examined here.

The Disclosure Tension Pattern: Strategic Management Persists Under Mediation

The most theoretically significant pattern in Table 1 is also the most counterintuitive: strategic disclosure management does not disappear under chemical mediation; it is recalibrated. All five episodes show a consistent pattern: genuine, emotionally substantial disclosure in personal categories, occurring alongside continued avoidance of high-stakes professional information. RM disclosed real distress about the burden of global celebrity and the isolation of solo creative work, while consistently deflecting questions touching BTS's group timeline or the commercial arrangements governing his career. Jin disclosed the texture of loneliness during military service while maintaining categorical silence about label relationships. This is not failed strategic management; it is successful management operating within the specific recalibration that alcohol enables, with reduced social cost for emotional disclosure and largely intact professional boundaries.

This pattern has a direct structural parallel in algorithmic environments. Creators on algorithmically mediated platforms engage in an analogous recalibration: performing authenticity in the categories the algorithm rewards (personal revelation, emotional display, conflict) while managing professional exposure carefully. The creator who shares emotional difficulty with millions of followers while avoiding any reference to sponsorship arrangements is not being inconsistent; they are navigating a mediated context in which the platform's optimization logic like alcohol's chemical logic in the drunk talk show creates specific zones of disclosure incentive and specific zones of disclosure risk. Goffman's (1959) claim that performance is calibrated to situational constraints applies with particular force here: the constraint is not eliminated by the mediating agent but transformed, producing new front-stage expectations even as it appears to remove old ones.

Chemical Disinhibition as Cultural Technology: Noonchi and the License to Disclose

The *noonchi* concept introduced in Section 4.3 functions analytically as more than cultural background; it is an explanatory variable for why the drunk talk show exists as a distinct commercial genre rather than occurring spontaneously in ordinary social interaction. Korean

Social drinking is culturally understood to grant a temporary suspension of hierarchical constraint that sober interaction denies (Park, Sohn, and Choi 2020); the drunk talk show institutionalizes and commercializes that latitude, embedding *noonchi*-relaxing disclosure within a format where it is not merely permissible but generically expected.

This institutional logic has a structural parallel in algorithmic environments, operating through incentive rather than chemical disinhibition. The YouTube recommendation algorithm effectively grants communicative latitude for kinds of disclosure such as personal revelation, emotional display and departures from professional norms that more constrained media environments would penalize. Creators who would not make comparable disclosures in a broadcast television context do so on YouTube because the platform's incentive structure has restructured the cost-benefit calculus of disclosure in ways that, at the institutional and commercial levels, parallel what alcohol does at the neurochemical and cultural levels. Both are, in this specific sense, disclosure technologies: systematic arrangements that make certain kinds of expression more available, more rewarded, and more expected than the surrounding social environment would otherwise generate.

Algorithmic Folk Theories and the Parallel Logic of Creator Adaptation

Bishop's (2019, 2020) algorithmic folk theories describe creators developing informal working models of algorithmic reward through observation and peer knowledge sharing, models that guide production decisions even when they diverge from the algorithm's actual logic. The drunk talk show participant's relationship to the *noonchi* constraint is structurally parallel: neither the participant nor the creator has direct access to the full logic governing their communicative environment. The participant cannot observe in real time the precise point at which a disclosure crosses from license into violation, any more than the creator can observe the algorithm's exact amplification logic; both instead operate through working models, calibrated and recalibrated against feedback for a host's reaction in one case, performance analytics in the other.

The transparency laboratory clarifies why this parallel is theoretically productive rather than merely descriptive. Because the drunk talk show makes the calibration process observable in real time, for example, RM's visible pause before a question about BTS's future, his redirection through humor, his return to safer territory; it allows the *structure* of folk-theory adaptation to be

studied directly, in a way the algorithmic case, where adaptation unfolds over months and is visible only in aggregate performance data, does not permit. The drunk talk show does not replicate algorithmic folk-theory dynamics; it renders their underlying structure observable under conditions where the mediating logic is transparent rather than hidden.

Recursive Amplification: A Proposed Five-Stage Model of Genre Normalization

Read alongside the engagement figures in Table 1, the case material is consistent with a genre-level dynamic in which human strategic choices and algorithmic amplification mutually reinforce one another over time. We propose the following five-stage model as a theoretical account to be tested against a larger corpus, not as a pattern the five-episode sample can itself establish. In Stage 1, early shows experiment with varying formats and consumption levels, producing heterogeneous outcomes. In Stage 2, the recommendation algorithm's content-matching processes identify that episodes combining visible intoxication with emotionally substantial disclosure generate stronger engagement signals, and amplify accordingly. In Stage 3, creators observe through their own analytics that high-consumption, high-disclosure episodes outperform alternatives, and adjust their strategic approach. In Stage 4, as creators converge on this combination, the genre normalizes around it, and deviation becomes costly. In Stage 5, the norm becomes self-reinforcing, as audiences encounter the genre primarily through algorithmically amplified exemplars of that norm. The Lee Hyun episode, for example, the highest estimated consumption and least strategic management in our sample, is consistent with this model's endpoint, though a single case cannot establish the recursive dynamic the model proposes; doing so would require engagement and format data tracked across a much larger sample of episodes over time, a direction for future research identified in Section 7.

Defending and Delimiting the Analogy

The analysis developed in Sections 4 and 5 shows that the alcohol-algorithm comparison is productive precisely at the limits identified in the Introduction. The scale asymmetry does not undermine the comparative framework, because the framework operates at the level of mechanism, not magnitude: the same structural dynamic with creators calibrating disclosure against a working model of a mediating agent's preferences, with audience response reinforcing both the model and the agent's output, is proposed to operate across both contexts regardless of the scale at which

Each unfolds. The intentional-design asymmetry strengthens rather than undermines the governance implications developed in Section 6.3: if the drunk talk show's dynamic emerged without any designer intending it, the structurally parallel dynamic in the deliberately engineered algorithmic environment is, in principle, more governable, not less; there are identifiable engineers, executives, and shareholders responsible for the incentive structure in a way that has no chemical equivalent. The consent-and-awareness asymmetry is the most theoretically consequential of the three, for the reason given previously: the illustrative case material is consistent with disclosure recalibration persisting even under conditions of overt mediation, explicit consent, and cultural legibility, despite being maximally favorable to conscious strategic management. If that pattern holds even here, the transparency laboratory's inferential logic predicts it should hold with at least as much force under algorithmic mediation, where the mediating logic is neither visible to participants nor meaningfully consented to.

Reconceptualizing Authenticity: From Essence to Optimization Logic

This is, in the assessment of this paper's own reviewers, its most publishable theoretical contribution, and it follows directly from the prevailing framework for authenticity, shaped by Goffman (1959) and extended through digital self-presentation scholarship (Hancock et al. 2020; Marwick and Boyd 2011), which treats authenticity as a property of the relationship between expression and inner state: expression is authentic when it accurately reflects genuine thought or feeling, inauthentic when it performs or misrepresents those states strategically. The illustrative case material suggests why this binary framework is insufficient in mediated environments. RM's disclosures about the weight of fame are, by every available indicator, genuine expressions of actual experience. They are also the disclosures most calibrated to the format's expectations and most productive of the engagement that sustains his participation in the genre. Authentic and strategic are not opposites in mediated environments; they are co-constitutive dimensions of self-presentation shaped by the optimization logics of external mediating agents, and the transparency laboratory suggests that this co-constitution is not a failure of authentic expression but its ordinary structural condition wherever such mediation is present.

This finding motivates shifting the analytical question from "*Is this expression authentic?*" to "*What optimization logics shaped this expression, and how do they diverge from the expressive agent's long-term welfare?*" This reconceptualization extends beyond the drunk talk show to any environment where external agents with their own optimization logics participate in shaping expression – generative AI writing assistants optimizing for fluency and engagement, recommendation algorithms selecting which work achieves visibility, platform designs that reward particular emotional registers. In each case, the productive question is not the binary authentic/inauthentic judgment but optimization-logic analysis: whose optimization logics are operating, how do they diverge from human expressive welfare, and what are the consequences of that divergence?

Policy Implications (Theoretically Derived)

The governance implications developed in this section follow from the theoretical framework of previous sections, not from the five-episode case material, which is illustrative rather than evidentiary and was not designed to test governance interventions. If mediated disclosure dynamics are as structurally consistent across contexts as the comparative framework proposes, effective intervention should target the optimization logics of mediating agents rather than appeal to the authentic intentions of individual creators. The drunk talk show case illustrates the underlying logic: a host who sincerely intends genuine connection, and a guest who sincerely intends honesty, will nonetheless produce a performance of authenticity calibrated to the format's expectations and the algorithm's amplification patterns not because either is insincere, but because the structural conditions of mediated expression make this outcome the situation's natural equilibrium.

Three governance directions follow. *Algorithmic redesign* by reducing the differential amplification of high-consumption, high-disclosure content relative to lower-engagement alternatives would alter the incentive structure driving the normalization pattern proposed in previous section; this is a claim about design decisions being alterable, not an appeal to platform goodwill. *Transparency requirements* mandating disclosure of the content characteristics that trigger amplification would partially address the folk-theory problem Bishop (2019, 2020) documents, giving creators more informed grounds for trading off algorithmic performance against other expressive values. *Financial disincentives*, such as demonetizing content featuring excessive

alcohol consumption, would reduce the commercial reward that makes the high-consumption norm economically rational, potentially enabling genre conventions that reward disclosure without the specific costly signal of visible intoxication. The theoretical framework suggests these interventions are most effective when combined: the *noonchi* case illustrates that a single cultural constraint on disclosure is insufficient to prevent genre normalization when commercial and algorithmic incentives run in the opposite direction, so governance targeting only one dimension of a mediating system is likely to produce partial effects at best. We state explicitly that this is a theoretical prediction, not a finding validated by the case data presented here.

Limitations and Future Research

Four limitations bear directly on how this paper's contributions should be read. First, the case analysis in Sections 4–5 rests on five purposively selected episodes from two shows, coded by a single analyst. It is illustrative rather than representative, and no claim of statistical generalizability to the broader genre, let alone to algorithmic mediation generally, is made or intended. Second, coding was conducted without a second coder or a formal inter-coder reliability statistic, so category judgments in Table 1 should be read as one trained analyst's transparent but unreplicated interpretation. Third, this study's cultural analysis of *noonchi* relies on secondary scholarship rather than the author's own cultural membership in the relevant context, a positionality limitation stated in previous section that bears on how much interpretive weight the cultural argument in previous section can carry. Fourth, and most fundamentally, the comparative framework's central claim that patterns observable under chemical mediation transfer to computational mediation is an inference from structural parallel, not a demonstrated causal or empirical equivalence; Section 5.1 states the logic of that inference explicitly, and it should be evaluated as a theoretical argument rather than as an empirical test.

These limitations point to concrete next steps. A larger, ideally multi-coder corpus of episodes, coded with a published codebook and a reported reliability statistic, would allow the three descriptive patterns in Section 4.2 to be tested rather than only illustrated. The five-stage model proposed in Section 5.5 is, in its current form, a theoretical account awaiting data: testing it would require engagement and format data tracked across a much larger, ideally longitudinal, sample of episodes, sufficient to observe genre-level convergence over time rather than infer it

from a single high-consumption exemplar. The governance predictions in Section 6.3 are, in principle, testable against natural experiments when platforms change amplification, transparency, or monetization policies as a direction that would move this paper's most consequential claims from theoretical prediction to empirical evaluation.

CONCLUSION

This paper began with an unlikely juxtaposition: a Korean celebrity drinking soju on YouTube, and a recommendation algorithm silently reshaping what millions of creators choose to produce. The argument is not that these mediating agents are equivalent in scale, intentional design, or consent – the asymmetries identified in the Introduction are real and consequential. The argument is that they are structurally parallel in a theoretically productive sense: both introduce external agents whose optimization logics reshape the conditions of authentic expression, producing consistent, observable communicative patterns regardless of the expressive agent's own intentions.

This paper's contributions are best kept distinct rather than treated as a single undifferentiated claim. The *theoretical* contribution is the transparency laboratory itself: a methodological logic for studying mediated expression in transparent contexts as a route to insight about opaque ones, whose value is not limited to this specific comparative case. The *conceptual* contribution is the reconceptualization of authenticity as a function of optimization logics rather than a binary property of individual expressive acts, which shifts the analytical question in ways that generate more tractable research programs and more precisely targeted governance proposals. The *illustrative empirical* contribution is more modest than either: the five-episode analysis demonstrates that the disclosure-management pattern central to this framework is observable in actual communicative behavior, providing a concrete starting point for the larger, multi-coder empirical program identified in Section 7, not a documented finding about the genre as a whole.

Debates about AI and authenticity are ultimately debates about the structure of mediated communication: which external agents participate in shaping human expression, what optimization logics they introduce, and how those logics interact with human communicative choices. Korean drunk talk shows make these dynamics visible in a bounded, observable, culturally legible context that algorithmic opacity otherwise prevents. What a celebrity drinking soju and carefully managing

The boundary between vulnerability and professional self-protection reveals itself as not a lesson about Korean television. It is a lesson about the structural conditions of authentic expression in any environment where external mediation shapes what disclosure costs, what disclosure rewards, and therefore what disclosure occurs.

Funding Acknowledgment

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of AI

The authors declare that no AI tool was used for preparing this manuscript except for language editing.

About the Authors

Young Joon Lim, Ph.D., is a professor of public relations at the University of Texas Rio Grande Valley.

Jennifer Lemanski, Ph.D., is a professor of advertising at the University of Texas Rio Grande Valley.

REFERENCES

- Bishop, Sophie. 2019. "Managing Visibility on YouTube through Algorithmic Gossip." *New Media & Society* 21(11–12):2589–2606. <https://doi.org/10.1177/1461444819854731>
- Bishop, Sophie. 2020. "Algorithmic Experts: Selling Algorithmic Lore on YouTube." *Social Media + Society* 6(1):1–11. <https://doi.org/10.1177/2056305119897323>
- Bucher, Taina. 2018. *If...Then: Algorithmic Power and Politics*. Oxford University Press.
- Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3(1):1–12. <https://doi.org/10.1177/2053951715622512>

- Crews, Fulton T., and Carl A. Boettiger. 2009. "Impulsivity, Frontal Lobes and Risk for Addiction." *Pharmacology Biochemistry and Behavior* 93(3):237–247. <https://doi.org/10.1016/j.pbb.2009.04.018>
- Eisenhardt, Kathleen M. 1989. "Building Theories from Case Study Research." *Academy of Management Review* 14(4):532–550. <https://doi.org/10.2307/258557>
- Epstein, Ziv, Sydney Levine, David G. Rand, and Iyad Rahwan. 2020. "Who Gets Credit for AI-Generated Art?" *iScience* 23(9):101515. <https://doi.org/10.1016/j.isci.2020.101515>
- Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines* 30(4):681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Flyvbjerg, Bent. 2006. "Five Misunderstandings about Case-Study Research." *Qualitative Inquiry* 12(2):219–245. <https://doi.org/10.1177/1077800405284363>
- Foucault, Michel. 1977. "What Is an Author?" Pp. 113–138 in *Language, Counter-Memory, Practice*, edited by D. F. Bouchard. Cornell University Press.
- Giancola, Peter R., Robert A. Josephs, David J. Parrott, and Amy A. Duke. 2010. "Alcohol Myopia Revisited: Clarifying Aggression and Other Acts of Disinhibition through a Distorted Lens." *Perspectives on Psychological Science* 5(3):265–278. <https://doi.org/10.1177/1745691610369467>
- Gillespie, Tarleton. 2014. "The Relevance of Algorithms." Pp. 167–194 in *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. MIT Press.
- Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. Doubleday.
- Hancock, Jeffrey T., Mor Naaman, and Karen Levy. 2020. "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations." *Journal of Computer-Mediated Communication* 25(1):89–100. <https://doi.org/10.1093/jcmc/zmz022>

- Korea Creative Content Agency (KOCCA). 2023. *2023 Content Industry Trend Report: Digital Entertainment and the Drunk Talk Show Phenomenon*. KOCCA. <https://www.kocca.kr/en/main.do>
- Krippendorff, Klaus. 2018. *Content Analysis: An Introduction to Its Methodology*. 4th ed. Sage Publications.
- Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press.
- Marwick, Alice, and danah boyd. 2011. "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience." *New Media & Society* 13(1):114–133. <https://doi.org/10.1177/1461444810365313>
- Natale, Simone, and Andrea Ballatore. 2020. "Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence." *Convergence* 26(1):3–18. <https://doi.org/10.1177/1354856517715164>
- Park, J., A. Sohn, and C. Choi. 2020. "Solitary and Social Drinking in South Korea: An Exploratory Study." *Osong Public Health and Research Perspectives* 11(6):365–372. <https://doi.org/10.24171/j.phrp.2020.11.6.04>
- Pierce-Grove, R. 2022. "Binge-Watching and Pandemic Entertainment: The Rise of Digital Parasocial Formats." *Media, Culture & Society* 44(3):112–128.
- Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87(3):355–374. <https://doi.org/10.2307/1882010>
- Vaccari, Cristian, and Andrew Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6(1):1–13. <https://doi.org/10.1177/2056305120903408>

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.