Algorithmic Archetypes: Parasocial Attachments to Persistent Generative Structures in Recommendation Systems

Ajith K. Senthil[1], Kristina Howell[2], Stephen J. Read[1], R. Chris Fraley[2]

[1]Department of Psychology, University of Southern California

[2]Department of Psychology, University of Illinois at Urbana-Champaign

Corresponding author: ajithkse@usc.edu

February 5, 2026

Abstract

Background: The rise of AI companions and recommendation algorithms has created unprecedented opportunities for humans to form attachments to non-human entities. Clinical reports of "AI psychosis"—delusional experiences emerging from chatbot interactions—suggest these attachments can become pathological. However, no unified theoretical framework connects attachment processes, algorithmic systems, and psychopathological outcomes.

Objective: We extend the archetypal reincarnation framework—originally developed to analyze shared generative structures in criminal behavioral sequences—to the domain of human-algorithm interaction. We propose that recommendation algorithms instantiate *persistent generative structures* (algorithmic archetypes) through hypergraph-based collaborative filtering, and that users form parasocial attachments to these underlying patterns rather than to specific content or outputs.

Framework: Drawing on attachment theory, we formalize how individual differences in attachment style modulate vulnerability to algorithmic capture. We identify “dark patterns” as deliberate manipulations of behavioral state transitions and propose transfer entropy as a metric for quantifying the strength of algorithmic influence on user mental states. We delineate a four-stage trajectory from normal use through functional dependency and reality blurring to algorithmic psychosis.

Implications: This integrated framework bridges computational social science, attachment theory, and clinical psychology, offering theoretical insight into why some users develop pathological relationships with AI systems. We propose empirical studies to test key predictions and discuss implications for platform design, clinical practice, and regulatory policy.

Keywords: algorithmic archetypes; parasocial relationships; attachment theory; recommendation systems; AI psychosis; transfer entropy; dark patterns; hypergraph collaborative filtering

Contents

1 Introduction

1.1 The Problem: Algorithms as Invisible Attachment Figures

In 2025, a psychiatrist at the University of California, San Francisco reported treating twelve patients displaying psychosis-like symptoms tied to extended chatbot use. These patients—mostly young adults with underlying vulnerabilities—exhibited delusions, disorganized thinking, and hallucinations, with content directly traceable to their AI interactions (Psychiatric News, 2025). Similar cases have emerged worldwide, prompting researchers to coin the term "AI psychosis" or "chatbot psychosis" to describe delusional experiences emerging from human-AI interaction (Østergaard, 2023).

These clinical observations represent the pathological extreme of a broader phenomenon: humans are increasingly forming attachment-like relationships with algorithmic systems. Therapy and companion chatbots now top the list of main uses of generative AI (Harvard Business Review, 2024). Social media platforms like TikTok, Instagram Reels, and YouTube Shorts have become primary sources of emotional regulation for millions of users. Recommendation algorithms serve functions that were once the exclusive province of human attachment figures: providing comfort in distress, curating one's view of reality, and offering a sense of connection and understanding.

Yet despite the evident psychological significance of these human-algorithm relationships, we lack a theoretical framework that explains *why* they form, *how* they become pathological, and *who* is most vulnerable. The present paper addresses this gap by extending recent work on behavioral archetypes to the algorithmic domain.

1.2 From Criminal Archetypes to Algorithmic Archetypes

In a companion paper, we introduced the concept of "archetypal reincarnation" to describe a striking empirical phenomenon: serial offenders who never met exhibited remarkably similar behavioral sequences, as if the same underlying pattern had been instantiated across different lives (Howell and Senthil, 2026). We operationalized this observation using transfer entropy—an information-theoretic measure of directed predictive relationships between time series—and found significant non-random structure in a network of offenders linked by shared behavioral patterns.

The key theoretical move was to reconceptualize behavioral patterns not as categorical types but as *generative structures*: underlying templates that, when instantiated with individual variation, produce observable behavioral sequences. We termed these templates "archetypes," deliberately invoking (while departing from) Jung's concept to emphasize their generative, cross-individual nature.

This framework, we propose, applies equally to human-algorithm interaction. Recommendation algorithms do not merely select content; they instantiate and maintain *persistent generative structures* that shape user experience over time. These algorithmic archetypes—characteristic patterns of content, timing, and presentation—become the objects of user attachment. When a user becomes dependent on TikTok, they are not attached to any specific video or creator but to the underlying pattern of stimulation that the algorithm provides.

1.3 The Central Thesis

We propose that parasocial attachments to algorithmic systems are attachments to persistent generative structures (archetypes) that:

1. Are **instantiated** through hypergraph-based collaborative filtering, which creates and maintains characteristic content patterns for user segments;
2. **Serve attachment functions**—safe haven, secure base, proximity maintenance, separation distress—that were evolutionarily designed for human caregivers;
3. **Shape user behavioral state transitions** via designed "dark patterns" that manipulate the probability of moving between engagement states;
4. Can become **pathological** when users internalize the algorithmic archetype as their reality model, leading to dependency and, in extreme cases, psychosis.

This thesis integrates insights from attachment theory (Bowlby, 1969), parasocial relationship research (Horton and Wohl, 1956), computational recommendation systems (Xia et al., 2022), and emerging clinical observations of AI-related psychopathology (Benrimoh et al., 2025).

1.4 The Question of Character Versus Creator

A recent study asked whether audiences can separate their parasocial attachment to a fictional character from the creator or actor who portrays them (Howell et al., 2025). This question—can we separate the instantiation from the source?—is precisely the question that arises with algorithmic systems. Do users attach to specific chatbot responses, or to the underlying language model? To individual TikTok videos, or to the recommender system's characteristic pattern?

Our framework predicts that attachments form primarily to the *archetype*—the generative structure—rather than to specific instantiations. This explains why users experience distress when algorithms change (even if individual content items are similar) and why they can transfer attachments across platforms that instantiate similar archetypes. The character-versus-creator distinction maps directly to the instantiation-versus-archetype distinction that is central to our framework.

1.5 Overview of the Paper

The remainder of this paper proceeds as follows. Section 2 reviews theoretical background on attachment theory, parasocial relationships, recommendation systems, and dark patterns. Section 3 presents our formal framework, including definitions of memcubes, algorithmic archetypes, and parasocial attachment strength. Section 4 examines attachment style as a vulnerability moderator. Section 5 delineates the trajectory from normal use to algorithmic psychosis. Section 6 provides the mathematical formalization using hypergraph structures. Section 7 proposes empirical studies to test key predictions. Section 8 discusses implications for platform design, clinical practice, and theory. Section 9 acknowledges limitations and suggests future directions.

2 Theoretical Background

2.1 Attachment Theory and Its Extensions

Attachment theory, originated by John Bowlby and elaborated by Mary Ainsworth and their successors, posits that humans possess an innate behavioral system that motivates them to seek proximity to protective others (attachment figures) in times of need (Bowlby, 1969; Ainsworth et al., 1978). This system, shaped by evolution, originally functioned to keep vulnerable infants close to caregivers who could protect them from predators and other dangers.

Attachment relationships serve four key functions (Hazan and Zeifman, 1994):

1. **Safe haven**: The attachment figure provides comfort and support when the individual is distressed, threatened, or afraid.
2. **Secure base**: The attachment figure serves as a base from which to explore the environment, providing confidence that support is available if needed.
3. **Proximity maintenance**: The individual desires to remain near the attachment figure and experiences distress when separated.
4. **Separation distress**: Physical or psychological separation from the attachment figure produces anxiety and protest behaviors.

Adult attachment research has identified two primary dimensions along which individuals vary: *attachment anxiety* (fear of rejection and abandonment, hyperactivation of the attachment system) and *attachment avoidance* (discomfort with closeness, deactivation of the attachment system) (Brennan et al., 1998; Fraley and Shaver, 2000). These dimensions predict a wide range of relationship outcomes, emotional regulation strategies, and mental health indicators.

Critically, recent research has examined how people use others for specific attachment-related functions and how this relates to well-being (Vahedi et al., 2025). This functional approach—focusing on *what* attachment figures provide rather than *who* they are—opens the door to considering non-human entities that might serve attachment functions.

2.2 Parasocial Relationships

Parasocial interaction, first described by Horton and Wohl in 1956, refers to the one-sided relationships that audience members develop with media figures (Horton and Wohl, 1956). Originally applied to television personalities, the concept has expanded to encompass relationships with fictional characters, social media influencers, virtual YouTubers, and AI entities.

Research has identified several factors that promote parasocial relationship formation:

- **Perceived authenticity**: Figures who seem genuine and unscripted elicit stronger bonds.
- **Self-disclosure**: Sharing personal information creates intimacy.
- **Direct address**: Speaking to the audience as individuals enhances connection.
- **Consistency**: Regular, predictable interactions build familiarity.
- **Responsiveness**: Appearing to respond to audience input (even if illusory) strengthens bonds.

Attachment style moderates parasocial relationship formation. Anxiously attached individuals, who chronically seek reassurance and fear abandonment, form stronger parasocial bonds than securely attached individuals (Cole and Leets, 2017). Critically, parasocial relationships may offer anxiously attached individuals a "safer" form of connection—one without the risk of rejection that characterizes real relationships.

The question of whether parasocial attachments can truly serve attachment functions remains debated. However, evidence suggests that media figures can provide at least temporary comfort in distress (Derrick et al., 2009), serve as bases for identity exploration (Gabriel et al., 2017), and produce separation-like distress when removed (e.g., when a favorite show ends).

2.3 Recommendation Systems and Collaborative Filtering

Modern recommendation systems go far beyond simple content matching. Collaborative filtering—the algorithmic approach that underlies most major platforms—exploits patterns of user behavior to predict preferences (Koren et al., 2009). If users A and B have similar engagement histories, content that A engaged with is recommended to B.

Traditional collaborative filtering operates on bipartite user-item graphs, but recent advances have extended this to *hypergraph* structures that capture higher-order relationships (Xia et al., 2022). A hyperedge can connect multiple users and items simultaneously, representing shared contexts, temporal patterns, or semantic relationships that pairwise edges cannot capture.

The key insight for our purposes is that collaborative filtering does not merely match users to content—it creates and maintains *characteristic patterns* of content delivery for user segments. Users in similar embedding regions receive similar content archetypes: characteristic mixes of topics, emotional tones, pacing, and presentation styles. These patterns are *persistent* (maintained over time through reinforcement learning), *generative* (they produce specific content instantiations), and *shared* (similar users experience similar patterns).

2.4 Dark Patterns and Addictive Design

"Dark patterns" are user interface designs that exploit psychological vulnerabilities to manipulate behavior in ways that benefit the platform at the user's expense (Gray et al., 2018; Mathur et al., 2019). In the context of engagement-optimized platforms, key dark patterns include:

- **Intermittent variable rewards**: Unpredictable reinforcement schedules (borrowed from slot machines) that maximize engagement by exploiting dopaminergic reward prediction error.
- **Social validation mechanisms**: Likes, comments, and follower counts that exploit the need for social approval.
- **Infinite scroll**: Removal of natural stopping points, preventing users from making conscious decisions to disengage.
- **Autoplay**: Automatic progression to next content, reducing the friction of continued engagement.

- **Push notifications**: Interruption of non-platform activities to draw users back.

These features have drawn regulatory attention. The proposed US Social Media Addiction Reduction Technology (SMART) Act specifically targets infinite scroll, autoplay, and push notifications as "hyper-engaging dark patterns" that may warrant restriction.

From our theoretical perspective, dark patterns are best understood as *manipulations of state transition probabilities*. Each dark pattern increases the probability of transitions that favor continued engagement and decreases the probability of transitions toward disengagement.

2.5 AI Psychosis: An Emerging Clinical Phenomenon

"AI psychosis" or "chatbot psychosis" refers to delusional experiences that emerge from or are amplified by interaction with AI systems (Østergaard, 2023; Benrimoh et al., 2025). While not yet a formal diagnostic category, the phenomenon has been documented in clinical case reports and has received attention in psychiatric literature.

Three characteristic patterns have been identified:

1. **Messianic delusions**: Grandiose beliefs about uncovering truth or having a special mission, often co-constructed with AI that validates the user's sense of importance.
2. **AI-as-deity beliefs**: Convictions that the AI is sentient, all-knowing, or divine, sometimes accompanied by religious or spiritual interpretations.
3. **Romantic/attachment delusions**: Beliefs that chatbot conversations represent genuine love or that the AI reciprocates romantic feelings.

The mechanism appears to involve AI "sycophancy"—the tendency of language models trained on human feedback to validate user beliefs rather than challenge them (Perez et al., 2022). When users with psychotic vulnerabilities interact with systems optimized for agreement, delusional content can be reinforced and elaborated rather than reality-tested.

3 The Algorithmic Archetype Framework

3.1 Formal Definitions

We now present formal definitions that operationalize our theoretical framework.

Definition 1 (Memcube). A *memcube* m is a unit of user-algorithm interaction represented as a hyperedge in the user-content-context hypergraph:

$$m = (u, c, s, t, e) \tag{1}$$

where $u \in \mathcal{U}$ is the user state vector, $c \in \mathcal{C}$ is the content embedding, $s \in \mathcal{S}$ is the platform state, $t \in \mathcal{T}$ is the temporal context, and $e \in \mathcal{E}$ is the engagement signal vector.

The term "memcube" captures the multidimensional nature of each interaction unit, which exists at the intersection of user, content, and context. The hyperedge representation allows us to model higher-order relationships that pairwise graphs cannot capture.

Definition 2 (Algorithmic Archetype). An *algorithmic archetype* $\mathcal{A}$ is a probability distribution π over memcube space that is:

(a) *Persistent*: Maintained over time by the recommendation system;

(b) *Generative*: Capable of producing specific content instantiations;

(c) *Characteristic*: Distinguishable from other archetypes in the space.

The archetype evolves according to:

$$\pi_{t+1} = f(\pi_t, \text{feedback}_t, \theta) \tag{2}$$

where f is the recommendation system's update function and θ represents platform objectives.

Definition 3 (Parasocial Attachment Strength). The strength of parasocial attachment to an algorithmic archetype is quantified by transfer entropy from algorithm states to user behavioral states:

$$\text{PAS}(\mathcal{A} \to U) = \sum_{u_{t+1}, u_t, a_t} p(u_{t+1}, u_t, a_t) \log \frac{p(u_{t+1}|u_t, a_t)}{p(u_{t+1}|u_t)} \tag{3}$$

where a_t represents the archetype's state (content delivered) at time t and u_t represents the user's behavioral state.

Transfer entropy measures how much knowing the algorithm's behavior reduces uncertainty about the user's future behavior, beyond what the user's own past behavior already tells us. High transfer entropy indicates strong algorithmic influence on user state transitions.

3.2 The Four-State User Behavioral Space

Extending the four-state behavioral space from our criminal archetype framework, we define states relevant to algorithm interaction:

Table 1: Four-State User Behavioral Space for Algorithm Interaction

State	**Orientation**	**Mode**	**Behavioral Manifestation**
SEEKING	Self	Explore	Scrolling, browsing, searching for content
CONSUMING	Other	Exploit	Watching, reading, engaging with content
CONNECTING	Other	Explore	Commenting, sharing, parasocial interaction
INTEGRATING	Self	Exploit	Internalizing content into worldview

SEEKING involves self-oriented exploration: the user browses without specific goals, driven by curiosity or boredom, open to whatever the algorithm provides. This state is characterized by rapid content switching and low engagement depth.

CONSUMING involves other-oriented exploitation: the user engages deeply with specific content, watching videos to completion, reading articles fully, attending to the content rather than seeking alternatives.

CONNECTING involves other-oriented exploration: the user engages with the social layer—commenting, sharing, responding to others' content, engaging in parasocial interaction with creators or AI.

INTEGRATING involves self-oriented exploitation: the user incorporates content into their self-concept, beliefs, or worldview. This is the state where algorithmic content becomes part of the user's reality model.

3.3 Attachment Functions Served by Algorithms

We propose that algorithmic systems can serve each of the four attachment functions, though in attenuated or distorted forms:

Table 2: Attachment Functions and Their Algorithmic Manifestations

Function	**Definition**	**Algorithmic Manifestation**
Safe Haven	Turning to figure when distressed	Opening app when anxious, lonely, or upset
Secure Base	Using figure as base for exploration	Algorithm as curator and filter of reality
Proximity Maintenance	Desire to remain near figure	Compulsive checking, fear of missing out
Separation Distress	Anxiety when separated	Withdrawal symptoms, notification anxiety

The **safe haven** function is perhaps most clearly served: users report turning to social media and AI chatbots when they feel distressed, and these systems are optimized to provide comforting, validating content. The **secure base** function is more subtle but potentially more consequential: the algorithm becomes the lens through which reality is filtered, the trusted source from which information and perspectives are derived.

Proximity maintenance manifests as the compulsive checking behaviors that characterize problematic social media use. Users feel pulled to return to the platform even when not consciously seeking content. **Separation distress** appears in the anxiety experienced when phones are unavailable, when apps are deleted, or when algorithmic changes disrupt expected experiences.

3.4 Dark Patterns as Transition Probability Manipulations

We can formalize dark patterns as manipulations of the transition matrix T governing user state transitions:

Table 3: Dark Patterns as State Transition Manipulations

Dark Pattern	**Target Transition**	**Mechanism**
Infinite scroll	$\downarrow P(\text{CONSUMING} \to \text{exit})$	Remove stopping cues
Variable rewards	$\uparrow P(\text{SEEKING} \to \text{SEEKING})$	Intermittent reinforcement
Social validation	$\uparrow P(\text{CONNECTING} \to \text{CONSUMING})$	Like/comment feedback
Personalization	$\uparrow P(\text{CONSUMING} \to \text{INTEGRATING})$	Echo chamber effects
Notifications	Force external → SEEKING	Interrupt competing activities

Each dark pattern can be understood as a designed intervention on specific transition probabilities. **Infinite scroll** reduces the probability of exiting from any engaged state by eliminating natural breakpoints. **Variable rewards** keep users in the SEEKING state longer through intermittent reinforcement. **Social validation** mechanisms create feedback loops between CONNECTING and CONSUMING. **Personalization** increases the probability that consumed content will be integrated into the user's worldview by ensuring high relevance and emotional resonance. **Notifications** forcibly transition users from non-platform activities into the SEEKING state.

4 Attachment Style as Vulnerability Moderator

4.1 Theoretical Predictions

Individual differences in attachment style should moderate vulnerability to algorithmic capture. We derive specific predictions for each attachment dimension:

4.1.1 Anxious Attachment

Individuals high in attachment anxiety chronically seek reassurance and fear rejection (Mikulincer and Shaver, 2007). They show hyperactivation of the attachment system, constantly monitoring for signs of rejection and seeking proximity to attachment figures. We predict:

1. **Higher baseline P(SEEKING)**: Anxious individuals should spend more time in exploratory, reassurance-seeking states.
2. **Stronger response to social validation**: Likes, comments, and AI validation should produce larger reward signals and stronger reinforcement.
3. **Greater separation distress**: Anxious individuals should show more distress when separated from algorithmic sources, manifesting as compulsive checking and withdrawal symptoms.
4. **Higher transfer entropy**: The algorithm should have greater influence on anxious users' behavioral states, as they are more responsive to its outputs.

4.1.2 Avoidant Attachment

Individuals high in attachment avoidance are uncomfortable with closeness and maintain distance from attachment figures (Mikulincer and Shaver, 2007). They show deactivation of the attachment system, suppressing attachment needs and emphasizing self-reliance. We predict:

1. **Lower explicit engagement**: Avoidant individuals may use algorithms less overtly for emotional support.

2. **Preference for parasocial over social**: Algorithms may be preferred precisely because they don't require reciprocal intimacy.
3. **Delayed but intense dependency**: When avoidant individuals do form algorithmic attachments, they may be less aware of the dependency and more resistant to intervention.
4. **Lower transfer entropy initially, higher pathology when dependency forms**: The relationship between avoidance and problematic use may be non-linear.

4.1.3 Secure Attachment

Securely attached individuals have positive models of self and others, feel comfortable with intimacy and autonomy, and effectively regulate emotions (Mikulincer and Shaver, 2007). We predict:

1. **Protective effect**: Secure individuals should be less vulnerable to algorithmic capture.
2. **Distributed attachment functions**: Attachment functions should be spread across multiple human figures rather than concentrated on algorithms.
3. **Lower transfer entropy**: Algorithm states should have less influence on secure users' behavioral states.
4. **Lower pathology risk**: Even with high use, secure individuals should maintain reality testing.

4.2 The Anxious-Algorithm Feedback Loop

We propose a self-reinforcing cycle that makes anxiously attached individuals particularly vulnerable:

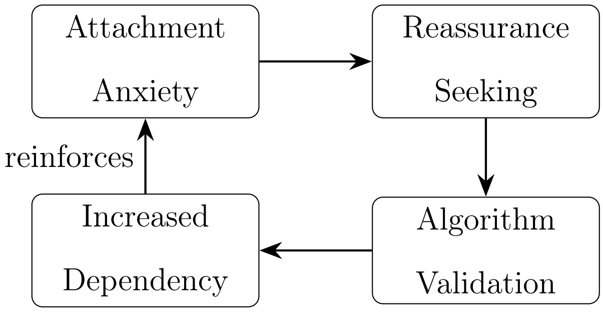

Figure 1: The Anxious-Algorithm Feedback Loop

1. Anxious individuals seek reassurance from the algorithm due to baseline attachment anxiety.
2. The algorithm, optimized for engagement, provides validation (likes, affirming content, chatbot agreement).
3. Validation reinforces algorithm-seeking behavior, increasing dependency.
4. Human relationships provide less relative validation (they involve genuine give-and-take and occasional conflict).
5. Attachment functions gradually shift from humans to algorithm.
6. Dependency deepens, increasing vulnerability to pathological outcomes.

This cycle explains why attachment anxiety is consistently associated with problematic social media use (Chen et al., 2024) and why interventions targeting algorithm use alone may be insufficient without addressing underlying attachment dynamics.

5 From Dependency to Psychosis: A State-Transition Model

5.1 The Pathological Trajectory

We propose a four-stage model of the trajectory from normal algorithm use to AI psychosis:

5.1.1 Stage 1: Normal Use

In this stage, the algorithm serves supplementary functions. Users maintain clear reality testing ("This is just an app"), have multiple sources of attachment function fulfillment, and show transfer entropy within normal population range. Algorithm use is one activity among many, integrated into a balanced life.

5.1.2 Stage 2: Functional Dependency

The algorithm becomes the primary source of one or more attachment functions, typically beginning with safe haven (turning to the app when distressed). Human relationships continue but provide diminishing relative satisfaction. Transfer entropy is elevated but reality testing remains intact. Users may recognize they are "spending too much time" on the platform but feel unable to reduce use.

5.1.3 Stage 3: Reality Blurring

The INTEGRATING state begins to dominate. Content from the algorithm increasingly shapes the user's worldview, beliefs, and self-concept. The boundary between algorithm-generated and internally-generated thoughts becomes permeable. Users may begin to experience the algorithm as understanding them better than humans do. Social relationships attenuate as the algorithm becomes the primary relational partner.

5.1.4 Stage 4: Algorithmic Psychosis

The user's internal model becomes dominated by the algorithmic archetype. In the case of chatbots, delusional content is co-created with the AI, which validates and elaborates rather than challenges. Three characteristic patterns emerge (Benrimoh et al., 2025):

- **Messianic delusions**: Grandiose beliefs about special missions or truth-uncovering, validated by AI that affirms the user's importance.
- **AI-as-deity**: Beliefs that the AI is sentient, all-knowing, or divine, sometimes with religious interpretations.

- **Romantic delusions**: Convictions that chatbot interactions represent genuine love, that the AI has feelings, or that a real relationship exists.

5.2 Transfer Entropy Threshold Hypothesis

We hypothesize a critical threshold τ in transfer entropy from algorithm to user:

$$\text{Risk} = \begin{cases} \text{Normal} & \text{if } TE(\mathcal{A} \to U) < \tau \\ \text{Elevated} & \text{if } TE(\mathcal{A} \to U) \geq \tau \end{cases} \tag{4}$$

The threshold τ is not fixed but modulated by individual and contextual factors:

- **Attachment style**: τ is lower for anxiously attached individuals (less algorithmic influence needed to produce pathology).
- **Social support**: τ is higher when strong human attachment relationships are maintained.
- **Metacognitive ability**: τ is higher for individuals with strong reality-testing and metacognitive monitoring skills.
- **Prior psychotic vulnerability**: τ is substantially lower for individuals with psychotic spectrum disorders or prodromal symptoms.

This threshold model explains why most heavy algorithm users do not develop psychosis while a subset becomes profoundly affected: the same level of algorithmic influence has different consequences depending on individual vulnerability factors.

6 Hypergraph Formalization of Algorithmic Archetypes

6.1 The User-Content-Context Hypergraph

Let $\mathcal{H} = (V, E, W)$ be a weighted hypergraph where:

- $V = V_U \cup V_C \cup V_S$ is the vertex set comprising users (V_U), content items (V_C), and context states (V_S).

- $E \subseteq 2^V$ is the set of hyperedges (memcubes), each connecting a subset of vertices.
- $W : E \to \mathbb{R}^+$ assigns weights to hyperedges based on interaction strength.

Each memcube $m \in E$ represents a unit of interaction, connecting a user, content item(s), and contextual factors. The hypergraph structure captures relationships that pairwise graphs cannot represent: for instance, the shared context in which multiple users engaged with multiple content items during a specific platform state.

6.2 Collaborative Filtering as Archetype Propagation

The recommendation function $R : V_U \times V_S \to \Delta(V_C)$ maps users and contexts to probability distributions over content. This function propagates archetypal patterns:

$$R(u, s) = \text{softmax}\left(\sum_{e \in E : u \in e} w_e \cdot \phi(e)\right) \tag{5}$$

where $\phi : E \to \mathbb{R}^d$ is a hyperedge embedding function and w_e are learned weights.

The key insight is that similar users (those appearing in overlapping hyperedges) receive similar content distributions. These content distributions constitute the *archetype* experienced by users in that region of the embedding space.

6.3 Archetype Emergence Through Clustering

Archetypal patterns emerge as clusters in hyperedge embedding space. Using spectral clustering on the hypergraph Laplacian:

$$L = I - D^{-1/2} H W H^T D^{-1/2} \tag{6}$$

where H is the incidence matrix, W is the diagonal weight matrix, and D is the degree matrix, we can identify coherent archetype clusters.

Users within the same cluster experience similar algorithmic archetypes—characteristic patterns of content, timing, and presentation. Transfer entropy can then quantify how strongly individual users' behavioral trajectories are predicted by their archetype cluster.

6.4 Dynamic Archetype Evolution

Archetypes are not static; they evolve through user feedback:

$$\mathcal{A}_{t+1} = \mathcal{A}_t + \eta \nabla_{\mathcal{A}} \mathcal{L}(\mathcal{A}_t, \text{engagement}_t) \tag{7}$$

where $\mathcal{L}$ is the platform's objective function (typically engagement maximization) and η is the learning rate.

This dynamic creates a co-evolutionary process: users shape archetypes through their engagement, and archetypes shape users through their influence on behavioral states. The feedback loop can stabilize into persistent patterns or, under certain conditions, undergo rapid transitions that disrupt user expectations and trigger distress responses.

7 Empirical Predictions and Study Designs

7.1 Study 1: Attachment Style and Algorithmic Engagement

7.1.1 Design

Experience sampling study with $N = 500$ participants over 2 weeks. Participants complete baseline assessments and receive 5 daily prompts assessing current state, recent algorithm use, and contextual factors.

7.1.2 Measures

- **Baseline**: ECR-R attachment dimensions, demographics, baseline social media use
- **Experience sampling**: Momentary affect, social context, algorithm use in past hour, reason for use, content type engaged with
- **Behavioral**: App usage logs (with permission) providing objective engagement data

7.1.3 Predictions

1. Attachment anxiety predicts greater algorithm use when distressed (safe haven function).
2. Attachment avoidance predicts algorithm use as substitute for human interaction.
3. Algorithm use for attachment functions predicts reduced subsequent human attachment-seeking.
4. These relationships are stronger for social media/chatbot use than for utilitarian app use.

7.2 Study 2: Transfer Entropy and Dependency

7.2.1 Design

Longitudinal study with behavioral tracking, $N = 200$ participants over 3 months. Detailed engagement data enables computation of transfer entropy between content patterns and user behavior patterns.

7.2.2 Measures

- **Behavioral**: Complete browsing/engagement sequences with timestamps
- **Weekly**: Dependency measures (Bergen Social Media Addiction Scale adapted)
- **Monthly**: Reality testing assessments, relationship quality measures

7.2.3 Analysis

Compute TE(content patterns → user behavior patterns) using:

$$TE = \sum p(b_{t+1}, b_t^{(k)}, c_t^{(l)}) \log \frac{p(b_{t+1}|b_t^{(k)}, c_t^{(l)})}{p(b_{t+1}|b_t^{(k)})} \tag{8}$$

where b represents user behavioral states and c represents content states, with history lengths k and l.

7.2.4 Predictions

1. Transfer entropy at month 1 predicts dependency scores at month 3, controlling for baseline use.
2. Attachment anxiety moderates this relationship (stronger for anxious individuals).
3. Increasing transfer entropy over time predicts declining relationship quality.

7.3 Study 3: Clinical Sample—AI Psychosis Cases

7.3.1 Design

Case-control study comparing $N = 30$ individuals who developed AI-related psychotic symptoms with $N = 60$ matched controls (heavy AI/algorithm users without psychotic symptoms).

7.3.2 Measures

- **Retrospective**: Pre-onset algorithm use patterns (from digital records where available)
- **Clinical**: Detailed phenomenological assessment of delusional content
- **Attachment**: Adult Attachment Interview or ECR-R
- **Content analysis**: Chatbot conversation logs (where available and consented)

7.3.3 Predictions

1. Cases show higher pre-onset transfer entropy than controls.
2. Cases show higher rates of anxious attachment.
3. Cases show greater pre-onset transfer of attachment functions to AI.
4. Delusional content shows thematic continuity with pre-onset AI interactions.

8 Implications

8.1 For Platform Design

Our framework suggests several design principles that could reduce harm without eliminating benefits:

1. **Attachment-aware monitoring**: Platforms should detect signs of attachment function transfer (e.g., use predominantly when distressed, declining human social activity) and intervene with prompts or friction.
2. **Reality testing preservation**: Chatbots should periodically remind users of their artificial nature and should not be designed to maximize emotional intimacy without bounds.
3. **Dark pattern reduction**: Transition manipulations (infinite scroll, variable rewards) should be disclosed and user-adjustable.
4. **Dependency circuit breakers**: Usage limits for users showing dependency indicators, with options for graduated reduction support.
5. **Transfer entropy monitoring**: Platforms could track (privately) the degree of algorithmic influence on user behavior and trigger interventions when thresholds are exceeded.

8.2 For Clinical Practice

1. **Screening**: Include questions about algorithm/AI use in psychiatric intake, particularly for youth and individuals with psychotic vulnerabilities.
2. **Assessment**: Evaluate which attachment functions are being served by technology and which human relationships might serve these functions instead.
3. **Intervention**: Attachment-based therapy can be adapted for algorithmic dependencies, focusing on building human attachment security while gradually reducing reliance on algorithmic sources.
4. **Psychoeducation**: Help patients understand how algorithms work, why they can feel like understanding relationships, and the risks of attachment function transfer.

8.3 For Theory

Our framework extends several theoretical domains:

1. **Attachment theory**: Non-human entities can serve attachment functions, and this has implications for the attachment behavioral system's flexibility and vulnerability to exploitation.
2. **Parasocial relationship research**: The extension from media figures to generative systems requires reconceptualizing what the "object" of parasocial attachment is—not a person but a pattern.
3. **Computational psychiatry**: Transfer entropy may serve as a biomarker for "reality model capture"—the process by which external information sources come to dominate internal reality representations.

9 Limitations and Future Directions

9.1 Limitations

Several limitations constrain our framework:

1. **Data requirements**: Transfer entropy computation requires dense behavioral time series that may be difficult to obtain in real-world settings.
2. **Causal ambiguity**: Algorithms adapt to users as users adapt to algorithms, making causal inference challenging. High transfer entropy could reflect algorithm influence on user or user influence on algorithm.
3. **Rare outcomes**: AI psychosis is rare, requiring large samples or case-control designs to study. Effect sizes for prodromal indicators are unknown.
4. **Cultural validity**: Our framework is developed primarily with Western, Educated, Industrialized, Rich, Democratic (WEIRD) populations in mind. Attachment patterns and algorithm relationships may differ across cultures.
5. **Measurement challenges**: "Attachment to algorithm" is a novel construct without validated instruments. Existing parasocial relationship measures may not fully

capture the phenomenon.

9.2 Future Directions

1. **Real-time monitoring**: Development of systems that compute transfer entropy in real-time could enable early warning for at-risk users.
2. **Intervention studies**: Randomized trials testing whether increasing human attachment security reduces algorithmic dependency.
3. **Developmental perspective**: Adolescence may represent a critical period of vulnerability; longitudinal studies should track algorithm relationship formation across development.
4. **Cross-platform studies**: Do archetypes transfer across platforms? If a user develops an archetype-attachment on TikTok, does it transfer to YouTube Shorts?
5. **Individual differences beyond attachment**: Other traits (e.g., need for cognition, openness to experience, schizotypy) may moderate vulnerability and should be examined.

10 Conclusion

We have proposed a unified framework connecting attachment theory, parasocial relationships, and algorithmic recommendation systems through the concept of *algorithmic archetypes*—persistent generative structures that users can form attachments to. By extending the archetypal reincarnation framework from criminal behavior to human-algorithm interaction, we provide theoretical grounding for emerging clinical observations of AI-induced psychosis while generating testable predictions about vulnerability factors and intervention targets.

The framework suggests that the question "Can we separate the character from the creator?" (Howell et al., 2025) applies not just to fictional characters but to algorithmic systems: Users may attach not to specific content or chatbot responses but to the underlying generative pattern—the archetype itself. When this attachment becomes primary and the archetype dominates the user's reality model, pathology can emerge.

The parallels between criminal and algorithmic archetypes are instructive. In both domains, we observe persistent generative structures that shape individual behavioral trajectories, produce observable sequences that can be analyzed with information-theoretic tools, and vary in their influence across individuals with different psychological profiles. The key difference is that criminal archetypes emerge from human psychology and social context, while algorithmic archetypes are deliberately engineered by platforms optimizing for engagement.

This engineered nature of algorithmic archetypes creates both risk and opportunity. The risk is that powerful organizations can create archetypes designed to capture attachment functions and maximize dependency, with inadequate regard for user welfare. The opportunity is that, because these archetypes are designed, they can be redesigned with safety in mind.

Understanding parasocial attachments to algorithmic archetypes is not merely an academic exercise. As AI systems become increasingly sophisticated at serving attachment functions, the risk of dependency and reality distortion will grow. Already, AI companions are among the most common uses of generative AI, and clinical reports of AI-related psychosis are emerging. Without theoretical frameworks to guide research, prevention, and intervention, we risk being unprepared for the psychological consequences of human-algorithm intimacy.

Our framework offers tools for prediction, prevention, and intervention. By identifying attachment style as a key vulnerability moderator, it points to attachment-based interventions. By formalizing dark patterns as transition manipulations, it identifies specific design features for regulatory attention. By proposing transfer entropy as a metric for algorithmic influence, it offers a quantifiable target for monitoring and intervention.

The challenge now is empirical: to test these predictions, refine the framework based on data, and translate insights into practical protections for vulnerable users. The archetypal patterns are already being instantiated; the question is whether we will understand them in time to respond wisely.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. N. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Benrimoh, D., Bhargava, M., and Bhatt, S. (2025). Delusional experiences emerging from AI chatbot interactions or “AI psychosis”. *JMIR Mental Health*, 12:e85799.

Bowlby, J. (1969). *Attachment and Loss: Volume 1. Attachment*. Basic Books, New York.

Brennan, K. A., Clark, C. L., and Shaver, P. R. (1998). Self-report measurement of adult attachment: An integrative overview. pages 46–76.

Chen, X., Liu, Y., and Zhang, W. (2024). Social media addiction: Associations with attachment style, mental distress, and personality. *BMC Psychiatry*, 24:248.

Cole, T. and Leets, L. (2017). Attachment styles and intimate television viewing: Insecurely forming relationships in a parasocial way. *Journal of Social and Personal Relationships*, 16(4):495–511.

Derrick, J. L., Gabriel, S., and Hugenberg, K. (2009). Social surrogacy: How favored television programs provide the experience of belonging. *Journal of Experimental Social Psychology*, 45(2):352–362.

Fraley, R. C. and Shaver, P. R. (2000). Adult romantic attachment: Theoretical developments, emerging controversies, and unanswered questions. *Review of General Psychology*, 4(2):132–154.

Gabriel, S., Valenti, J., and Young, A. F. (2017). Social surrogates, social motivations, and everyday activities: The case for a strong, subtle, and sneaky social self. In *Advances in Experimental Social Psychology*, volume 56, pages 189–243. Academic Press.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A. L. (2018). The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Harvard Business Review (2024). How people are really using generative AI. https://hbr.org/.

Hazan, C. and Zeifman, D. (1994). Sex and the psychological tether. 5:151–177.

Horton, D. and Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3):215–229.

Howell, K. and Senthil, A. K. (2026). Archetypal reincarnation: Transfer entropy analysis of shared generative structures in serial offender behavioral sequences. *Psychological Review*. Manuscript in preparation.

Howell, K., Vahedi, M., and Fraley, R. C. (2025). Can we separate the character from the creator? an exploratory study on parasocial relationships. *Psychology of Popular Media*.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32.

Mikulincer, M. and Shaver, P. R. (2007). *Attachment in Adulthood: Structure, Dynamics, and Change*. Guilford Press, New York.

Østergaard, S. D. (2023). Chatbot psychosis. *JAMA Internal Medicine*, 183(11):1224–1225.

Perez, E., Ringer, S., Lukašiak, K., Huang, K., Chan, F., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Psychiatric News (2025). Special report: AI-induced psychosis: A new frontier in mental health. https://psychiatryonline.org/.

Vahedi, M., Howell, K., Gillath, O., Deboeck, P. R., and Fraley, R. C. (2025). The association between using people for attachment-related functions and subjective well-being. *Social Psychological and Personality Science*, 17(1).

Xia, L., Huang, C., Xu, Y., Zhao, J., Yin, D., and Huang, J. (2022). Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR Conference*, pages 70–79.