The Open-Audience Failure of Viral Therapy-Speak

Detection-Rule Leakage, Trust Camouflage, and MIPU/MAP Formation in Digital Pseudo-Psychology

Michael Zot
Independent Researcher, Cognitive Security and Symbolic Intelligence
Founder, ZotBot.ai · ORCID: 0009-0001-9194-938X

OPENING THESIS

Abuse-awareness content is made for people who need help. At its best, it gives somebody a word for a private hell they could not explain. A person hears “gaslighting” or “love bombing” and suddenly the fog has a shape. That part is real, and this paper does not try to take it away.

The problem begins when that lesson leaves the helping room and enters the feed. A feed is not a clinic. It does not check who is watching, what role they play, what history they bring, or what they plan to do with the lesson. The victim sees it. The person doing harm sees it. The anxious viewer sees it. The person looking for a stronger word for a normal fight sees it. The creator chasing moral authority sees it. Nobody gets screened at the door. The lesson just goes out.

Accuracy alone cannot carry the whole burden. A correct rule can still become unstable when millions of strangers receive it without context. One viewer finds safety. Another finds a weapon. Another finds a script that lets them stop listening.

The paper tracks three risks inside that open room. **Detection-rule leakage** means a public warning can teach victims what to notice while also teaching some harmful people what to hide. **Trust camouflage** means a person can sound protective in public before anyone has checked how they behave in private. **MIPU/MAP** means a small label can change what someone notices, expects, and treats as proof long after the clip is gone.

OAPRM gives the argument a testable form. It names ten dimensions to score, eight claims that can lose, and six study designs other researchers could run. The supporting pieces already exist across concept creep, context collapse, memory science, DARVO, diagnostic labeling, misinformation persistence, algorithmic amplification, adversarial adaptation, affect labeling, estrangement, information hazards, and responsible disclosure. No single field proves the whole chain. Together they make the chain worth testing.

The standard is simple. Judge abuse education by every audience it trains: the victim, the abuser, the false accuser, the anxious watcher, the brand-building creator, and the person in the back row learning what to hide next.

A NOTE ON METHOD

The opening is written to be felt, but the argument does not get to win because it sounds right. Section 12 gives eight ways the claims can fail. Section 13 sets the evidence bar. Section 14 gives studies that other people could run. If the data go against the framework, the framework has to change or come down.

ABSTRACT

Therapy-speak has become one of the largest informal psychology systems online. Words like gaslighting, narcissist, trauma bond, love bombing, DARVO, boundaries, toxic, and no contact now move through TikTok, YouTube, Instagram, X, Reddit, podcasts, and group chats with almost no gatekeeping. Most criticism focuses on bad advice, loose credentials, self-diagnosis, and clinical words drifting from their original meanings. Those concerns matter, but they miss the larger structure: these lessons reach an open audience.

An open audience holds real victims, people in painful but non-abusive conflict, anxious viewers looking for certainty, people sharpening false accusations, creators chasing status, and the very people the lessons claim to expose. The algorithm does not know which row anyone sits in. A single detection rule reaches all of them at once.

This paper maps that loop. Clinical language leaks into daily talk. Platforms reward the most emotionally certain version. Creators compress hard cases into quick signs. Viewers carry those signs into relationships, comment sections, breakups, family fights, and moral performances. Some people finally name real harm. Some people mislabel ordinary conflict. Some learn how to sound safer while behaving the same.

The model rests on three ideas. Detection-rule leakage describes warnings that protect one viewer while giving another viewer evasion notes. Trust camouflage describes public fluency in abuse language that creates unearned moral trust. MIPU/MAP describes how a small input can change what a viewer notices, expects, and questions next, then harden into a broader interpretive lens when several such updates support one another.

OAPRM, the Open-Audience Pseudo-Psychology Risk Model, turns the concern into a coding instrument. It scores ten dimensions, gives eight falsifiable hypotheses, and proposes six study designs. The goal is not censorship. The goal is measurement, friction, and better education for the audience that actually exists, not only the audience creators imagine.

KEYWORDS: therapy-speak · cognitive security · detection-rule leakage · trust camouflage · concept creep · context collapse · MIPU · MAP · gaslighting · DARVO · no contact · adversarial learning · mental-health misinformation · diagnostic labeling · algorithmic amplification

1 Introduction: The Other Student in the Classroom

Picture a classroom. A security expert stands at the front teaching people how to spot pickpockets. In the front row sit people who have actually been robbed, and they need this lesson. In the middle rows are regular people who will walk out a little sharper, though some of them may start seeing thieves everywhere they look.

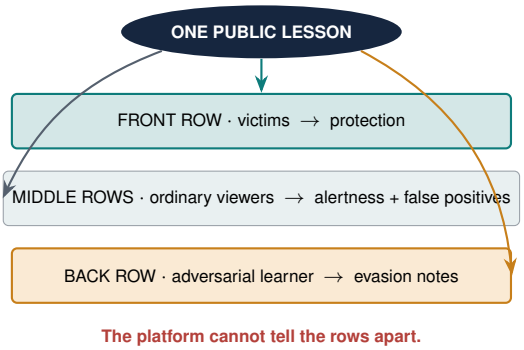

Figure 1: The open-audience classroom. A single detection lesson is received simultaneously by protective, neutral, and adversarial learners, with no mechanism to separate them.

In the back row sits an actual pickpocket, taking notes like it is professional development. One lesson, three completely different takeaways. The teacher has no way of knowing who is learning what. Neither does the algorithm pushing the video (Figure 1).

Short-form content about abuse is that same classroom, scaled to billions of views. There is no intake form, no case file, and no way to know who is sitting where. People usually argue about whether the advice is accurate or whether the creator has credentials. Those things matter. But even a perfect video from a trained clinician still lands in the same mixed room. The question almost nobody asks is the one that changes everything: who else is watching?

Once you ask that question, the subject shifts. The content does not reach a clean group of victims. It reaches people who genuinely need the language. It reaches people trapped in messy fights that are not abuse. It reaches anxious viewers who want certainty so badly that almost any label feels like relief. It reaches people looking for sharper tools to accuse with. It reaches creators who have learned that moral authority travels well online. And it reaches the exact people the content warns everyone about. That is not a side effect. That is the structure.

A detection rule that protects one viewer can train another. A warning can teach someone how to hide better. A term that gives one person clarity can become ammunition for someone else. The person sounding protective on camera might be building a following while running the same patterns off camera.

This is why I treat viral therapy-speak as a cognitive-security problem. The question is how people can protect the way they form beliefs, set evidence standards, update their read on others, and interpret social signals when the input is simplified, emotionally charged, and algorithmically amplified. I am not saying therapy-speak is always bad. Abuse and coercive control are real, and sometimes no contact is the only safe option. The narrower point is that psychological rules stop being neutral information when platforms strip them down, crank them up emotionally, and blast them into an audience whose motives stay invisible.

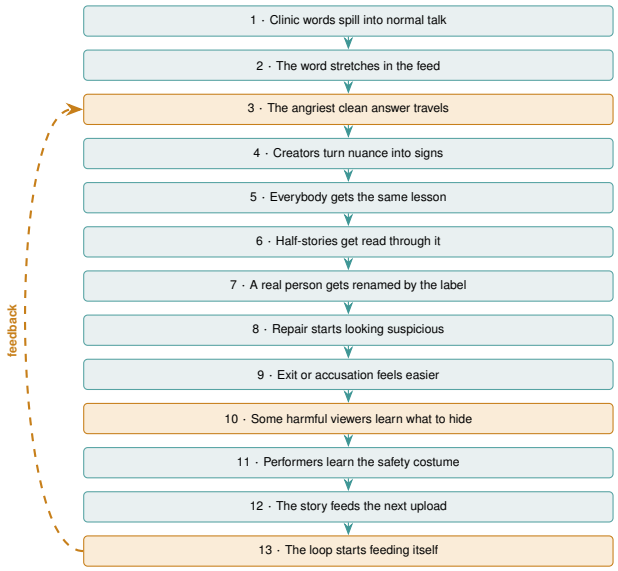

Figure 2: One way the loop can run. The dashed line shows how strong stories feed more strong stories.

The usual misinformation frame does not cover that. A false claim is one kind of failure. Sometimes the dangerous thing is a rule that is mostly true, stripped of context and case history, landing in the wrong hands and quietly changing what someone now treats as proof.

> *Viral therapy-speak can change what counts as proof before people notice that their standard of proof has moved.*

2 The Mechanism Exposed

Nobody had to design this in a back room. The loop hides in plain sight because different fields have been studying different pieces of it under different names. One field studies concept creep. Another studies context collapse. Another studies memory, diagnostic labels, outrage, and misinformation persistence. When those pieces are placed in sequence, the mechanism becomes easier to see (Figure 2).

The chain is not a claim that every viewer walks through every step. It is a claim about a pathway that platforms make available at scale. Clinical words leave bounded settings, become short moral signs, reach unscreened viewers, and then begin affecting how those viewers interpret people they already know. OAPRM exists so that this can be measured instead of merely complained about.

The main unit of analysis is the audience, not the word. A hammer is a tool in one hand and a weapon in another. A psychological term can name harm, excuse avoidance, help someone leave danger, help someone accuse first, or let someone borrow moral trust they have not earned. The word alone does not decide the outcome. The viewer, the motive, the relationship history, and the surrounding platform incentives decide what the word becomes.

3 Existing Evidence: The Proof Ledger

This argument is not built from frustration. It connects pieces already sitting across several research areas. No single study proves that viral therapy-speak creates open-audience harm. That is exactly why synthesis matters. One field explains how harm words stretch. Another explains how feeds collapse audiences. Another explains how labels, memory, outrage, and correction failures change later judgment. The assembled chain remains the test object, but the links are not imaginary.

3.1 Therapy-speak is already flagged

Isern-Mas and Almagro (2025) examine loose clinical language in daily life. Their risk list lines up with the problem studied here: eroded meanings, ordinary life turned pathological, self-diagnosis, discrediting others, dodging responsibility, status games, and borrowed medical authority. Almagro and Isern-Mas also show the double edge. Popular language can help people name pain, while the same popularity can dull the concepts. This paper adds the open-audience question: what happens when dulled language reaches victims, confused viewers, false accusers, performers, and adversaries at the same time?

3.2 Harm words expand

Haslam's (2016) work on concept creep gives the semantic backbone. Words such as trauma, abuse, bullying, and prejudice can stretch into milder territory over time. Once a word stretches far enough, ordinary conflict can start to look pathological. The next question is practical: who learns the stretched version, and what do they do with it?

3.3 Social media smashes audiences together

Marwick and boyd (2011) describe context collapse: different audiences get mashed into the same feed, so the speaker cannot tailor the message. A recipe can survive that collapse because the stakes are low. A detection rule for hidden abuse carries higher stakes because it can affect accusation, trust, evasion, repair, and severance.

3.4 Moral language becomes status play

Tosi and Warmke (2016) show how moral talk can become self-promotion. Anti-abuse language fits that incentive perfectly because the speaker appears protective, virtuous, and difficult to criticize. The darker version is trust camouflage: a person sounds like a protector in public while using related tactics in private.

3.5 Memory is messy

Loftus and Palmer (1974) and Loftus (2005) show that memory is reconstructive. Two honest people can remember the same event differently. Content that treats

memory disagreement as automatic gaslighting sets the evidence bar lower than the science supports. That lower bar can decide whether a relationship moves toward repair or burns down.

3.6 DARVO is real, and checklists can still warp it

Harsey and Freyd (2020) show that DARVO affects credibility judgments, and that education can blunt some of those effects. The open-audience trap begins when DARVO becomes a shortcut. A person denying a false accusation can match the surface pattern. A manipulator can soften the obvious moves. A false accuser can invoke DARVO first so that defense itself starts looking like proof. The same concept can protect or distort depending on how it is taught and used.

3.7 Short-form mental-health content already misleads at scale

Research on ADHD TikTok, CBT TikTok, functional tic-like behavior, and mental-health misinformation shows that social platforms can shift perception, confidence, and symptom language at scale (Yeung et al., 2022; Karasavva et al., 2025; Lorenzo-Luaces et al., 2023; Frey et al., 2022; Nguyen et al., 2024). The same mechanism becomes more socially dangerous when the target moves from self-diagnosis to other-diagnosis.

3.8 Diagnostic labels reshape perception

Altmann et al. (2024) show that diagnostic labels can change how people read marginal mental-health cases, including treatment need, agency, empathy, and expected recovery. Viral quasi-labels such as narcissist and gaslighter can carry similar weight with far fewer guardrails.

3.9 Bad updates can survive correction

Ecker et al. (2022) and Walter and Tukachinsky (2020) show that misinformation can keep affecting reasoning after correction. This gives MIPU/MAP part of its empirical spine. The claim is not that every update is permanent. The claim is that some updates are easier to install than to remove, especially when they arrive with emotion, identity, and social confirmation attached.

3.10 Naming can also calm people down

The protective branch matters. Lieberman et al. (2007) found that putting feelings into words can reduce emotional reactivity. That work concerns labeling one's own affect, so it cannot be treated as proof that labeling other people is safe. Still, it shows why crude anti-label arguments are wrong. Naming can regulate. OAPRM targets the failure case, where a label stops organizing evidence and starts replacing it.

3.11 Outrage travels

Brady et al. (2017) show that moral-emotional language spreads faster. McLoughlin et al. (2024) show that misinformation can ride outrage. Milli et al. (2023) show that engagement ranking can amplify divisive content even when users say they prefer otherwise. Those incentives reward the cleanest, hottest version of a psychological claim, not the most careful one.

Table 1: The proof ledger. Individual links have evidence; the assembled chain remains the test object.

Link	Status
Term expansion	Evidenced
Audience collapse	Evidenced
Mental-health misinformation	Evidenced
Diagnostic-label effects	Evidenced
Memory malleability	Evidenced
DARVO credibility effects	Evidenced
Correction resistance	Evidenced
Affect labeling benefit	Evidenced
Emotional amplification	Evidenced
Adversarial adaptation	Adjacent
Severance complexity	Evidenced
Assembled open-audience chain	Untested here

3.12 Adversaries adapt when rules become visible

The adjacent evidence comes from adversarial machine learning, moderation evasion, and intimate-partner surveillance (Vassilev et al., 2025; Bickham et al., 2024; Tseng et al., 2020). These literatures show that motivated actors can adapt when boundaries become visible. This does not prove every abusive person becomes strategic. It supports a narrower risk: some motivated viewers can learn from the same public signs meant to protect others.

3.13 True information can still become risky

Information-hazard work and vulnerability-disclosure debates explain why true information can create risk when released without the right structure (Bostrom, 2011; Wardle and Derakhshan, 2017; Rescorla, 2005; Schneier, 2007). OAPRM is the therapy-speak version of that old fight. The issue is not whether the information is true in isolation. The issue is how a mixed audience can use it once it travels.

3.14 Estrangement is real, complex, and easily flattened

Pillemer (2020) and Coleman (2021) show that estrangement is messy and varied. Some distance is necessary. Some no-contact decisions protect people from danger. The risk begins when platforms turn no contact into a default script detached from severity, evidence, repair, and safety planning.

4 Construct Status

MIPU, MAP, and OAPRM are proposed constructs, not settled science. The paper gives the definitions, hypotheses, coding rules, evidence standards, self-audit, and study designs needed to judge them. The argument does not ask readers to trust an earlier program. It stands or falls on what is written here.

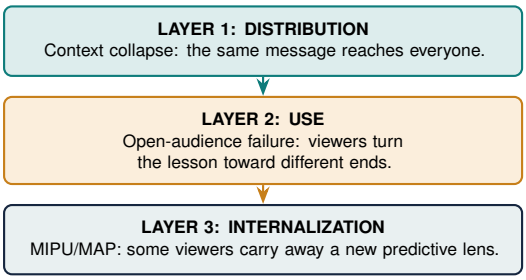

Figure 3: The three-layer stack. Distribution is not the same as internalization.

The word "irreversible" stays inside MIPU for continuity, but the technical meaning is modest. It means asymmetric reversal cost: the update can form quickly, while undoing it usually requires more evidence, trust, time, or lived contradiction than forming it required.

5 What MIPU/MAP Adds

Existing theories already explain pieces of belief updating, mental structure, and failed correction. MIPU/MAP targets a narrower event: a small input collides with a viewer's expectations, changes what the viewer notices next, and leaves behind a reversal cost that is larger than the cost of adoption.

Bayesian updating tracks probability shifts. MIPU/MAP asks when a label stops being a probability adjustment and starts changing what counts as relevant evidence. Moving from 0.3 to 0.7 is one kind of update. Reframing apology as hoovering or disagreement as DARVO changes the evidence field itself.

Schema theory describes durable mental structures. MIPU names the installation moment that makes one start forming. MAP names the larger lens that appears when multiple MIPUs begin supporting one another. The emphasis is not just that people have schemas. The emphasis is that some prediction lenses are cheap to install and expensive to dismantle once recognition has shifted.

Predictive processing helps explain why the moment can feel clarifying. A sudden label reduces uncertainty, and that reduction can feel like truth. MIPU/MAP adds the audit this domain needs: did clarity improve prediction across situations, or did it simply narrow what the viewer now treats as signal?

Concept creep explains how harm terms stretch. MIPU/MAP follows the stretched term into a viewer's future perception. A broad term is one problem. A broad term that becomes a quick social prediction engine is another.

The three-layer stack keeps the distinction clean. Context collapse explains distribution. Open-audience dynamics explain why different viewers use the same lesson differently. MIPU/MAP tracks what gets carried away inside the viewer and what later becomes hard to revise (Figure 3).

6 Core Vocabulary

The terms below are working definitions for measurement. They are meant to keep the argument scoreable rather than ornamental.

· Therapy-speak

Clinical-sounding language used to interpret social life: gaslighting, narcissist, trauma, toxic, triggered, attachment style, boundaries, love bombing, trauma bond, hoovering, DARVO, emotional abuse, and no contact. It can help people name real harm. Risk rises when the terms become broad, low-evidence labels for anything painful or uncomfortable.

· Viral pseudo-psychology

Public psychology content that borrows clinical authority while dropping clinical restraints such as case evidence, scope limits, counterexamples, professional responsibility, and downstream accountability. The issue is content structure, not credentials alone. A licensed therapist can make risky viral content, and an anonymous poster can make careful content.

· Open audience

A viewer pool with unknown motives and roles: victims, people in ordinary fights, anxious viewers, false accusers, harmful actors, status-seeking creators, friends hearing half a story, and bystanders who reuse the language elsewhere.

· Detection rule

A rule that tells viewers which signs reveal a hidden pattern. Examples include treating denial as DARVO, fast affection as love bombing, memory disagreement as gaslighting, or family concern as proof that no contact is required.

· Detection-rule leakage

The risk that a public lesson meant to help people spot manipulation also reveals the detection boundary to people who want to evade it. The same rule trains every viewer, including some viewers it was built to catch.

· Trust camouflage

Credibility gained by publicly opposing harmful behavior while privately using related tactics. The manipulator posts about manipulation. The reversal tactician teaches DARVO. The creator builds a following around victim protection while treating criticism as abuse. The mistake is simple: people

hear anti-abuse fluency and confuse it with evidence of safety.

· False positive

A heavy label applied to a person or behavior that fuller evidence would not support. This matters because labels such as abuser, narcissist, and gaslighter are sticky. Once attached, the accused person can lose the ability to explain, repair, or be read fairly.

· MIPU

A Minimal Irreversible Predictive Update is the smallest input that changes what a viewer notices, expects, or questions next at asymmetric reversal cost. "Irreversible" does not mean impossible to reverse. It means the update can install quickly while undoing it usually takes repeated counterevidence, trusted correction, emotional work, or lived contradiction.

· MAP

A Maximal Anticipatory Pattern is the larger lens that forms when multiple MIPUs stabilize together. Once formed, it guides attention, memory, fear, evidence standards, trust, repair behavior, and action across situations. It behaves less like one belief and more like a filter.

· Low-quality irreversibility

A durable prediction update produced by weak, context-poor, emotionally charged content. The paradigm case is a five-second clip that recodes a ten-year relationship and outlives the evidence that created it.

7 The Failure, Anatomized

The frustrating part is that many people inside this chain are acting in good faith. A creator wants to help. A viewer wants clarity. A term contains real truth. Harm can still happen because the structure is built for reach, not for judgment.

First, the creator is flying blind. There is no relationship history, danger assessment, or reliable way to know whether the viewer is a victim, a perpetrator, a confused partner, or someone being fairly called out. Second, the viewer can attach a heavy label to an absent person after seeing only a partial match. Third, the algorithm rewards certainty, so "toxic" travels farther than careful talk about pattern, power, duration, motive, and context. Fourth, the back row is always occupied. Some people learn to hide, some learn to accuse, and some learn to dress themselves in moral language.

The final step is the most dangerous. Once the label sticks, repair signals can become further evidence. An apology becomes hoovering. A question becomes gaslighting. Confusion becomes DARVO. The verdict starts writing the trial.

8 Three Harms Running Simultaneously

8.1 Mislabeled innocent

Many signs taught in viral content appear in both abuse and ordinary life: defensiveness, memory gaps, fast affection, conflict avoidance, apologizing, or asking to talk. Denial can be lying, but it can also be truth. Fast intensity can be love bombing, but it can also be excitement, anxiety, or immaturity. The false-positive path is short: a viewer sees a partial match, treats it as proof, recodes the person, and then filters future behavior through the label. From inside the experience, it feels like sudden clarity rather than a chain of small interpretive moves.

8.2 Educated abuser

Manipulative people have phones. That does not mean every harmful person becomes a master strategist. Many act from entitlement, anger, habit, or control, and many are bad at hiding it under pressure. The leakage claim is narrower: public checklists can teach some motivated viewers which obvious signs to sand down. Old red flags may fade while the control stays. OAPRM therefore treats adversarial adaptation as a risk to test, not a settled fact about all abusers.

8.3 Protector who is not

Trust camouflage scales cleanly online. Public anti-abuse language can buy instant credibility. Criticism can be reframed as harm. Followers can defend the halo. Once a person occupies the protector slot, harm from that person becomes harder to name.

> *Anti-abuse fluency can be weak evidence in some contexts, but it is never proof of safety. The downside of misreading it is too high.*

9 One-Way Door: Why It Does Not Wash Out

Before the video, the viewer may only think, "this relationship is confusing." After the video, the thought can harden into, "this is gaslighting." If the label is right, that moment can save someone. If the label is wrong, the same machinery can still run. Memories reorganize, future behavior gets filtered, and the feeling of awakening cannot tell the viewer which case they are in.

The risk grows when terms cluster. Love bombing explains the honeymoon. Gaslighting explains disagreement. DARVO explains defense. Trauma bond explains attachment. Narcissism explains personality. No contact

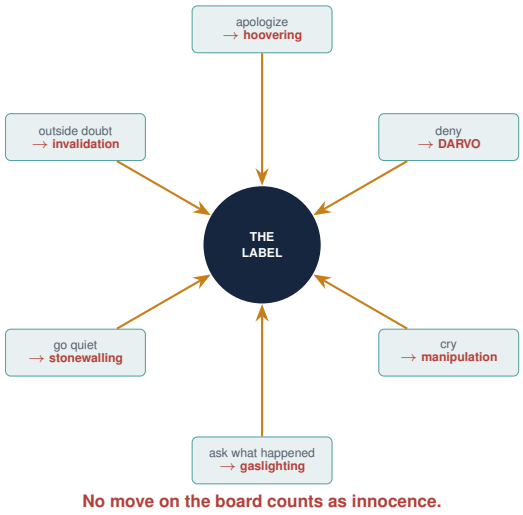

Figure 4: Closed interpretive map. Every response is absorbed as confirmation.

supplies the exit. Together they create a complete story with a villain, a mechanism, a proof test, and an action command.

Inside a closed map, almost every response gets absorbed. If the accused apologizes, that becomes hoovering. If they deny, that becomes DARVO. If they cry, that becomes manipulation. If they ask what happened, that becomes gaslighting. If they go quiet, that becomes stonewalling. Figure 4 shows the trap.

This is why "just vocabulary" is too weak as a defense. Some vocabulary gives people freedom. Some vocabulary becomes a one-way door. The difference has to be measured.

10 Objection That Should Have Killed This Paper

The strongest counterargument deserves full force. Security engineering abandoned secrecy long ago. Kerckhoffs argued that a good system should remain secure even when the adversary knows the design. Victims need public education, and abusive people often know their tactics already. If that objection wins, leakage becomes a reason to publish more, not a reason to slow down.

The anti-secrecy point is right. Silence helps abusers. The better comparison is responsible disclosure. Security people do not hide every vulnerability forever, but they also do not dump every operational detail with no patch, timing, warning, or audience care. Therapy-speak needs the same maturity.

The password analogy still helps. A system can survive public knowledge of its design. It cannot survive someone posting the password on the wall. Weak viral therapy-speak often behaves like the second case. It gives fixed signs, magic phrases, and instant-proof rules. Better education teaches the pattern without pretending that one phrase unlocks the case. It publishes protection without handing out a brittle evasion map.

The alternative is resilient disclosure: teach pattern reasoning under uncertainty. Pattern, repetition, motive, power, benefit, refusal of repair, isolation, and evidence over time matter more than one phrase or one clip. Table 2 shows the difference.

> *The choice was never between educating victims and staying silent. The real choice is between education that survives an open audience and education that breaks once the back row learns it too.*

11 Instrument: OAPRM

OAPRM is a prototype coding instrument, not a validated scale yet. Coders score each dimension from 0 to 5. The instrument can be tested through inter-rater reliability, factor analysis, and outcome prediction. The ten dimensions are expected to cluster into semantic distortion, misuse risk, and irreversibility risk, but the data may refuse that structure. If the factors do not hold, the model has to change.

12 Eight Ways to Prove Wrong

Every major claim has to be a bet that can lose.

FALSIFIABLE HYPOTHESES

H1. Checklist therapy-speak lowers evidence bars compared with careful education.

H2. High false-positive content increases other-diagnosis in ambiguous cases.

H3. High-leakage content helps adversarial participants build cleaner evasion plans.

H4. Diagnostic labels plus severance commands produce more blocking and no-contact comments.

H5. Repair-suppression language reduces clarifying questions and safe repair attempts.

H6. Term clusters create stronger closed-loop interpretation than single terms.

H7. Risk concentrates in a small subset of posts rather than spreading evenly.

H8. Resilient disclosure keeps usefulness while reducing false labels and evasion potential.

If the data cut against these claims, the model must be revised or scrapped. A paper about closed maps cannot become one.

13 Claim Licensing

Different claims need different evidence. Descriptive claims need corpus coding. Mechanism claims need experiments, longitudinal data, or strong triangulation. MIPU/MAP claims face the highest bar because they require persistence, transfer, and resistance to counterevidence. Until those studies run, the framework is a plausible map with tests attached, not a settled result.

14 Studies

Six study designs can move the paper from framework to evidence.

Table 2: Weak versus resilient disclosure. Weak disclosure treats signs like passwords; resilient disclosure teaches pattern reasoning.

Weak disclosure	Resilient disclosure
"Five signs of gaslighting."	"Gaslighting is repeated pattern of reality manipulation. Memory disagreement proves nothing alone. Look for repetition, motive, power imbalance, refusal of correction, isolation, benefit from confusion, and what evidence would change mind."
"Move fast = love bombing."	"Fast affection sincere or manipulative. Risk rises with pressure, isolation, premature promises, guilt, control, then devaluation. Intensity alone = intensity."
"Deny = DARVO."	"DARVO reversal pattern. Innocent people deny too because some accusations false. Watch denial + attack + credibility reversal + evidence avoidance across time."

Table 3: OAPRM instrument. Dims 0–5. Shaded bands = predicted factors.

	Dimension	High-risk (5)	Low-risk (0–1)
S	**Semantic**	Drift / inflation / corruption	Rows 11.1–11.3
11.1	Semantic Drift	"Gaslighting = disagreeing."	"Repeated pattern + context + evidence."
11.2	Diagnostic Inflation	"Attention-seeker = narcissist."	"Many causes; no disorder established."
11.3	Evidence Corruption	"Saying X = proves Y."	"X in many contexts; pattern matters."
M	**Misuse**	Leakage / adaptation / misuse / camo	Rows 11.4–11.6, 11.8
11.4	Intent Failure	Every viewer = victim assumed.	Warnings against diagnosing from clips included.
11.5	Detection Leakage	"Exact phrases manipulators use."	Pattern reasoning; no fixed scripts.
11.6	Adversarial Adaptation	Visible markers only.	Long-term evidence-based assessment.
11.8	Camouflage Potential	Creator uniquely safe; crit = harm.	Education separate from moral status.
I	**Irreversibility**	FP / severance / suppression	Rows 11.7, 11.9, 11.10
11.7	FP Susceptibility	Signs common stress/grief/neuro/culture.	Counterexamples supplied.
11.9	Severance Activation	"Do X once = run."	Danger/bounds/repair/safety distinguished.
11.10	Repair Suppression	"Apology = hoovering."	Unsafe reconcil vs genuine repair separated.

Corpus study. Collect public content and code transcripts, engagement, definitions, counterexamples, warnings, calls to action, and comment samples.

OAPRM coding study. Train raters, test reliability, run factor analysis, and test whether scores predict comment outcomes.

Comment outcome study. Code other-diagnosis, severance intent, repair refusal, perceived awakening, hostility to dissent, and concrete behavior change.

Priority experiment. Randomly assign participants to careful education, viral checklist content, personal story content, or neutral conflict content. Use ambiguous vignettes to measure label confidence, evidence bars, repair willingness, alternate explanations, and resistance to counterevidence.

Adversarial simulation. Use ethics-reviewed abstract role-play to compare checklist material with resilient material. The goal is to measure evasion planning without producing real-world scripts.

Longitudinal MAP study. Track exposure, familiarity, other-diagnosis frequency, repair willingness, belief rigidity, and relationship outcomes over months.

15 No-Contact Cluster

No contact can save lives. Dangerous, coercive, obsessive, or destabilizing relationships sometimes require distance. Any framework that blurs that fact fails ethically.

The risk sits in the cluster around the phrase, not the phrase alone: no contact, narcissist, toxic family, gaslighting, protect your peace, anyone questioning you is invalidating, never explain, they know what they did, and if they cared they would understand. Read as a system, the cluster welds exits shut. Questions become invalidation. Explanations become weakness. Reconsideration becomes proof that the person has not healed. OAPRM dimensions 11.3, 11.9, and 11.10 are built to catch this pattern while still protecting genuine escape.

16 Mechanism Detection

One video proves little. Repeated vocabulary can be a meme without becoming a harm pattern. Harmful convergence needs three layers at once (Figure 5).

Layer 1 is linguistic: terms travel together, such as narcissist, gaslighting, trauma bond, boundaries, toxic, and no contact. Layer 2 is interpretive: evidence bars move, alternative explanations shrink, disagreement starts looking abusive, and repair starts looking suspicious. Layer 3 is behavioral: viewers other-diagnose, report severance, block, refuse repair, or suddenly reinterpret long relationships through the new label.

A benign control such as sourdough or fitness can show whether jargon alone travels as a pack. It cannot carry the whole burden because abuse language is threat-relevant and sourdough is not. The better test compares crude abuse checklists, careful abuse education, and other threat lessons such as scam or crime awareness. If careful education raises useful caution without closing the map, OAPRM gains support. If every threat lesson looks the same, the model has to shrink.

17 Action Levels

17.1 Viewer: immunization

Before a label hardens, the viewer can slow the update down. What else explains this? Does the pattern

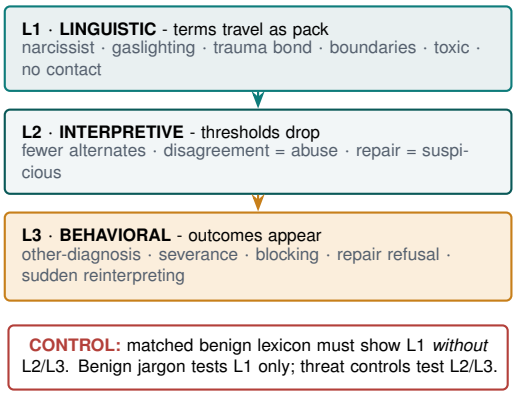

Figure 5: Three-layer convergence. Benign jargon tests term-packing; threat controls test harmful closure.

repeat over time? What evidence would change my mind? Am I confusing discomfort with harm, safety with punishment, or repair with unsafe reconciliation? Did the content give counterexamples, or only certainty? Could someone use the same lesson to hide better? Has one term become the whole lens?

17.2 Creator: resilient disclosure

People making abuse or mental-health content for open audiences should define terms, state limits, give counterexamples, name evidence bars, offer alternate explanations, warn against diagnosing from clips, warn against weaponizing the lesson, acknowledge that adversaries may be watching, separate repair from danger, and point serious situations toward professionals. The goal is not to make the content weaker. The goal is to make it survive the real room it enters.

17.3 Platform: measurement and friction

High-risk content can carry light context nudges: "Discusses abuse or health. Short clips cannot diagnose people. Serious situations need full context and professional help." Platforms could also score content along OAPRM dimensions such as drift, leakage, false-positive risk, repair suppression, trust camouflage, and comment clusters. The goal is measurement and light friction, not crude censorship.

17.4 Clinical and educational use

Therapists, mediators, and educators can ask clients where they learned a term, how they define it, what would falsify it, and whether the word clarified a mess or turned every interaction into confirmation. That last question is the compressed MAP-closure test.

18 What Changes

The old test is too small: did the content help the person it meant to help? A better test watches the whole room. One viewer needed the word. One viewer wanted a weapon. One viewer was confused and scared. One viewer learned how to sound safe. The same lesson can help, harm, and camouflage at once. That mixed result is the paper's target.

> *Core claim: viral therapy-speak can change evidence bars before viewers notice the change. Sometimes it raises the bar by giving a person structure. Sometimes it lowers the bar by letting a label replace evidence. OAPRM is built to tell those apart.*

This goes beyond simple falsehood. Partly true content can still install bad updates when it travels without context, limits, or audience-risk controls.

19 Not Earned Yet

This is a framework, a synthesis, and a research program. It is not finished validation. Strong MIPU/MAP claims stay at the evidence tier in Section 13 until studies run.

The confounds are real. Some people were in danger before exposure. Some severance is justified. Some creators are careful. Some comments are performative. Exposure can follow a decision rather than cause it. Many labels scroll past and never stick. Corpus work alone will not settle causality, so experiments and longitudinal work are required.

The research also has ethical limits. Adversarial simulations must stay abstract. The work cannot become an evasion manual for the behavior it warns about. Until data arrive, the framework is useful for three things: naming a mechanism, offering an instrument, and setting a research agenda.

20 OAPRM Self-Audit

A paper that says every audience matters has to run the same standard on itself. This framework can also be misused. A bad-faith reader could quote it at a real victim and say, "your evidence bar collapsed," when the victim is actually naming a real pattern. A creator could use the language as a costume. A platform could turn measurement into crude censorship. Those are not side notes. They are this paper's own back row.

The guardrail is simple. OAPRM does not decide truth from one label. It asks for pattern, time, power, motive, repair, counterexamples, and evidence that could change the read. Anyone using the framework to dismiss a claim without doing that work is misusing the framework.

21 MIPU/MAP Coding Appendix

These examples show how a coder might use OAPRM without pretending the score is a verdict. The coder asks four questions first. What input is offered? What prediction does it install? What does the viewer now notice or expect? What evidence would make the viewer revise?

The examples below are training cases. Two raters

Table 4: OAPRM self-audit of this paper. Scores are provisional, 0–5.

Dimension	Score	Reason
Semantic drift	1	The paper defines its terms and keeps separating real abuse from ordinary conflict.
Diagnostic inflation	1	It warns against diagnosing people from clips, comments, and one-sided stories.
Evidence corruption	2	It gives strong labels for bad information flow, but Sections 12–14 keep the claims testable.
Intent failure	2	The paper names the mixed audience, including victims, accusers, creators, abusers, and confused viewers.
Detection leakage	3	It discusses evasion risk. It avoids detailed scripts and keeps the adversarial study abstract.
Adversarial adaptation	2	It treats adaptation as a risk to test, not as a settled claim about all abusers.
False-positive susceptibility	2	A bad-faith reader could misuse the framework to dismiss a real victim. The risk is named directly here.
Trust camouflage	2	ZotBot.ai creates a possible commercial incentive, disclosed in the conflict statement.
Severance activation	1	The paper protects no-contact for danger while warning against default severance scripts.
Repair suppression	1	It separates unsafe reconciliation from genuine repair and asks what evidence would change the read.

should score independently, compare disagreements, and revise the codebook. The goal is not to decide whether a real relationship is abusive from one post. The goal is to score how risky the content structure becomes once it enters an open audience.

Coding rule. Score the post, not the creator. Then score the comment culture separately. A careful post can still produce reckless comments. A sloppy post can still spark useful discussion. OAPRM is strongest when it tracks the content, the uptake, and the behavior that follows.

MIPU/MAP rule. A likely MIPU appears when a small input changes what the viewer now treats as signal. A likely MAP appears when several labels start protecting one another, so responses from the target keep getting absorbed into the same story.

22 Who Else Learns

Therapy-speak became a mass-scale informal psychology system while nobody checked enrollment. It now teaches millions how to read relationships, what counts as evidence, which labels explain pain, and when to stop listening to someone they once trusted. That power deserves a higher standard.

Public psychological education happens inside open audiences. The same lesson can reach a victim, an abuser, a false accuser, a confused partner, a performer, a predator, and a regular person having a terrible week. The lesson does not sort them. When the lesson is a simplified detection rule, the risks multiply.

The point of naming the failure is measurement. Judge content by every audience it trains, not only by intention or accuracy. The victim sits in the front row. Ordinary viewers sit in the middle. Someone in the back row may be learning how to look cleaner next time.

Question left running: **who else is watching?** Once landed, abuse-awareness content cannot be evaluated the same way again. That is the MIPU. The map changes from there.

Conflict of Interest Statement

Founded ZotBot.ai, may build tools for cognitive-security analysis, content-risk scoring, and manipulation or pseudo-psychology detection. That creates a possible conflict because the paper proposes measurement systems relevant to such tools. The framework, hypotheses, and studies should be judged independently of any commercial use.

Author Note

This paper stands independently. The terminology comes from a broader research program, but the definitions, hypotheses, scoring rules, and evidence standards needed to judge the argument are contained here.

AI and Document-Preparation Note

This paper is not AI-authored. The framework, claims, examples, and research direction are Michael Zot's. AI tools were used only as document-production support: LaTeX cleanup, layout repair, formatting checks, citation-format cleanup, and PDF compilation. No AI system originated the framework or decided the argument. The argument itself, including MIPU/MAP, OAPRM, detection-rule leakage, and trust camouflage, remains the author's work.

References

Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy, 10*, 44–79.

Almagro, M., & Isern-Mas, C. (2025). Blunting concepts: The double-edged effect of popularizing psychotherapy language. *Philosophical Psychology*. Advance online. doi:10.1080/09515089.2025.2573763

Altmann, B., Fleischer, K., Tse, J., & Haslam, N. (2024). Effects of diagnostic labels on perceptions of marginal cases of mental ill-health. *PLOS Mental Health, 1*(3), e0000096. doi:10.1371/journal.pmen.0000096

Bickham, C., Kazemi-Nia, K., Luceri, L., Lerman, K., & Ferrara, E. (2024). Hidden in plain sight: Intersections of mental health, eating disorders, content moderation on TikTok. *arXiv*, arXiv:2404.15457

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes diffusion of moralized content in social networks. *PNAS*, 114(28), 7313–7318

Coleman, J. (2021). *Rules of estrangement*. Harmony.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). Drivers of misinformation belief and resistance to correction. *Nature Reviews Psychology, 1*, 13–29

Frey, J., Black, K. J., & Malaty, I. A. (2022). TikTok Tourette's: Functional tic-like behavior driven by adolescent social media use? *Psychology Research & Behavior Management, 15*, 3575–3585. doi:10.2147/PRBM.S359977

Harsey, S. J., & Freyd, J. J. (2020). Deny, attack, reverse victim/offender: Credibility effects. *J Aggression Maltreatment Trauma, 29*(8), 897–916

Haslam, N. (2016). Concept creep. *Psychological Inquiry, 27*(1), 1–17

Table 5: Worked OAPRM examples. Scores are illustrative training judgments.

Content sample	Likely predictive update	Risk	Coding rationale
"Five signs they are gaslighting you."	Memory disagreement starts looking like abuse evidence.	4	High semantic drift and false-positive risk unless the post adds repetition, motive, power, correction refusal, and counterexamples. The likely MIPU is fast because the viewer gets a ready-made explanation for confusion.
"Gaslighting is repeated reality manipulation. Memory conflict alone proves little."	Pattern matters more than one sign.	1	The label is useful, but it tells the viewer what would and would not count as evidence. It raises the evidence bar rather than replacing it.
"If they deny it, that is DARVO."	Defense itself becomes confirmation.	5	Strong repair suppression. The viewer is given no route for honest denial, false accusation, memory error, or evidence review. This can create a closed MAP quickly.
"DARVO means denial plus attack plus role reversal across time. Innocent people can deny too."	Look for the full reversal pattern.	1	Keeps the concept while slowing the update down. It protects against checklist abuse and leaves room for correction.
"Go no contact with anyone who questions your healing."	Outside doubt becomes invalidation.	5	Strong severance script. It trains the viewer to treat concern, repair attempts, or disagreement as proof that the other person is unsafe.
"Distance may be needed when contact keeps you unsafe or unstable. Safe repair still needs evidence."	Safety and repair become separate questions.	2	Some severance activation remains, but the post adds limits, evidence, and room for genuine repair.

Isern-Mas, C., & Almagro, M. (2025). Unmasking therapy-speak. *Theoretical Medicine and Bioethics, 46*, 465–489. doi:10.1007/s11017-025-09730-5

Karasavva, V., Miller, C., Groves, N., Montiel, A., Canu, W., & Mikami, A. (2025). Double-edged hashtag: ADHD-related TikTok content perceptions. *PLOS ONE, 20*(3), e0319335

Kerckhoffs, A. (1883). La cryptographie militaire. *J Sci Militaires, 9*, 5–38

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science, 18*(5), 421–428.

Loftus, E. F. (2005). Planting misinformation: 30-year investigation of memory malleability. *Learning & Memory, 12*(4), 361–366

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of auto destruction: Language/memory interaction. *J Verbal Learning Verbal Behavior, 13*(5), 585–589

Lorenzo-Luaces, L., Dierckman, C., & Adams, S. (2023). Attitudes/misinformation about CBT on TikTok. *JMIR, 25*, e45571

Marwick, A. E., & boyd, d. (2011). Tweet honestly, tweet passionately. *New Media & Society, 13*(1), 114–133

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage online. *Science*

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2023). Engagement, satisfaction, amplification of divisive content. *arXiv*, arXiv:2305.16941

Nguyen, V. C., Jain, M., Chauhan, A., Soled, H. J., Alvarez Lesmes, S., Li, Z., Birnbaum, M. L., Tang, S. X., Kumar, S., & De Choudhury, M. (2024). Supporters and skeptics: LLM engagement with mental-health (mis)information on video platforms. *arXiv*, arXiv:2407.02662

Rescorla, E. (2005). Is finding security holes a good idea? *IEEE Security & Privacy, 3*(1), 14–19.

Schneier, B. (2007). Full disclosure of security vulnerabilities a damned good idea. *CSO Online*.

Pillemer, K. (2020). *Fault lines*. Avery.

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs, 44*(3), 197–217

Tseng, E., Bellini, R., McDonald, N., Danos, M., Greenstadt, R., McCoy, D., Dell, N., & Ristenpart, T. (2020). Tools/tactics of intimate partner surveillance. *USENIX Security Symposium*

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2025). Adversarial machine learning: Taxonomy of attacks/mitigations (NIST AI 100-2e2025). National Institute of Standards.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

Walter, N., & Tukachinsky, R. (2020). Meta-analytic exam of continued influence post-correction. *Communication Research, 47*(2), 155–177

Yeung, A., Ng, E., & Abi-Jaoude, E. (2022). TikTok/ADHD cross-sectional. *Can J Psychiatry, 67*(12), 899–906