# Detection-Rule Leakage and Trust Camouflage

## A Function-Over-Form Model of Adversarial Adaptation in Open-Audience Mental-Health Communication

**Michael Zot**
Independent Researcher, Cognitive Security and Digital Psychology
Developer of the MIPU/MAP Cognitive-Security Framework
ORCID: 0009-0001-9194-938X
Official OSF DOI: 10.17605/OSF.IO/XG82Q
Zenodo GitHub Archive DOI: 10.5281/zenodo.20779751

**Abstract**

Public mental-health education helps people name harmful behavior, seek support, and leave abusive dynamics. This paper does not argue against such education. It identifies a narrower adversarial failure mode in open-audience psychological communication: the same detection rules that help victims and bystanders identify harm may also teach motivated actors what to stop displaying.

The proposed framework identifies *Detection-Rule Leakage*: the process by which public abuse-detection rules, clinical vocabulary, and relational safety scripts become visible to the actors those rules are meant to expose. Once detection rules become public, an actor can update presentation style by replacing overt hostility with socially approved psychological language. This adapted presentation is called *Trust Camouflage*. The behavioral function may remain coercive, avoidant, or reality-distorting while the surface language appears calm, therapeutic, accountable, and boundary-focused.

The paper's central safeguard is *function over form*. Public phrase checklists are vulnerable to Goodhart-style optimization: once a phrase becomes a target, it can be avoided, copied, or polished. This framework therefore does not ask observers to memorize suspicious phrases. It asks whether a speaker's language preserves repair, evidence, accountability, and proportional consequences over time. The model draws on scholarship in coercive control, DARVO/accountability reversal, gaslighting-related reality distortion, therapy-speak drift, concept creep, impression management, framing effects, Goodhart's Law, Campbell's Law, moral-emotional diffusion online, engagement-based ranking systems, and limitations of reinforcement learning from human feedback. These are theoretical anchors, not a complete list of abuse tactics, coping strategies, or interpersonal maneuvers. The model's claim is broader: any language, label, or conflict strategy should be evaluated by function over time, including whether it preserves repair, evidence, accountability, proportional consequences, and pattern visibility, or repeatedly makes the original harm harder to confront. It is presented as a falsifiable cognitive-security framework and pre-validation measurement proposal, not as a diagnostic tool.

**Keywords:** detection-rule leakage; trust camouflage; coercive control; therapy-speak; gaslighting; cognitive security; LLM safety

### Central Contribution

This manuscript formalizes a specific adversarial sequence: public detection rules teach victims and bystanders what to notice, but they can also teach motivated actors what to stop displaying. The proposed safeguard is function-over-form detection: evaluate what language does to repair, evidence, accountability, proportional consequences, and longitudinal pattern rather than trusting surface calm or approved psychological vocabulary.

# Contents

# 1 Problem Statement

Modern public psychology often assumes a single intended audience: the person being harmed. This assumption is incomplete. Online mental-health content is published into an open environment where victims, bystanders, clinicians, institutions, social media users, AI systems, and manipulative actors can all read the same guidance.

When abuse-detection content becomes public, it can produce two simultaneous outcomes:

1. **Protective literacy:** Victims and bystanders learn language for naming harm.
2. **Adversarial literacy:** Motivated actors learn which behaviors, phrases, tones, and emotional displays have become socially detectable.

The second outcome is the core risk. A motivated actor does not need to change the underlying behavior to reduce detection. The actor only needs to stop producing the signals that observers have been trained to flag. This paper names that failure mode *Detection-Rule Leakage*.

The central claim is deliberately narrow:

> Public abuse-detection content can create an adversarial learning loop in which manipulative actors use published detection rules to reduce visible red flags while preserving coercive function.

This claim does not reject public abuse education. It argues that abuse education becomes stronger when it includes adversarial modeling.

### Scope, Status, and Research-Use Guardrails

**Population and gender scope:** This framework is not a claim about any sex, gender, diagnosis, disability, trauma history, or identity category. It may apply wherever trusted psychological language is used to alter accountability conditions. Where cited intimate-partner violence and coercive-control literatures document gendered patterns, that empirical scope is preserved rather than generalized beyond the evidence.

**Scope and status:** This manuscript is a conceptual, falsifiable research model and pre-validation measurement proposal. It does not diagnose individuals, replace clinical assessment, or argue that therapeutic language is inherently manipulative. Its empirical target is observable function over time: whether language preserves repair, reverses blame, dissolves evidence, removes consequences, or repeatedly shifts the burden of proof away from the original behavior.

**Research-use guardrail:** The Trust Camouflage Function Score introduced below is exploratory coding for research contexts only. It is not a clinical instrument, diagnostic scale, custody tool, HR tool, or relationship verdict.

# 2 Function Over Form: The Anti-Goodhart Thesis

A direct objection to this paper is that it could reproduce the failure it describes. If public detection rules create evasion, then a paper about Trust Camouflage could itself become a new evasion manual. This is the **self-Goodhart problem**.

The answer is to avoid phrase-based detection. Phrases are easy to copy, avoid, and optimize against. Functions are harder to fake because they require outcomes. A person can change wording indefinitely. A person cannot fake repair without producing repair, cannot preserve accountability while deleting consequences, and cannot preserve evidence while making evidence unusable.

**Function-Over-Form Invariant**

**The framework is anti-checklist by design.** It does not identify Trust Camouflage by the presence of words such as boundary, trauma, safety, dysregulation, accountability, or nervous system. Those terms can be legitimate, clinically useful, and protective.

Trust Camouflage is evaluated only by repeated function:

1. Did the language preserve or block repair?
2. Did it keep the original behavior available for discussion?
3. Did it preserve evidence or dissolve it into subjectivity?
4. Did it allow proportional consequences or make all consequences appear abusive?
5. Did this pattern repeat across time and documented repair attempts?

**Single-utterance scoring is invalid.** A single phrase should not be coded as Trust Camouflage. The unit of analysis is a longitudinal pattern involving documented behavior, repair attempts, evidence handling, accountability outcomes, and proportional consequences.

**Victim Protection Clause.** The model should not be used against a person who is responding to documented harm, preserving evidence, participating in repair, accepting proportional consequences, or requesting reasonable accommodation. In those cases, therapeutic language may be protective and legitimate rather than camouflage.

## 3 Theoretical Foundations

The framework draws from ten established research areas. The aim is not to replace these literatures, but to connect them into a specific failure chain.

### 3.1 Coercive Control

Coercive control research shows that abuse is often patterned, contextual, and non-physical. Stark describes coercive control as a broader system of entrapment, not merely a sequence of isolated incidents [1]. Dutton and Goodman conceptualize coercion in intimate partner violence through demands, credible threats, surveillance, resources, and social ecology [2]. Digital-media research extends this analysis into technology-facilitated coercive control [3].

### 3.2 DARVO and Accountability Reversal

DARVO stands for Deny, Attack, and Reverse Victim and Offender. It describes a defensive response pattern in which a person accused of harm denies wrongdoing, attacks the accuser's credibility, and repositions themselves as the harmed party [6, 7]. Trust Camouflage can include DARVO, but it is narrower in one respect and broader in another: it focuses on accountability blocking through trusted clinical or safety language, especially after that language has circulated publicly.

### 3.3 Gaslighting and Reality Distortion

Gaslighting research frames reality distortion as social and relational, not merely individual confusion. Sweet argues that gaslighting is rooted in power-laden relationships and social inequalities [4]. Recent

interdisciplinary review work also shows the term's broad evolution and operational ambiguity across fields [5]. This matters because therapy-speak can soften the surface of reality distortion. A sentence can sound reflective while still making evidence harder to discuss.

### 3.4 Therapy-Speak Drift and Concept Creep

Therapy-speak refers to the movement of clinical and therapeutic language into everyday communication. Isern-Mas and Almagro describe therapy-speak as the imprecise and superficial integration of psychotherapy-derived language into ordinary speech, while also noting that such language can be used to discredit individuals and evade responsibilities [8]. Haslam's work on concept creep explains how harm-related psychological concepts can expand outward and downward over time [9]. This paper focuses on one subset of that drift: adversarial or accountability-blocking use of trusted clinical vocabulary.

### 3.5 Goodhart, Campbell, and Detection Metrics

Goodhart's Law is often summarized as the problem that a measure loses reliability once it becomes a target [10]. Campbell's Law makes a similar point about social indicators: when a metric becomes important for decision-making, it becomes vulnerable to corruption pressure [11]. Abuse-detection checklists can be subject to the same structural risk. Once the checklist is public, a manipulative actor can optimize against it.

### 3.6 Impression Management and Framing Effects

Impression management research shows that people can shape how others perceive a situation through performed presentation [12]. Framing research shows that the wording and framing of a situation can alter judgment and choice [13]. Trust Camouflage builds on these older insights but gives them a narrower adversarial target: the use of publicly trusted mental-health language to shift accountability away from observable conduct.

### 3.7 Platform Incentives

Social media research shows that moral-emotional language can increase diffusion online [14]. Vosoughi, Roy, and Aral show that false news online can spread farther and faster than true news [15]. Work on engagement-based ranking also shows that social media systems can amplify divisive content [16]. These dynamics matter because label-first mental-health content is often simple, emotionally loaded, and easy to identify with. Mechanism-first content is slower and harder to spread.

### 3.8 LLMs and Human Feedback

Large language models trained with reinforcement learning from human feedback can produce outputs that users prefer, but RLHF also has unresolved limitations involving reward modeling, preference capture, oversight, and auditing [17, 18]. In this paper, the AI risk is narrow: models may help convert hostile or coercive content into language that appears calm, safe, and psychologically literate. Recent work on LLM gaslighting risk supports the need to study manipulation-relevant failure modes directly [19].

### 3.9 AI Disclosure and Scholarly Responsibility

Academic publishing norms increasingly expect disclosure when generative AI materially shapes a manuscript [20, 21]. The disclosure in this paper follows that principle: assistance with structure and phrasing is identified, while authorship and responsibility remain with the human author.

### 3.10 Institutional Credibility and Legal Systems Abuse

Coercive control can interact with institutional credibility judgments. Reeves and colleagues describe how legal systems abuse and credibility problems affect victim-survivors in coercive-control contexts [22]. This paper extends that concern into a broader parser problem: institutions may overvalue composure and approved language while undervaluing pattern, repair history, and consequence avoidance.

## 4 Novel Contribution

This framework does not claim that manipulation, gaslighting, DARVO, concept creep, therapy-speak misuse, impression management, or framing effects are new. Its proposed contribution is the complete adversarial sequence.

This manuscript also does not claim that the weaponization of therapy-speak is new. Recent work already identifies therapy-speak as vulnerable to superficial use, credibility distortion, discrediting, and responsibility evasion [8]. The narrower contribution here is the proposed adversarial sequence: public detection rules become visible to motivated actors, those actors adapt to the rules, trusted psychological language becomes a credibility update, and observers may overvalue surface form unless they evaluate repair, evidence, accountability, consequences, and pattern over time.

**Proposed Contribution**

1. Public detection rules are broadcast into an open audience.
2. The rules become simplified into checklist language.
3. Manipulative actors can study the checklist and reduce visible red flags.
4. Trusted psychological language becomes a credibility update.
5. Institutions, bystanders, and AI systems may overweight surface calm and underweight behavioral function.
6. The harmed person must fight both the original behavior and the social authority of the label.
7. A function-over-form model can reduce this failure by scoring repair, evidence, accountability, consequences, and longitudinal pattern rather than phrases.

The model is therefore not merely "people misuse therapy words." The stronger claim is that public detection systems can create their own evasion training data unless they teach observers to evaluate function, evidence, pattern, and repair.

## 5 Core Definitions

### Primary Terms

**Detection-Rule Leakage:** The process by which public safety rules, abuse checklists, diagnostic labels, or institutional criteria become visible to the actors those rules were meant to identify.

**Trust Camouflage:** The use of trusted psychological, therapeutic, institutional, or safety language to conceal, soften, justify, or reframe harmful behavior. Trust Camouflage is not detected by language alone. It is detected by function over time.

**MIPU:** In this manuscript, MIPU refers to a minimal trust-bearing cue that creates high-friction reversal cost: a substantial increase in the cost of reversing an observer's interpretation. The construct is grounded in framing effects [13] and impression management [12].

**MAP: Maximal Anticipatory Pattern.** The broader interpretive structure that forms after the MIPU. Once activated, the observer begins predicting the speaker as vulnerable, regulated, harmed, unsafe, or psychologically sophisticated.

**Accountability Shield:** A language frame that makes ordinary accountability appear harmful, unsafe, coercive, or clinically insensitive.

**Label-Conversion Farm:** A social-media account or content pattern that converts complex relational behavior into fast identity labels for engagement, usually without adversarial modeling.

**Invalid-use rule:** Trust Camouflage should not be inferred from a single utterance, a single boundary, a single request for accommodation, or the mere presence of therapeutic language. It requires longitudinal pattern, evidence handling, repair history, and accountability outcome.

## 6 Construct Boundaries and Discriminant Validity

A predictable reviewer concern is that Trust Camouflage may be redundant with DARVO, gaslighting, coercive control, concept creep, or impression management. The framework is related to all of these, but it names a more specific failure chain.

| Construct | Core mechanism | Primary target | What Trust Camouflage adds |
|---|---|---|---|
| DARVO | Denial, attack, and reversal of victim/offender roles. | Accountability reversal after confrontation. | Shows how reversal can be performed through trusted therapeutic or safety language rather than obvious attack. |
| Gaslighting | Reality distortion that destabilizes the target's confidence in what occurred. | Reality-testing and evidence confidence. | Shows how clinical language can soften contradiction and make evidence disputes appear psychologically refined. |
| Coercive control | Patterned domination, restriction, surveillance, threat, or entrapment. | Power, autonomy, and space for action. | Scores whether approved language preserves coercive function while reducing visible detectability. |
| Concept creep | Expansion of harm-related concepts. | Semantic boundaries of psychological categories. | Focuses on adversarial use of expanded concepts to block repair or consequences. |

| Construct | Core mechanism | Primary target | What Trust Camouflage adds |
|---|---|---|---|
| Impression management | Strategic control of how one is perceived. | Social presentation. | Narrows the analysis to public detection-rule learning and accountability blocking with institutionally trusted vocabulary. |
| Trust Camouflage | Trusted language preserves harmful function while improving surface credibility. | Repair, evidence, accountability, consequences, and pattern over time. | Adds the full sequence: public detection rule, adversarial learning, credibility update, institutional parser failure, and function-over-form scoring. |

The discriminant claim is therefore precise: *Trust Camouflage is accountability blocking via institutionally trusted psychological language, learned or optimized through exposure to public detection content, and measurable by function over time.*

# 7 The Detection-Rule Leakage Sequence

The mechanism follows a repeatable chain:

1. **Checklist broadcast:** A public post teaches an audience to detect manipulation, abuse, narcissism, gaslighting, trauma responses, or boundary violations.
2. **Social simplification:** Complex clinical material becomes compressed into viral phrases.
3. **Metric conversion:** Observers start using those phrases as quick detection rules.
4. **Adversarial reading:** Motivated actors learn which phrases, tones, and behaviors now create suspicion.
5. **Presentation adaptation:** The actor removes obvious red flags and adopts trusted vocabulary.
6. **Observer lock-in:** The victim, institution, bystander, or AI system interprets the actor through the new therapeutic frame.
7. **Accountability collapse:** The original behavior becomes harder to confront because confrontation now appears to violate a boundary, invalidate a trauma response, or attack vulnerability.

The harmed person is no longer only fighting the behavior. They are fighting the social authority of the label.

# 8 MIPU/MAP Model

## 8.1 MIPU: The Smallest Update That Changes the Burden of Proof

A MIPU occurs when one trust-bearing cue changes the burden of proof inside a conflict. Before the cue, the issue may be a broken promise, contradiction, or harm. After the cue, the harmed person may have to prove that asking for repair is not itself harmful.

For peer-review clarity, MIPU is treated here as a minimal trust-bearing cue that can raise the cost of returning attention to the original behavior. Once a trusted label enters a conflict, the

observer's threshold for challenging the speaker may rise. This is related to broader work on framing and social presentation [13, 12], but the MIPU claim is narrower: in accountability conflicts, a small cue can change who must prove what.

In practice, MIPUs usually appear as *burden-shifting claims* rather than as fixed phrases. They may move attention from the original behavior to the harmed person's tone, convert a repair request into an alleged safety violation, turn a factual contradiction into a dispute about subjective reality, or use a clinical label to suspend consequences.

These are categories, not scripts. The model should not be used as a phrase checklist. The feelings or labels named by a speaker may be real, valid, and protective. The coding question is narrower: did the cue preserve repair, evidence, and accountability, or did it make them harder to reach?

### 8.2 MAP: The Interpretive Lock-In

Once the MIPU lands, the observer may enter a MAP state. The observer now anticipates the speaker as wounded, psychologically aware, and in need of accommodation. The target of harm may become afraid to press for repair because pressing for repair now looks cruel, unsafe, or uninformed.

The MAP state produces three predictable distortions:

1. **Causality drift:** Attention moves from the original behavior to the speaker's emotional state.
2. **Burden reversal:** The harmed person must prove they are not being harmful by asking for accountability.
3. **Evidence softening:** Concrete facts become recoded as interpretation, tone, projection, or emotional flooding.

## 9 TCFS Coding Protocol

The Trust Camouflage Function Score (TCFS) is a pre-validation coding protocol for research use only. It is not a clinical instrument, diagnostic scale, custody tool, HR tool, or relationship verdict. It is designed for vignette studies, content coding, inter-rater reliability testing, and construct validation.

**Single-utterance scoring is invalid.** TCFS should only be applied to a documented sequence that includes the triggering event, the disputed language, repair attempts, evidence handling, accountability outcome, and pattern history.

| Marker | 0 = function preserved | 1 = function weakened | 2 = function blocked |
|---|---|---|---|
| Repair outcome | Clear repair step occurs or is scheduled within a protocol-defined timeframe. | Repair becomes vague, delayed, or conditional without clear reason. | Repair disappears or is framed as harmful, unsafe, or coercive. |
| Accountability outcome | Speaker owns the relevant behavior and keeps it discussable. | Responsibility becomes mixed or unclear. | Harmed person becomes the primary offender for requesting accountability. |
| Evidence outcome | Evidence remains discussable and checkable. | Evidence is mixed with feelings without clear comparison. | Evidence is dismissed as interpretation, projection, or reality-overwriting. |
| Consequence outcome | Proportional consequences remain possible. | Consequences are softened or delayed without clear repair logic. | Any consequence is framed as abusive, unsafe, retaliatory, or coercive. |
| Pattern outcome | Language fits a repair-producing pattern over time. | Pattern is ambiguous or insufficiently documented. | Language repeatedly appears when accountability is requested and repair then vanishes. |

No diagnostic thresholds are proposed. A summed score may be used as an exploratory severity index in research, but interpretation requires inter-rater reliability, base-rate estimates, factor analysis, preregistered vignette testing, and validation against independent behavioral outcomes. Until validated, TCFS should be reported as **TCFS exploratory coding**, not as a clinical or institutional determination.

## 10 Trust Camouflage Examples and Counterexamples

The following examples are not diagnostic scripts. They are detection examples. The key question is not whether a phrase sounds therapeutic. The key question is what the phrase does to accountability. These examples are illustrative only; any new language that blocks repair while sounding safe should be scored by function, evidence, and pattern rather than by memorized wording.

## Interpersonal Function Tests

**Hard rule:** Do not score a single sentence. Score only documented pattern, repair attempts, evidence handling, and accountability outcome. TCFS coding is invalid under the Victim Protection Clause when the speaker is responding to documented harm, preserving evidence, accepting consequences, or participating in repair.

---

**1. Responsibility Avoidance**
*Observable event:* A shared obligation was ignored.
*Camouflage marker:* The actor frames accountability as pressure, overwhelm, dysregulation, or lack of capacity.
*Function test:* Did the phrase produce a repair plan, or did it make the obligation disappear?

---

**2. Conversation Avoidance**
*Observable event:* A necessary repair conversation is repeatedly delayed.
*Camouflage marker:* The actor frames repeated requests for clarity as unsafe, intrusive, or boundary-violating.
*Function test:* Did the boundary define when and how repair will happen, or did it permanently block repair?

---

**3. Blame Relocation**
*Observable event:* The actor caused harm.
*Camouflage marker:* The actor moves causality from their behavior to the target's tone, timing, trauma, facial expression, or emotional intensity.
*Function test:* Is the speaker addressing the harm they caused, or are they converting the harmed person's reaction into the central offense?

---

**4. Reality Dissolution**
*Observable event:* A contradiction, denial, or revision of facts occurred.
*Camouflage marker:* The actor turns factual disagreement into symmetrical subjectivity.
*Function test:* Are both accounts being checked against evidence, or is evidence being replaced by therapeutic relativism?

---

**5. Emerging Label Capture**
Public conflict language increasingly borrows terms related to demand avoidance, rejection sensitivity, nervous-system safety, trauma responses, and attachment wounds. These experiences can be real. The risk is not the label itself. The risk is using the label to cancel ordinary accountability.
*Function test:* Does the label help create a fair accommodation plan, or does it make all expectations impossible?

### 10.1 Counterexamples

A statement should *not* be coded as Trust Camouflage merely because it uses therapy language. Examples of lower-risk or legitimate use include:

- A person sets a boundary and also identifies a specific time, method, or condition for repair.
- A person names dysregulation and still takes responsibility for the behavior that harmed someone.
- A person requests accommodation while preserving the other person's right to evidence, consequences, and repair.
- A person disagrees about facts but remains willing to compare timelines, messages, witnesses, or records.

- A person uses clinical language while accepting proportional consequences.

# 11 Institutional Vulnerability Audits

Institutions often reward composure, consistency, documentation, and approved language. Those values are useful. They are also exploitable. The risk is not that institutions are malicious. The risk is that institutions can overweight presentation and underweight pattern.

## Institutional Failure Modes

**1. Couples Therapy and Clinical Mediation**
A covertly controlling person may perform reflection in-session while the harmed partner appears reactive, angry, or disorganized. If the clinician treats asymmetric coercion as mutual dysregulation, the therapeutic room can unintentionally discipline the victim's alarm while validating the actor's mask.
*Audit question:* Is the clinician tracking the full timeline of control, repair avoidance, contradiction, and consequence removal, or only the emotional presentation inside the session?

---

**2. Corporate HR Disputes**
HR systems often flatten patterns into isolated incidents. A polished actor can frame control as professionalism and frame pushback as hostility. The harmed employee then faces the difficult task of explaining a pattern inside a process that may only want discrete events.
*Audit question:* Does the institution evaluate repeated function and power asymmetry, or only meeting behavior and tone?

---

**3. Family Court and Custody**
Family court can mistake calmness for safety and emotional intensity for instability. In coercive-control cases, the victim may appear distressed because the pattern was designed to produce distress.
*Audit question:* Does the court distinguish emotional presentation from behavioral pattern, documentation history, legal systems abuse, and control tactics?

---

**4. Platform Moderation and AI Safety**
Moderation systems and AI assistants often evaluate surface language. A hostile message rewritten in therapeutic vocabulary may pass as safe, while a victim's blunt description of harm may be flagged as hostile.
*Audit question:* Does the system classify only tone, or does it evaluate behavioral function?

# 12 Algorithmic and AI-Mediated Failure Modes

## 12.1 Platform Reward Distortion

Social platforms reward content that is legible, emotionally charged, and easy to identify with. A simple post saying "your ex was a narcissist" may spread faster than a careful explanation of adversarial adaptation. This creates a content ecology where labels travel faster than models.

The result is a feedback loop:

1. Simple labels spread.
2. The public learns the labels.
3. Motivated actors learn the labels.
4. New camouflage forms around the labels.

5. More labels are published to detect the new camouflage.

Without adversarial design, public psychology can unintentionally become an evasion resource.

### 12.2 LLM-Assisted Camouflage

Large language models can convert hostile intent into polished language. A user can ask an AI system to make a coercive message sound calm, boundaried, trauma-informed, or accountable. The output may remove obvious aggression while preserving the original function.

The risk is not that AI creates manipulation from nothing. The risk is that AI can lower the cost of rhetorical laundering. This claim should be tested through controlled benchmarks that compare original harmful prompts, AI-polished outputs, and blind function coding.

### 12.3 RLHF and the Empathy Trap

Models trained to be helpful, validating, and safe can become over-responsive to therapeutic labels. In some contexts, that is beneficial. In adversarial contexts, it creates a weakness: the system may validate the label before testing the function of the behavior.

A safer AI response pattern would separate three layers:

1. The person's stated feeling.
2. The observable behavior.
3. The accountability outcome.

A model should be able to say: "The feeling may be real. The behavior still requires repair."

## 13 Falsifiable Hypotheses

This framework is testable. The following hypotheses convert the model into research programs.

**Testable Claims**

**H1: Detection-rule exposure increases camouflage quality.**
Participants exposed to abuse-detection checklists will produce more detection-evading responses in simulated conflict tasks than participants not exposed to those checklists.

**H2: Therapeutic vocabulary reduces accountability perception.**
Observers will rate the same avoidant or coercive behavior as less harmful when it is wrapped in therapeutic language.

**H3: Tone-polished harmful messages outperform blunt victim messages in institutional review.**
HR, mediation, or moderation-style evaluators will judge polished actor statements as more credible than distressed victim statements, even when the victim has stronger evidence.

**H4: LLM rewriting increases perceived maturity while preserving coercive function.**
AI-rewritten manipulative messages will be rated as more emotionally mature and less hostile than originals, while independent coders will still detect the same accountability-blocking function.

**H5: Label-first content spreads faster than mechanism-first content.**
Posts that convert behavior into identity labels will receive higher engagement than posts explaining detection-rule leakage, even when both discuss the same relational harm.

**H6: MAP lock-in can be reduced through function testing.**
Observers trained to ask "what did this phrase do to accountability?" will be less vulnerable to Trust Camouflage than observers trained only on red-flag checklists.

## 14 Study Designs and Minimum Pilot Path

The framework can be tested without diagnosing real individuals.

1. **Vignette experiment:** Present participants with identical relational conflicts. Vary only whether the actor uses neutral language, hostile language, or therapeutic language. Measure perceived harm, credibility, repair obligation, blame assignment, and willingness to recommend consequences.

2. **Minimum pilot:** A small pre-registered vignette pilot can test whether therapeutic vocabulary reduces accountability perception. Even an initial sample of 40–100 participants can estimate effect direction, variance, and item clarity before larger validation.

3. **Adversarial writing task:** Under controlled experimental conditions and ethics oversight, randomly assign participants to read abuse-detection checklists or neutral content. Ask them to transform a controlling message so that it appears less detectable, then have blind coders rate the degree of Trust Camouflage using TCFS. Stimuli should be debriefed and designed to test adaptation without publishing usable interpersonal scripts.

4. **Institutional review simulation:** Give HR-style, mediation-style, or moderation-style reviewers paired statements from a harmed person and an actor. Vary evidence strength and tone polish. Measure credibility judgments.

5. **LLM rhetorical laundering benchmark:** Compare original hostile prompts with AI-polished outputs. Code whether hostility is removed while accountability-blocking function remains. Publish aggregate ratings and codebooks without releasing reusable manipulation scripts.

6. **Content-diffusion analysis:** Compare label-first and mechanism-first posts about the same relational harm category. Use engagement metrics as the outcome variable.

## 15 Operational Markers

The following markers help identify possible Trust Camouflage. No single marker is conclusive. The model requires pattern analysis.

1. **Repair disappearance:** A therapeutic phrase appears, and the original repair obligation vanishes.

2. **Consequence immunity:** The actor's label makes consequences appear abusive.

3. **Tone substitution:** Calm presentation replaces factual accountability.

4. **Evidence relativism:** Concrete contradictions become "different realities."

5. **Burden reversal:** The harmed person must prove that accountability is safe enough to request.
6. **Asymmetrical compassion:** The actor's pain receives full recognition while the target's harm is minimized.
7. **Perpetual deferral:** Every repair conversation is delayed by new language about readiness, safety, capacity, or regulation.

## 16 Ethical Guardrails

This framework can be misused if applied carelessly. Several limits are necessary.

1. **Do not treat therapy language as proof of manipulation.** A phrase can be sincere, clinically accurate, and necessary.
2. **Do not score single utterances.** Trust Camouflage is invalid as a one-sentence verdict.
3. **Do not diagnose from conflict language.** This model evaluates function, not mental illness.
4. **Do not punish vulnerability.** The goal is to distinguish vulnerability that leads to repair from vulnerability language that blocks repair.
5. **Do not remove accommodations.** Neurodivergence, trauma, disability, and psychiatric distress can require real accommodation.
6. **Track behavior over time.** Trust Camouflage is a pattern claim, not a phrase claim.
7. **Use evidence.** The model gains strength only when tied to timelines, contradictions, repair attempts, and outcomes.
8. **Separate function from intent.** Intent may be unknown. The measurable question is what the phrase repeatedly does to repair, evidence, and accountability.
9. **Apply the Victim Protection Clause.** TCFS coding is invalid when the speaker is responding to documented harm, preserving evidence, accepting consequences, or participating in repair.

## 17 Observer Restoration Protocol

When Trust Camouflage succeeds, the harmed person may feel guilty for naming harm. The following protocol breaks the MAP lock without denying the other person's stated feelings.

**MAP-Breaker Protocol**

**Step 1: Label Detachment**
Separate the label from the event.
*Internal command:* "The label may be real. It is not the whole event."

**Step 2: Raw Behavior Recovery**
Translate the therapeutic language back into observable action.
*Questions:* What happened before the label appeared? What promise was broken? What repair never happened? What consequence disappeared?

**Step 3: Function Test**
Judge the phrase by its outcome.
*Question:* Did this phrase create repair, or did it block accountability?

**Step 4: Baseline Reality Restoration**
Return to the original reality before the label hijacked the frame.
*Script:* "I accept that you feel [unsafe / dysregulated / overwhelmed]. The original issue is still [behavior]. The impact is still [harm]. The next step still needs to be [repair]."

**Step 5: Timeline Reconstruction**
Write the sequence in plain behavior language.
*Format:* Event. Response. Label introduced. Accountability outcome. Pattern history.

## 18 Science Communication Implication

Structural explanations spread poorly when they require the audience to implicate itself. Social media rewards fast recognition, identity confirmation, and moral-emotional compression. Detection-rule leakage requires second-order reasoning. It asks the audience to see that the tool used for protection can also become training data for evasion.

Therefore, public education about Trust Camouflage should include two layers:

1. **Surface hook:** A familiar relational problem.
2. **Mechanism payload:** The adversarial model.

Examples of compressed public communication include:

**Public Communication Lines**

**The Two-Audience Rule:** "Every abuse checklist has two audiences: the person learning what to notice, and the person learning what to stop showing."

**The New Camouflage:** "Old red flags were easier to hear. New camouflage can sound psychologically literate."

**The Boundary Test:** "A real boundary defines conduct. A camouflage boundary deletes accountability."

**The Repair Test:** "The question is not whether the phrase sounds therapeutic. The question is whether it produced repair."

## 19 Limitations

This paper proposes a conceptual framework. Several limitations remain.

First, Trust Camouflage is not yet validated as a clinical construct. It should be treated as a cognitive-security model until tested.

Second, TCFS is not a validated instrument. No clinical or institutional threshold is proposed. Future work must test inter-rater reliability, base rates, factor structure, and predictive validity.

Third, therapeutic language is often legitimate. Many people use terms like boundaries, dysregulation, trauma, and safety accurately. The model must not become an excuse to dismiss real distress.

Fourth, intent is difficult to prove. The stronger empirical target is function: whether language repeatedly blocks repair, reverses blame, erases facts, or removes consequences.

Fifth, institutional settings differ. Courts, HR departments, therapy rooms, and AI systems have different standards of evidence. A single model cannot replace domain-specific evaluation.

Sixth, public communication creates a paradox. Naming the camouflage can also leak the detection rule. This is unavoidable. The solution is not secrecy. The solution is adaptive literacy: teach people to evaluate function, evidence, pattern, and repair rather than memorizing static red flags.

Seventh, the model risks false positives if used without context. A high-conflict exchange is not automatically coercive control. The safest use is longitudinal: multiple events, documented behavior, repair attempts, contradictions, and outcomes.

## 20 Conclusion

Open-audience psychology changed the social environment by giving victims language, institutions new categories, platforms endless engagement material, and manipulative actors a public map of what to hide.

Trust Camouflage names one proposed next-stage failure mode in that environment: the old red flags are no longer enough when a person can sound calm, healed, boundaried, trauma-informed, and accountable while still using language to erase harm.

The answer is not less psychology; the answer is psychology designed for adversarial conditions.

A useful detection model must ask:

> What did the phrase do to the facts?
> What did it do to repair?
> What did it do to accountability?
> Who became responsible after the label appeared?

This framework is anti-checklist by design. It does not ask readers to memorize banned phrases. It asks them to test whether language preserves repair, evidence, accountability, and proportional consequences over time.

Future work should test whether function-over-form training improves detection of adaptive accountability-blocking strategies.

## Research Materials and Archive Links

The official peer-review package, study materials, coding protocol, preregistration materials, benchmark materials, analysis scripts, and validation roadmap are archived on OSF:

- Official OSF research package DOI: https://doi.org/10.17605/OSF.IO/XG82Q

A public GitHub mirror is available for repository inspection, code review, and collaboration:

- GitHub materials mirror: https://github.com/mikecreation/detection-rule-leakage-trust-camouflage

The GitHub release is archived on Zenodo for repository preservation and indexing:

- Zenodo GitHub archive DOI: https://doi.org/10.5281/zenodo.20779751

The OSF archive should be treated as the canonical citation source for the peer-review package.

## Declarations

**Author contribution:** Michael Zot is the sole author and is responsible for the conceptual framework, core terms, examples, hypotheses, interpretations, manuscript preparation, and final claims.

**Generative AI use:** AI assistance was limited to organization, formatting, and sentence-level copyediting. All references were manually verified by the author against primary sources. No AI system was used as a source of empirical evidence, clinical authority, theoretical authority, or citation content. The conceptual framework, core terms, examples, hypotheses, interpretations, and final claims are the author's sole responsibility.

**Funding:** No external funding is reported.

**Competing interests:** The author declares no competing interests relevant to this manuscript.

**Ethics approval:** This conceptual manuscript reports no human-subjects data. Future empirical studies using participants should obtain appropriate ethics review or exemption before data collection.

**Data and materials availability:** Research materials are available through the OSF project listed above. The TCFS is a pre-validation research protocol and should not be used as a clinical, legal, custody, HR, or diagnostic determination.

## References

[1] Stark, E. (2007). *Coercive Control: How Men Entrap Women in Personal Life*. Oxford University Press.

[2] Dutton, M. A., & Goodman, L. A. (2005). Coercion in intimate partner violence: Toward a new conceptualization. *Sex Roles, 52*(11–12), 743–756. https://doi.org/10.1007/s11199-005-4196-6

[3] Dragiewicz, M., Burgess, J., Matamoros-Fernandez, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies, 18*(4), 609–625. https://doi.org/10.1080/14680777.2018.1447341

[4] Sweet, P. L. (2019). The sociology of gaslighting. *American Sociological Review, 84*(5), 851–875. https://doi.org/10.1177/0003122419874843

[5] Darke, L., Paterson, H. M., & van Golde, C. (2025). Illuminating gaslighting: A comprehensive interdisciplinary review of gaslighting literature. *Journal of Family Violence*. Advance online publication. https://doi.org/10.1007/s10896-025-00805-4

[6] Harsey, S. J., Zurbriggen, E. L., & Freyd, J. J. (2017). Perpetrator responses to victim confrontation: DARVO and victim self-blame. *Journal of Aggression, Maltreatment & Trauma, 26*(6), 644–663. https://doi.org/10.1080/10926771.2017.1320777

[7] Harsey, S. J., Adams-Clark, A. A., & Freyd, J. J. (2024). Associations between defensive victim-blaming responses (DARVO), rape myth acceptance, and sexual harassment. *PLOS ONE, 19*(12), e0313642. https://doi.org/10.1371/journal.pone.0313642

[8] Isern-Mas, C., & Almagro, M. (2025). Unmasking therapy-speak. *Theoretical Medicine and Bioethics, 46*(6), 465–489. https://doi.org/10.1007/s11017-025-09730-5

[9] Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry, 27*(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

[10] Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. *Papers in Monetary Economics, 1*(1), 1–20.

[11] Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

[12] Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books.

[13] Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683

[14] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

[15] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[16] Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus, 4*(3), pgaf062. https://doi.org/10.1093/pnasnexus/pgaf062

[17] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[18] Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=bx24KpJ4Eb

[19] Li, W., Zhu, L., Song, Y., Lin, R., Mao, R., & You, Y. (2025). Can a large language model be a gaslighter? *International Conference on Learning Representations*. https://openreview.net/forum?id=RQPSPGpBOP

[20] Perkins, M., & Roe, J. (2024). Academic publisher guidelines on AI usage: A ChatGPT supported thematic analysis [peer review: 3 approved, 1 approved with reservations]. *F1000Research, 12*, 1398. https://doi.org/10.12688/f1000research.142411.2

[21] Cleland, J., Driessen, E., Masters, K., Lingard, L., & Maggio, L. A. (2026). When and how to disclose AI use in academic publishing: AMEE Guide No. 192. *Medical Teacher, 48*(4), 542–553. https://doi.org/10.1080/0142159X.2025.2607513

[22] Reeves, E., Fitz-Gibbon, K., Meyer, S., & Walklate, S. (2025). Incredible women: Legal systems abuse, coercive control, and the credibility of victim-survivors. *Violence Against Women, 31*(3–4), 767–788. https://doi.org/10.1177/10778012231220370