Detecting Regulatory States in Natural Language: A Validation Study of the Four-Mode Gradient Framework

PREPRINT — Not Peer Reviewed

Anna Paretas-Artacho
Independent Researcher; TEG-Blue Research
research@teg-blue.org
ORCID: 0009-0005-2394-7162

Pre-registration: https://osf.io/f4x6y
January 2026 | Version 1.0

Abstract

This study presents the first empirical validation of the Four-Mode Gradient, a theoretical framework proposing that human regulatory states can be classified into four modes—Connection, Protection, Control, and Domination—based on perceived safety or threat. Using natural language analysis of 10,000+ Reddit "Am I The Asshole" (AITA) posts, we tested whether these four modes could be detected using validated psychological constructs from polyvagal theory, attachment research, contempt markers (Gottman), and moral disengagement theory (Bandura).

Results: All four modes were successfully detected in natural language. Mode classifications correlated meaningfully with external community judgments (verdicts), suggesting the framework captures real conflict dynamics. Clustering analysis confirmed three distinct mode clusters (Connection, Control, Domination), with Protection functioning as a transitional state rather than a separate cluster.

Critical Discovery: Analysis of 1,294 posts where authors responded to negative feedback revealed that self-awareness—operationalized through "complexity markers" (ability to hold multiple perspectives)—significantly differentiated those who de-escalated (22.2%) from those who escalated toward Control/Domination (33.8%). De-escalators showed 78% higher complexity marker rates than escalators.

Conclusion: The Four-Mode Gradient can identify not only current regulatory state but, more importantly, capacity to return to Connection when challenged. This has implications for conflict de-escalation, relationship assessment, therapeutic intervention, and online moderation.

Keywords: emotional regulation, polyvagal theory, nervous system states, natural language processing, self-awareness, conflict escalation

1. Introduction

1.1 Background

Human behavior in conflict situations has traditionally been analyzed through personality traits, attachment styles, or diagnostic categories. However, emerging research in polyvagal theory (Porges, 2011) and interpersonal neurobiology (Siegel, 2012) suggests that behavior is better understood through the lens of nervous system states—momentary regulatory positions that shift based on perceived safety or threat.

The Four-Mode Gradient framework (Paretas-Artacho, 2026) proposes that human regulatory organization can be mapped onto four modes:

1. Connection — Ventral vagal activation; social engagement; capacity for empathy, repair, and collaboration
2. Protection — Sympathetic activation; threat response; withdrawal, hedging, or defensive positioning
3. Control — Cognitive elaboration of Protection; attempts to manage threat through rules, demands, and blame
4. Domination — Extreme cognitive elaboration; contempt, dehumanization, and willingness to cause harm

This framework synthesizes research from polyvagal theory (Porges), attachment theory (Bowlby, Ainsworth), trauma research (van der Kolk, Herman), contempt dynamics (Gottman), moral disengagement (Bandura), and power research (Keltner, Stark).

1.2 Research Questions

This study addresses three primary questions:

1. Can the four regulatory modes be detected in natural language using validated psychological constructs?
2. Do detected modes correlate with external judgments of behavior (community verdicts)?
3. What distinguishes individuals who escalate toward harm from those who return to Connection when challenged?

1.3 Theoretical Framework

The Four-Mode Gradient proposes that Protection is the biological foundation of threat response, while Control and Domination represent cognitive elaborations that occur when: (a) Protection alone fails to restore equilibrium, (b) the individual lacks self-awareness to return to Connection, and (c) the cognitive system overrides emotional-somatic feedback.

This suggests that the critical variable is not current mode but **capacity to move on the gradient**—specifically, capacity to return to Connection when challenged.

2. Methods

2.1 Design

Pre-registered observational study using natural language analysis (OSF: https://osf.io/f4x6y).

Critical Design Decision: Rather than imposing the four-mode structure on data, we operationalized each mode using established psychological constructs and tested whether predicted patterns emerged organically.

2.2 Data Source

Primary Dataset: Kaggle AITA Dataset (nird96)
Source: Reddit "Am I The Asshole" community
Sample: 10,000+ posts for mode detection; 11,670 posts for verdict analysis; 1,294 posts for escalation analysis

Advantages: Natural language describing real interpersonal conflicts; external judgment (community verdict) provides ground truth; edit sections allow analysis of response to challenge; large sample enables statistical analysis.

2.3 Detection Schema

Four modes operationalized with validated psychological constructs:

Table 1. Detection Schema

Mode	Theoretical Sources	Key Markers
Connection	LIWC affiliation; polyvagal safety cues (Porges); attachment linguistics	we/us/together; curiosity questions; acknowledgment; repair attempts
Protection	Threat-response research; anxiety markers; withdrawal patterns	hedging; minimizing; escape language; uncertainty markers
Control	Power discourse analysis; LIWC certainty; demand patterns	should/must; absolutes; blame frames; invalidation
Domination	Gottman contempt research; moral disengagement (Bandura)	contempt terms; dehumanization; identity attacks; dismissal

2.4 Self-Awareness Markers

To test the hypothesis that self-awareness determines escalation trajectory, we developed a marker dictionary based on metacognitive research:

Table 2. Self-Awareness Markers

Category	Function	Example Markers
Ownership	Self-responsibility	"I was wrong," "my mistake," "I should have"
Complexity	Multiple perspective-taking	"both," "I can see why," "from their perspective"

Deflection	Other-blame	"you don't understand," "I already explained"
Attack	Escalation	"you're wrong," "you people," "ridiculous"

2.5 Analysis Procedures

1. Mode Detection: Each post scored for presence of markers in all four categories; dominant mode assigned based on highest proportion
2. Clustering Analysis: K-means clustering (k=2 through k=6) on mode proportion vectors to test whether four clusters emerge naturally
3. Verdict Correlation: Chi-square analysis of mode distribution by community verdict
4. Escalation Analysis: Mode transition from original post to edit was classified as escalation (toward Control/Domination), de-escalation (toward Connection), or stable
5. Self-Awareness Comparison: Marker rates compared between escalators and de-escalators

3. Results

3.1 Mode Detection Confirmed

All four modes were successfully detected across 10,000+ posts.

Table 3. Mode Distribution (N=10,000+)

Mode	Frequency (Any Presence)	As Dominant Mode
Connection	36.5%	51.4%
Protection	22.0%	14.8%
Control	29.9%	29.7%
Domination	11.6%	4.1%

Interpretation: Connection dominated as expected—individuals describing conflicts tend to frame themselves favorably.

3.2 Modes Correlate with External Judgment

Table 4. Dominant Mode by Community Verdict (N=11,670)

Mode	YTA Posts	NTA Posts	Difference
Connection	53.3%	48.6%	+4.7%
Protection	16.4%	13.5%	+2.9%
Control	26.8%	33.1%	-6.3%
Domination	3.5%	4.8%	-1.3%

Psychological Interpretation: YTA posters use more Connection/Protection language (self-favorable framing, defensive justification). NTA posters describe more Control/Domination—they are reporting what was done to them. This confirms the schema captures real conflict dynamics, not random noise.

3.3 Clustering Analysis

Table 5. Clustering Silhouette Scores

k	Silhouette Score	Interpretation
2	0.341	Best statistical fit
3	0.307	Good

4	0.264	Acceptable (pre-registered)
5	0.266	Similar to k=4
6	0.239	Diminishing returns

Key Finding: Three modes cluster clearly (Connection, Control, Domination). Protection does not form a separate cluster—it co-occurs with other modes, suggesting it functions as a transitional state.

3.4 Escalation Analysis

Sample: 1,294 YTA posts containing edit sections (where author responds after receiving YTA verdict).

Table 6. Escalation Analysis (N=1,294)

Response Pattern	Count	Percentage
Escalated (toward Control/Domination)	437	33.8%
De-escalated (toward Connection)	287	22.2%
Same mode	570	44.0%

When challenged with negative feedback, approximately **one-third escalate** while **one-fifth de-escalate**.

3.5 Self-Awareness Markers Distinguish Trajectories

Table 7. Self-Awareness Marker Comparison

Marker Type	Escalators	De-escalators	Difference
Ownership	0.485	0.415	-0.07
Complexity	**0.121**	**0.216**	**+0.095 (78% higher)**
Deflection	0.048	0.042	-0.006
Attack	0.055	0.056	+0.001

Key Finding: Complexity markers—the ability to hold multiple perspectives ("I can see both sides")—significantly differentiate de-escalators from escalators. De-escalators showed **78% higher complexity marker rates** than escalators.

Notably, ownership markers were slightly higher in escalators, suggesting that surface-level acknowledgment ("I was wrong") without genuine complexity may be performative rather than indicative of actual self-awareness.

4. Discussion

4.1 Summary of Findings

1. Four modes detectable: The detection schema successfully identified all four regulatory modes in natural language using validated psychological constructs.

2. Modes correlate with external judgment: Mode classifications aligned meaningfully with community verdicts, confirming the schema captures real conflict dynamics.

3. Protection is transitional: Clustering analysis did not support Protection as a distinct category. Instead, Protection appears to function as a modifier or transitional state.

4. Self-awareness predicts trajectory: Complexity markers—operationalizing the capacity to hold multiple perspectives—significantly differentiated those who de-escalated from those who escalated when challenged.

4.2 The Self-Awareness Gate

The data support a model where self-awareness functions as a gate determining Protection's trajectory. When Protection is challenged, individuals with high self-awareness (high complexity markers) tend to return to Connection, while those with low self-awareness tend to elaborate into Control or Domination.

This has significant implications: **the critical variable is not current state but capacity to return to Connection when challenged.**

4.3 Practical Applications

Conflict de-escalation: Early detection of escalation trajectory through complexity marker analysis
Relationship assessment: Can this person receive feedback? Predicts relationship safety
Therapeutic intervention: Target self-awareness and perspective-taking capacity directly
Online moderation: Flag high-escalation linguistic patterns for intervention
Education: Develop complexity/perspective-taking as learnable skill

4.4 Limitations

1. **Single data source** — Results require replication across different platforms and contexts
2. **Self-report bias** — AITA posts represent author's framing; actual behavior may differ
3. **English language only** — Marker dictionaries validated for English; cross-cultural validity untested
4. **Cross-sectional** — Cannot establish causality; longitudinal studies needed
5. **Protection clustering** — Failure to find distinct Protection cluster may reflect measurement limitations

4.5 Future Research

• Cross-source replication (Reddit API access pending)
• Predictive validity: Can escalation trajectory be predicted from initial post before feedback?

- Somatic markers: Can disconnection from emotional-somatic feedback be detected in language?
- Intervention studies: Does increasing complexity/perspective-taking capacity reduce escalation?

5. Conclusion

This study provides initial empirical support for the Four-Mode Gradient framework. All four regulatory modes were detectable in natural language, and mode classifications correlated meaningfully with external judgments of behavior.

The central contribution is the identification of self-awareness—operationalized through complexity markers—as the key variable distinguishing those who escalate toward harm from those who return to Connection when challenged. This shifts focus from state detection to trajectory prediction: **the question is not "what mode are you in?" but "can you move back to Connection when challenged?"**

This has implications for understanding human conflict, designing interventions, and potentially identifying individuals at risk of causing harm before escalation occurs.

References

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193-209.

Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development.* Basic Books.

Gottman, J. M. (1994). *What predicts divorce? The relationship between marital processes and marital outcomes.* Lawrence Erlbaum Associates.

Herman, J. L. (1992). *Trauma and recovery: The aftermath of violence.* Basic Books.

Keltner, D. (2016). *The power paradox: How we gain and lose influence.* Penguin.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* University of Texas at Austin.

Porges, S. W. (2011). *The polyvagal theory: Neurophysiological foundations of emotions, attachment, communication, and self-regulation.* W. W. Norton.

Siegel, D. J. (2012). *The developing mind: How relationships and the brain interact to shape who we are* (2nd ed.). Guilford Press.

Stark, E. (2007). *Coercive control: How men entrap women in personal life.* Oxford University Press.

van der Kolk, B. A. (2014). *The body keeps the score: Brain, mind, and body in the healing of trauma.* Viking.

Author Note

This research was conducted independently without institutional affiliation or funding. The TEG-Blue framework was developed through synthesis of established psychological research, clinical observation, and systematic framework development supported by AI research tools.

The author declares no conflicts of interest.

Correspondence: research@teg-blue.org
Framework: https://teg-blue.com
Pre-registration: https://osf.io/f4x6y
ORCID: 0009-0005-2394-7162