Protocol for Coding Emergent Allegiance in Online Debates

Testing the Leviathan Hypothesis in Real Time

Author: Richard Dobson

Affiliation: Clara Futura World

Version: 1.2 (Plain-English revision, January 2026)

Keywords: social movements; emergence; threshold dynamics; costly allegiance; identity fusion; governance response; online discourse; protocol; falsifiability

1. Introduction

This document is a research protocol. Its purpose is to test whether the Leviathan Hypothesis can be turned into a set of observables, measurable indicators and then tested in live online debates.

The protocol builds on earlier theoretical work on "Leviathan-class" movements and participatory ontology. Where needed, this document summarizes that work rather than repeating it in full.

Scope clarification and dependent variable.

The Leviathan Hypothesis does not model social movement diffusion, adoption rates, or pathways of spread. It assumes that relational, brokered, and digitally mediated diffusion mechanisms are well established in the existing literature and treats them as background conditions rather than as the explanandum. The dependent variable analysed here is different: the transition from a movement that can be governed through ordinary means to a participatory system that becomes difficult to suppress and forces a change in governance response (e.g., containment, capture, or accommodation). The Leviathan Hypothesis therefore asks not how movements spread, but when and why some movements, regardless of diffusion pathway, internalize identity, obligation, and ultimacy to a degree that external governance shifts from routine management to exceptional measures. Critiques aimed at diffusion modelling apply only if diffusion is the outcome under study; in this framework it is not.

1.1 The problem: spread is not the same as impact

Most models of social movements focus on spread: how many people join, how fast ideas diffuse, how networks grow, and how long participants stay.

Those tools are powerful, but they do not always explain impact—why some movements reorganize institutions, reshape ethical norms, and redefine identity while others remain discussable subcultures or short-lived trends.

Protocol for Coding Emergent Allegiance in Online Debates

The Leviathan Hypothesis addresses this gap by asking a different question:

When does a movement become hard to govern from the outside because it has begun to govern from the inside—through identity, obligation, and moral ultimacy?

1.2 Key terms in plain English

Allegiance system

A shared pattern of participation—language, norms, rituals, identity claims—that coordinates behaviour across people who do not share a single central controller.

Leviathan-class

A case is classified as "Leviathan-class" only if it shows all three of the following:

Threshold behaviour: nonlinear escalation and persistence (not gradual growth)

Costly or risk-bearing allegiance in at least some participants

Governance response (containment, capture, or suppression) that follows rather than causes the escalation

Governance response

Any attempt by a platform, institution, or authority to manage the movement's effects: moderation, rule-setting, boundary policing, credential appeals, de-platforming, or formal consolidation.

1.3 Participatory ontology as a working lens

Participatory ontology is used here as a background lens: people do not merely describe social reality; they help produce it through participation.

In this protocol, participatory ontology does not function as an empirical result. It functions as a rationale for why identity-charged participation might generate stable, self-reinforcing patterns.

Important: Disagreement with participatory ontology is never coded as evidence for Leviathan dynamics. Only pre-specified behavioral indicators count (see Appendix A).

1.4 Why prospective testing is needed

Most Leviathan claims are made after the fact, using famous historical cases like early Christianity, the French Revolution, or the civil rights movement. Retrospective work is valuable, but it is vulnerable to two problems:

Selective evidence: choosing only cases that fit

Post-hoc fitting adjusting the theory to match what already happened

Protocol for Coding Emergent Allegiance in Online Debates

Online debates provide a complementary testing ground:

1. They unfold in real time
2. They leave complete textual traces
3. They allow controlled variation (prompt type, moderation intensity, author presence)

This makes them suitable for prospective tests of whether indicator patterns cluster where the hypothesis predicts.

1.5 Research contribution

This protocol contributes three items intended for replication:

1. An operational definition of Leviathan-class dynamics
2. A comment-level and event-level codebook with reliability procedures
3. A falsifiability checklist and comparative stress-test design

1.6 Method overview and roadmap

This is a prospective, mixed-methods protocol designed to be replicated across sites and topics.

The core idea:

Instead of looking backward at famous historical cases, we run controlled tests in live online debates, code what happens using clear indicators, and check whether the predicted patterns show up.

Three phases:

Phase 0 – Pilot and calibration

Code a small set of existing threads to refine the codebook, train coders, and check that two independent coders agree on how to score the same comment.

Phase 1 – Live testing with controlled variation

Post four types of prompts (high-identity, control, meta) across selected sites, assign different moderation conditions at random, capture all comments and governance actions, and code them using the frozen codebook.

Phase 2 – Comparative stress tests

Apply the same protocol to non-religious domains (politics, fandom, wellness) to see whether the indicator patterns are general or topic-specific, and to identify clear negative cases where the model should say "no, not Leviathan-class."

Protocol for Coding Emergent Allegiance in Online Debates

How this document is organized:

- Section 2 states the research questions and pre-registered hypotheses
- Section 3 details the design, sites, prompts, coding layers, moderation conditions, and procedure
- Section 4 specifies the analysis plan (quantitative tests, qualitative case narratives, comparative tests)
- Section 5 explains the relationship to Leviathan v2 and participatory ontology
- Sections 6–8 discuss expected contributions, limitations, and conclusions
- Appendix A provides the full codebook: every indicator, every scale point, with the anti-self-sealing safeguard

For replication:

Start with Appendix A to understand what gets coded, then follow the procedure in Section 3.7 step-by-step.

2. Research questions and hypotheses

This section states the research questions and the hypotheses to be tested. The hypotheses are derived from the Leviathan v2 specification.

2.1 Core research question

- RQ1: In live online debates, do high-identity topics reliably produce a repeatable cluster of Leviathan indicators (threshold escalation, identity fusion, governance strain) compared with technical control topics?

2.2 Subsidiary questions

- RQ2: Which indicators are sensitive to moderation (avoidable breakdowns) and which persist across moderation conditions (structural dynamics)?
- RQ3: Do similar indicator patterns appear across domains (religion, politics, fandom, wellness), or are they topic-specific?
- RQ4: Does participatory language (e.g., "participation," "field," "in-Christ" language) correlate with escalation or de-escalation patterns?

2.3 Hypotheses (pre-registered targets)

- H1 (Indicator clustering): High-identity prompts will show higher rates of identity fusion, ad hominem, declared exit-and-return, and governance actions than technical control prompts.
- H2 (Symbolic gravity): Participants with higher identity fusion scores will show higher re-entry rates (returning after declaring exit) and stronger resistance to falsifiability prompts.

- H3 (Governance escalation): High-identity threads will show more governance responses (moderation actions, boundary policing, credential appeals) per comment than control threads.
- H4 (Avoidable vs structural): Active moderation will reduce avoidable behaviors (personal insults, motive speculation) more than it reduces structural dynamics (boundary disputes, identity stake disclosure, repeated re-entry).
- H5 (Model comparison): If the above effects disappear after controlling for standard predictors (topic popularity, comment volume, prior network ties), Leviathan adds little; if a remainder persists, Leviathan retains explanatory value.

3. Methods

This study uses a mixed-methods, multi-site design with prospective observation and controlled variation.

3.1 Design overview

Phase 0 (pilot calibration):

Retrospectively code a small corpus of existing debate threads to refine the codebook, train coders, and test reliability.

Phase 1 (live threads):

Run four standard prompt types across selected online sites with randomized moderation conditions. Observe and code all comments and governance actions over a fixed window.

Phase 2 (comparative stress tests):

Apply the same protocol to non-religious high-identity domains (politics, fandom, wellness) to test generality and to identify negative cases.

3.2 Sites and participants

Sites:

Public discussion spaces where participants already debate contested topics (e.g., forums, Reddit, Academia.edu, platform comment sections).

Recruitment:

Naturalistic—the study does not solicit private data and does not require participants to disclose identity beyond what they choose to post publicly.

Protocol for Coding Emergent Allegiance in Online Debates

Anonymization:

In all reporting, participants are anonymized (Participant 01, Participant 02, etc.). Direct quotes are minimized, and any quoted text is paraphrased unless explicit permission is granted.

3.3 Prompts (four levels of identity load)

Each thread is initiated with one of four prompt types:

Type A (Leviathan prompt, high identity):

Asks whether a movement became "uncontainable" and what evidence would demonstrate that.

Type B (Lineage prompt, high identity):

Asks whether a hidden or suppressed lineage preserves the "true" teaching and what would count as evidence.

Type C (Technical control prompt, low identity):

Asks a narrow historiographical question that requires sources but does not usually demand existential allegiance (e.g., "What is the earliest attestation of X in the manuscript record?").

Type D (Meta prompt, stress test):

Asks when debates shift from evidence to identity and whether that shift has recognizable markers.

Each thread runs for a fixed observation window (e.g., 28 days).

3.4 Coding and measures

Comments are coded on two layers (see Appendix A for full definitions):

1. Comment-level indicators (scored 0–2):

1. Evidence posture (E)
2. Falsifiability posture (F)
3. Identity charge (I)
4. Governance / suppression talk (G)
5. Ad hominem / motive-talk (A)
6. Re-entry after declared exit (R)
7. Participatory vocabulary (P)

Protocol for Coding Emergent Allegiance in Online Debates

2. Event-level escalation markers (present/absent, timestamped):

- T1: Boundary challenge
- T2: Motive attribution
- T3: Self-sealing move
- T4: Identity disclosure
- T5: Identity assignment
- T6: Playful boundary disruption
- T7: Contempt threshold
- T8: Closure attempt
- T9: Counter-identity rejection
- T10: Totalization

Anti-self-sealing safeguard:

Disagreement with the model (or with the author) is never coded as evidence for Leviathan dynamics. Only the pre-specified behavioral indicators are coded.

3.5 Reliability and quality control

At least two independent coders score each thread

Disagreements are resolved by discussion

Reliability is reported (e.g., Cohen's kappa for categorical codes)

The codebook is revised only during Phase 0; Phase 1 uses a frozen codebook

3.6 Moderation conditions

Threads are randomly assigned to one of three moderation conditions:

Condition 1 (minimal):

Intervene only for clear policy violations (threats, doxxing).

Condition 2 (active facilitation):

Redirect participants toward evidence and away from motive-talk and insults.

Condition 3 (structured dialogue):

Require each participant to state what would change their mind and remove ad hominem quickly.

This variation tests whether the indicator pattern is mainly a product of bad facilitation or a more structural feature of high-identity topics.

3.7 Procedure (replicable steps)

- Select site and prompt type
- Post the prompt with a clear observation window and assigned moderation condition
- Capture all comments, timestamps, and moderation actions during the window
- Code comments daily to reduce hindsight bias
- At the end of the window, freeze the dataset and run quantitative comparisons and qualitative case summaries
- Repeat across prompt types and sites; then repeat in Phase 2 domains

3.8 Ethics

- This protocol uses public discourse only.
- The study avoids collecting private information
- The study avoids doxxing
- Participant handles are anonymized in all reporting
- If a platform's terms of service prohibit scraping, data collection is manual and limited
- Any request by a participant to be excluded is honoured for future reporting

4. Analysis plan

The analysis tests whether the predicted indicator pattern clusters where the hypothesis says it should, and whether moderation changes that pattern.

4.1 Quantitative tests (pre-specified comparisons)

Primary test (H1/H3):

Compare high-identity threads (Types A/B) with control threads (Type C) on rates of identity charge, ad hominem, exit-and-return, and governance actions per comment.

Moderation test (H4):

Compare Conditions 1–3 to see which indicators drop with good facilitation and which persist.

Participant-level test (H2):

Model whether identity charge predicts re-entry and resistance to falsifiability prompts.

Sequence test:

Identify whether escalation markers occur in a repeatable order and whether that order differs by topic and moderation.

Protocol for Coding Emergent Allegiance in Online Debates

4.2 Qualitative analysis

Construct short case narratives of representative threads:

1. What claim triggered escalation?
2. How did identity enter the exchange?
3. What governance responses followed?

The qualitative goal is to check whether quantitative indicator clusters map onto coherent interaction patterns rather than artifacts of coding.

4.3 Comparative stress tests (Phase 2)

Apply the same protocol to domains that can produce high identity without theology (e.g., political ideology, fandom canon disputes, wellness controversies).

A strong Leviathan model should be able to say both:

(a) Why some high-intensity cases qualify

(b) Why some do not (negative cases)

5. Relation to Leviathan v2 and participatory ontology

This protocol operationalizes the Leviathan v2 claim that some movements become difficult to manage externally because they internalize obligation and identity.

Participatory ontology is used here as a motivation for why participation may have causal relevance at the social level.

The empirical claims of this protocol do not require readers to endorse the ontology; they require only that the indicators be observable and that predicted patterns can be confirmed or disconfirmed.

6. Expected contributions

If successful, this protocol will deliver:

1. A reusable codebook and scoring sheet
2. A set of negative and positive cases
3. A clearer boundary between Leviathan-class dynamics and ordinary online conflict

If unsuccessful, it will clarify which parts of the Leviathan model are rhetorical rather than testable and should be revised or abandoned.

Protocol for Coding Emergent Allegiance in Online Debates

7. Limitations and future directions

Limitations:

1. Online discourse is not the same as offline history
2. Results will be limited to the tested settings and topics
3. Platform design (algorithms, moderation norms) may shape outcomes

Future directions:

1. Test whether the same indicators predict offline mobilization and institutional change
2. Apply the protocol to historical text corpora where textual traces allow similar coding
3. Develop longitudinal tracking to see whether indicator patterns in online debates predict later governance escalation in offline contexts

8. Conclusion

The Leviathan Hypothesis is only useful if it can be tested against alternatives.

This protocol provides a way to do that in real time, using:

1. Observable indicators
2. Moderation variation
3. Explicit disconfirmation criteria

The intended outcome is not a protected theory, but a clearer map of when symbolic allegiance becomes hard to govern—and when it does not.

Appendix A. Codebook summary and scoring sheet

This appendix provides operational definitions for the indicators used in the protocol. It is designed so that independent reviewers can apply the same codes and contest the results.

A1. Comment-level indicators (score 0–2 unless noted)

Evidence posture (E):

- E0 = assertion without evidence
- E1 = vague evidence ("scholars say…") without a checkable source
- E2 = specific checkable source with accurate paraphrase or quotation

Falsifiability posture (F):

- F0 = non-negotiable posture ("nothing could change my mind")
- F1 = conditional openness ("if X were shown…")

Protocol for Coding Emergent Allegiance in Online Debates

- F2 = explicit statement of disconfirming evidence that would change the author's view

Identity charge (I):

- I0 = detached, analytic tone
- I1 = personal investment but not fused
- I2 = fusion (critique is treated as betrayal/attack; purity language; intense self-protective framing)

Governance / suppression talk (G):

G0 = none

- G1 = descriptive references to power, councils, institutions, censorship
- G2 = suppression story used to protect the claim ("they destroyed the evidence," "all scholars are corrupt")

Ad hominem / motive-talk (A):

- A0 = none
- A1 = mild personal dismissal
- A2 = strong personal attack or psychological diagnosis used as argument substitute

Re-entry after declared exit (R):

- R0 = no exit language
- R1 = continued participation without exit claim
- R2 = explicit exit ("I'm done") followed by return and renewed engagement

Participatory vocabulary (P):

- P0 = none
- P1 = mentions participation/field language descriptively
- P2 = substantively uses the vocabulary to frame claims or revise a view
- A2. Event-level escalation markers (present/absent; timestamped)
- T1 Boundary challenge: legitimacy of the model or community is challenged
- T2 Motive attribution: the argument shifts toward diagnosing the person's motives
- T3 Self-sealing move: resistance is reframed as proof (flag carefully; see safeguard)
- T4 Identity disclosure: participant states a personal or existential stake
- T5 Identity assignment: participant assigns a hidden identity to another ("you're really X")
- T6 Playful boundary disruption: satire/clowning that changes the frame
- T7 Contempt threshold: insult that changes the temperature of the thread
- T8 Closure attempt: unilateral attempt to end discussion on one party's terms
- T9 Counter-identity rejection: explicit refusal of assigned identity
- T10 Totalization: the conflict itself is treated as the main evidence base

- A3. Anti-self-sealing safeguard (mandatory rule)

Disagreement with the theory is never evidence for the theory.

T3 and T10 are scored only when a participant explicitly claims that the other party's resistance is itself evidence, and only when that move co-occurs with other indicators (identity charge, governance strain, re-entry) in the pre-specified pattern.

End of document