

Toward a Framework for Cultural Immunology: Computational Approaches to Understanding Societal Resilience

A research programme for measuring how societies maintain shared meaning under stress: formally defined text-based indicators, a validation design with known ground truth, and an ethical governance framework.

Saket Poswal

Author

Sep 13, 2025 • Updated: Sep 29, 2026

ABSTRACT

Societies differ in how well they maintain shared meaning under stress, whether that stress comes from economic crisis, political conflict or the reorganisation of public discourse by digital platforms. This paper proposes Cultural Immunology as a research programme for studying that capacity with computational methods. The immune metaphor has a scholarly lineage, in psychological inoculation against persuasion and misinformation, the behavioural immune system, and epidemiological models of cultural transmission, but existing work acts mainly on individuals. Cultural Immunology asks a population-level question: can stress and resilience in a society's shared meanings be measured from the texts it produces, and do such measures anticipate outcomes such as declining institutional trust? The paper makes three contributions. First, it defines three indicators formally: a Semantic Coherence Index based on aligned group-specific word embeddings, a Discourse Separability Index based on how easily a speaker's group can be inferred from their language, and an Institutional Trust Trajectory combining survey and text measures. Second, it sets out a validation design anchored to a known result, the sharp rise in partisan divergence in U.S. congressional speech from the early 1990s, before extension to Indian parliamentary debates. Third, it proposes an ethical and governance framework, grounded in the risks the programme itself creates, including authoritarian capture and technocratic simplification. The paper is a programme proposal: the indicators have not yet been computed, and it states the conditions under which the approach should be judged to have failed.

Author

Saket Poswal • [ORCID 0009-0009-8574-9953](#) • [Google Scholar](#)

Type	Analysis (working paper, not peer reviewed)
Published	13 September 2025
Last revised	29 September 2026
Method	Research programme design: formal definition of text-based indicators using aligned word embeddings and group-separability estimators; validation design with known ground truth; agent-based modelling of cultural transmission; ethical and governance analysis
License	CC BY 4.0
DOI	10.5281/zenodo.23011016
Keywords	cultural immunology · computational social science · cultural analytics · societal resilience · semantic coherence · word embeddings · discourse polarisation · institutional trust · cultural evolution · inoculation theory · research ethics · data governance

Working paper · Version 2.0 (29 September 2026)

Revision note. Version 2.0 substantially revises the September 2025 proposal. The research questions and the ethical framework are retained. The revision situates the programme in prior work it previously omitted (inoculation theory, the behavioural immune system, cultural evolution and established models of cultural transmission); replaces informally described indicators with formal definitions based on established embedding and group-separability methods; replaces a non-falsifiable narrative-diversity measure; adds a validation design anchored to a known empirical result; and states failure conditions. The author line has been corrected.

1. Introduction

Some societies absorb shocks such as recessions, contested elections, pandemics and rapid technological change while keeping enough shared meaning to argue productively about what to do. Others fragment. Anderson (1983/2006) described nations as “imagined communities” held together by shared narratives; Geertz (1973) described culture as “webs of significance” that people spin and inhabit. If shared meaning is what allows collective life to continue through disagreement, then how it holds up under stress is a question of practical as well as scholarly importance.

The digitisation of public discourse makes a new kind of evidence available. Books, newspapers, parliamentary debates and social media leave textual traces of shared meaning at a scale no ethnographer could read, and computational social science has developed methods for analysing them

(Lazer et al., 2009; Michel et al., 2011; DiMaggio et al., 2013; Manovich, 2020). Word embeddings, in particular, have been shown to recover culturally meaningful structure: dimensions of social class across a century of books (Kozlowski et al., 2019) and the evolution of gender and ethnic stereotypes tracked against census data (Garg et al., 2018).

This paper proposes Cultural Immunology as a research programme that brings these methods to bear on one question: can stress and resilience in a society's shared meanings be measured from its texts, and do such measures anticipate social outcomes? It contributes:

1. A positioning of the immune metaphor within existing scholarship, clarifying what is and is not new (Section 2).
2. Three formally defined indicators built on established methods (Section 3).
3. A validation design that tests the indicators against a known result before applying them to new settings, with explicit failure conditions (Section 4).
4. An ethical and governance framework addressing risks the programme itself creates (Section 6).

It is a proposal. No indicator has yet been computed, and the value of the programme depends on whether the validation in Section 4 succeeds.

2. Related Work: The Immune Metaphor and Its Lineage

2.1 Inoculation against persuasion

The idea that minds can be “immunised” has a long history in social psychology. McGuire's (1964) inoculation theory proposed that exposure to weakened forms of a persuasive attack, together with refutation, builds resistance to later, stronger attacks. The approach has been revived for misinformation: in six preregistered experiments and a field study on YouTube with more than 22,000 participants, short videos explaining manipulation techniques improved people's ability to recognise them and to discern trustworthy from untrustworthy content (Roozenbeek et al., 2022).

2.2 The behavioural immune system

Schaller and Park (2011) describe a *behavioural immune system*: psychological mechanisms that detect cues of infection and trigger avoidance. Because it responds to superficial cues, it can generate prejudice toward people who pose no real threat. This literature is a warning as much as an inspiration. Defensive responses to perceived cultural “threats” can themselves be harmful, and a programme that labels some cultural change as a threat inherits that risk.

2.3 Cultural transmission and evolution

Sperber (1996) proposed an *epidemiology of representations*, explaining culture by the causes and effects of the spread of ideas through populations. Boyd and Richerson (1985) developed formal models of

cultural evolution using the methods of population biology. Agent-based models show how simple transmission rules produce population-level structure: in Axelrod's (1997) model, agents who become more similar through local interaction can produce stable global polarisation, and bounded-confidence models show how the willingness to be influenced only by similar others yields consensus, polarisation or fragmentation depending on a single parameter (Hegselmann & Krause, 2002). Centola (2018) showed that behaviours requiring reinforcement from several contacts spread differently from simple contagions.

2.4 Social resilience

Adger (2000) defined social resilience as the ability of groups or communities to cope with external stresses and disturbances arising from social, political and environmental change, drawing an analogy with ecological resilience.

2.5 What is new here

Existing immune-style work acts mainly on individuals (inoculation) or explains mechanisms of transmission (cultural evolution). Cultural Immunology asks a population-level measurement question: can the state of a society's shared meanings, whether cohering, fragmenting or recovering, be tracked from its texts, and are those measures informative about outcomes? The metaphor is used with its limits stated (Section 2.6); what matters is whether the measures work.

2.6 Limits of the metaphor

Cultural change is often desirable; "health" and "pathology" are contested, normative categories in culture; diversity and conflict are frequently signs of vitality rather than disease; and, as the behavioural immune system shows, defensive responses can target the wrong things. Throughout this paper, *stress* means measurable strain on shared meaning, not a judgement that any particular change is bad.

3. Indicators

Each indicator is defined for a society partitioned into groups G (for example, political parties, regions or language communities) observed over time periods t .

3.1 Semantic Coherence Index (SCI)

Question: do different groups use key concepts to mean the same thing?

Construction.

1. For each group g and period t , train word embeddings on that group's corpus.

2. Align each group's embedding space to a reference space using orthogonal Procrustes on a set of high-frequency anchor words assumed to be stable, as in diachronic semantic change research (Hamilton et al., 2016).
3. For a pre-registered set C of civic concepts (for example, *freedom*, *justice*, *nation*, *development*, *rights*), compute the cosine similarity of each concept's vectors across every pair of groups.

Definition. SCI_t is the mean, over concepts c in C and group pairs (g, h) , of $\cos(v_{c,g,t}, v_{c,h,t})$, after alignment.

Complementary measure. Following Kozlowski et al. (2019), each concept can also be projected onto interpretable cultural dimensions (for example, a moral dimension defined by antonym pairs such as *good–evil*), and group differences on each dimension reported. This shows *how* meanings diverge, not only *that* they do.

Known risk. Embedding similarities are unstable across training runs and corpus samples (Antoniak & Mimno, 2018). SCI must therefore be reported as a mean with bootstrap confidence intervals over multiple training seeds and corpus resamples.

3.2 Discourse Separability Index (DSI)

Question: how easily can one tell which group a speaker belongs to from what they say?

The first version of this paper proposed a Narrative Diversity Score whose interpretation (“very low or very high scores may indicate stress”) could not be falsified. It is replaced by an established measure.

Gentzkow, Shapiro and Taddy (2019) defined partisanship in speech as the ease with which an observer could infer a speaker's party from a single utterance, and developed an estimator that corrects the severe finite-sample bias of naive approaches. Applied to U.S. congressional speech from 1873 to 2016, their measure rose sharply from the early 1990s.

Definition. DSI_t is the expected posterior probability, under the Gentzkow–Shapiro–Taddy estimator, that a neutral observer correctly identifies a speaker's group from one utterance in period t . It generalises from parties to any partition of speakers.

SCI and DSI capture different things. DSI measures divergence in *what is said* (word and phrase choice); SCI measures divergence in *what shared words mean*. A society could show high DSI with high SCI (different topics, shared meanings) or low DSI with low SCI (same words, different meanings).

3.3 Institutional Trust Trajectory (ITT)

Question: is trust in institutions rising or falling, and how fast?

Definition. ITT_t is the slope of a local regression of institutional trust on time over a moving window, estimated from (a) representative surveys where available and (b) a text-based measure of expressed trust, validated against (a) in periods where both exist.

3.4 Narrative structure (exploratory)

Computational narratology can measure structural features of narratives at scale (Boyd et al., 2020). Whether narrative structure in public discourse carries information about social stress is an open question; it is included as an exploratory measure without a pre-specified hypothesis.

4. Validation Design

A new indicator is credible only if it recovers something already known. The programme therefore proceeds in three stages, each with a stated failure condition.

4.1 Stage 1: known ground truth (U.S. Congress)

Data: U.S. congressional speech, for which Gentzkow et al. (2019) documented a sharp rise in partisan divergence beginning in the early 1990s.

Test: compute SCI for civic concepts across Democratic and Republican speech, by congressional session.

Prediction: SCI should decline over the same period in which DSI rises, and the decline should exceed variation across seeds and resamples.

Failure condition: if SCI shows no change beyond its confidence interval during a period of well-documented polarisation, SCI does not measure what it claims to, and the programme should revise or abandon it before proceeding.

4.2 Stage 2: convergent validity with surveys

Test: correlate text-based measures of trust and semantic divergence with survey measures of institutional trust and affective polarisation for the same populations and periods.

Failure condition: no reliable association with survey measures after accounting for measurement error.

4.3 Stage 3: extension to India

Data: Lok Sabha debates, which are publicly available across several decades and in multiple languages, together with English- and Hindi-language newspaper corpora.

Test: compute SCI and DSI across parties and regions, and examine whether changes anticipate changes in survey-measured institutional trust.

Additional requirement: multilingual alignment and translation choices introduce further measurement error; results must be reported separately by language before any cross-language aggregation.

4.4 Simulation layer

Agent-based models in the tradition of Axelrod (1997) and Hegselmann and Krause (2002) can generate synthetic discourse with known levels of fragmentation. Applying the indicators to simulated populations whose true state is known provides a further check: an indicator that cannot recover known fragmentation in simulation should not be trusted on real data.

5. Research Questions and Hypotheses

- RQ1. Do SCI and DSI measure distinct aspects of cultural divergence? *H1: SCI and DSI are correlated but not redundant; each explains variance in trust outcomes not explained by the other.*
 - RQ2. Do declines in semantic coherence precede declines in institutional trust? *H2: Changes in SCI lead changes in survey-measured trust, controlling for economic conditions.*
 - RQ3. Do shocks leave lasting traces? *H3: After major shocks (for example, the 2008 financial crisis), SCI falls and then partially recovers; the speed of recovery varies across societies and is associated with pre-shock institutional trust.*
 - RQ4. Does platform design matter? *H4: Discourse on platforms with engagement-ranked feeds shows higher DSI than comparable discourse on chronologically ordered platforms.* This must be tested with care, since the relationship between platform exposure and polarisation is not straightforward: exposure to opposing views can increase rather than reduce polarisation (Bail, 2021).
-

6. Ethics and Governance

Measurement of culture at scale is dual-use. Tools that reveal fragmentation could help communities understand themselves; the same tools could help states or firms monitor, target or manipulate populations (Zuboff, 2019). Scott (1998) documents how states that make societies “legible” through simplified measurement have repeatedly caused harm. These risks are the reason to specify governance before building anything.

6.1 Principles

- Cultural sovereignty. Communities define their own cultural goals; no indicator defines cultural “health”.
- Democratic legitimacy. Uses of the research are governed with the participation of affected communities, with rights to contest findings.

- Cognitive liberty. No application may target individuals or be used to manipulate beliefs.
- Beneficence with humility. Following principles for trustworthy AI, including beneficence, non-maleficence, autonomy, justice and explicability (Floridi et al., 2018), results are presented as uncertain measurements, not diagnoses.

6.2 Safeguards

- Aggregate only. Indicators are computed for groups and periods, never for individuals; no individual-level profiling.
- Public and open. Methods, code and aggregate results are open, so that claims can be checked and misuse is visible.
- No intervention layer. The first version proposed a “Cultural Strengthening Recommendation System”. It is withdrawn from the programme: the project measures and reports, and any decision to act rests with communities and democratic institutions.
- Independent oversight. An oversight board including community representatives, social scientists, ethicists and human-rights experts can halt any application.

6.3 Risks

Risk	Example	Mitigation
Authoritarian capture	Using DSI to identify dissenting regions for surveillance	Aggregate-only design; public methods; oversight with power to halt
Technocratic simplification	Treating SCI as a target for policy	State that indicators are diagnostic, not targets; report uncertainty
Pathologising difference	Labelling minority meanings as “incoherence”	Report divergence descriptively; never rank groups
Measurement error presented as fact	Unstable embeddings producing spurious trends	Bootstrap intervals; multi-seed estimation; validation gates

7. Limitations

- No results yet. The indicators are defined but not computed; their validity is an empirical question.
- Texts are not culture. Written, public discourse over-represents elites, literate groups and dominant languages; oral and non-textual culture is largely invisible to these methods.
- Correlation, not causation. Even if indicators track trust, observational designs cannot establish that semantic change causes social outcomes.
- Measurement instability. Embedding-based measures are sensitive to corpus size and training choices (Antoniak & Mimno, 2018); small corpora may not support reliable estimates.
- Normative risk. Any measure of “coherence” can be read as a value judgement in favour of conformity. The programme tries to resist that reading, but it cannot fully control how results are used.

8. Conclusion

Cultural Immunology is proposed not as a system for managing culture but as a disciplined attempt to measure one of its properties: how well shared meaning holds under stress. The programme stands or falls on validation. Its central indicator must first recover a well-documented case of polarisation before it is applied anywhere new, and it should be abandoned if it cannot. If it can, the result would be a transparent, open, aggregate-only instrument that helps societies see changes in their own shared meanings, and a governance framework designed from the start to keep that instrument from becoming a tool of control.

References

- Adger, W. N. (2000). Social and ecological resilience: Are they related? *Progress in Human Geography*, 24(3), 347–364.
- Anderson, B. (2006). *Imagined communities: Reflections on the origin and spread of nationalism* (Revised ed.). Verso. (Original work published 1983)
- Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6, 107–119.
- Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2), 203–226.
- Bail, C. A. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.
- Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.
- Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32), eaba2196.
- Centola, D. (2018). *How behavior spreads: The science of complex contagions*. Princeton University Press.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606.

- Floridi, L., Cows, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644.
- Geertz, C. (1973). *The interpretation of cultures*. Basic Books.
- Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4), 1307–1340.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1489–1501).
- Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).
- Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5), 905–949.
- Lazer, D., Pentland, A., Adamic, L., et al. (2009). Life in the network: The coming age of computational social science. *Science*, 323(5915), 721–723.
- Manovich, L. (2020). *Cultural analytics*. MIT Press.
- McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, 1, 191–229.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34), eabo6254.
- Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, 20(2), 99–103.
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.
- Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Blackwell.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.