I. What Constitutes a National Threat (Expanded Definition)

A post (or pattern of posts) is treated as a national threat when it satisfies at least one of the following:

Threatens life at scale

Threatens state authority or territorial integrity

Threatens critical systems needed for civilian survival

Threatens social cohesion at a mass level

Threatens economic or financial stability

Enables hostile foreign or proxy operations

Transitions from speech → coordination → execution

II. Expanded Threat Taxonomy (Ranked)

🟥 Tier 1 — Existential / Catastrophic Threats (Highest Severity)

1. Mass Casualty Violence Advocacy

Severity: Extreme

Risk Weight: 95–100

Includes

Calls for genocide, ethnic cleansing

Advocacy of mass murder

Glorification of large-scale killings with encouragement

Post Signals

“Wipe them out”

“They all deserve to die”

Dehumanization + violence

Why Critical

Historically precedes mass atrocities

Rapid offline translation

2. Terrorism (Direct or Indirect Support)

Severity: Extreme

Risk Weight: 90–100

Includes

Recruitment messaging

Praise of terrorist attacks

Ideological justification

Logistics or fundraising hints

Signals

Martyr narratives

Symbols, slogans, manifestos

“Join the struggle”, “support the cause”

3. Critical Infrastructure Sabotage

Severity: Extreme

Risk Weight: 90–100

Includes

Power grids

Water systems

Hospitals

Telecoms

Transport hubs

Signals

Naming infrastructure

Discussing vulnerabilities

Encouraging disruption

Why Critical

One act can affect millions

Cascading failures

🟧 Tier 2 — Severe National Stability Threats

4. Coordinated Insurrection or Armed Rebellion

Severity: Very High

Risk Weight: 80–95

Includes

Armed uprising calls

Overthrow of government

Seizure of state facilities

Signals

“We take the streets”

“This regime falls”

Organized rhetoric + timing

5. Election Interference & Democratic Subversion

Severity: Very High

Risk Weight: 75–90

Includes

False election results

Calls to reject lawful outcomes

Delegitimization of voting systems

Signals

“The election is fake”

“Do not accept the results”

Fake announcements

6. Targeted Threats to State Officials

Severity: Very High

Risk Weight: 75–90

Includes

Threats against presidents, MPs, judges, military

Doxxing + hostility

Signals

Naming individuals

Fixation + escalation

“They will pay”

🟨 Tier 3 — High-Risk Societal Destabilization

7. Ethnic / Religious Mobilization Toward Violence

Severity: High

Risk Weight: 65–80

Includes

Collective blame

Calls for exclusion or punishment

Framing groups as existential threats

Signals

“They are invading us”

“They control everything”

Dehumanization narratives

8. Large-Scale Disinformation Campaigns

Severity: High

Risk Weight: 60–80

Includes

Fake emergencies

False coup announcements

Pandemic or security misinformation

Signals

“Breaking: Army has taken over”

Fabricated government statements

9. Economic Warfare & Financial Destabilization

Severity: High

Risk Weight: 60–75

Includes

Bank run incitement

Market manipulation narratives

False collapse claims

Signals

“Withdraw all your money now”

“Banks will close tomorrow”

🟦 Tier 4 — Emerging / Latent Threats

10. Radicalization Pipelines

Severity: Medium

Risk Weight: 45–65

Includes

Ideological grooming

Extremist framing without violence yet

Signals

Absolutist ideology

Glorification of violence “in theory”

11. Coordinated Hate Networks

Severity: Medium

Risk Weight: 40–60

Includes

Harassment campaigns

Narrative amplification

Signals

Repetitive messaging

Shared slogans

Networked behavior

12. Foreign Influence & Proxy Narratives

Severity: Medium

Risk Weight: 40–65

Includes

Foreign strategic messaging

Narrative laundering

Signals

Synchrony across accounts

Language mirroring foreign state lines

🟩 Tier 5 — Non-Threat / Protected Expression

13. Political Criticism

Severity: None

Risk Weight: 0–20

14. Satire, Art, Protest Speech

Severity: None

Risk Weight: 0–15

III. Risk Scoring (Detailed)

Composite Risk Formula (Conceptual)

Risk =

0.30 * Intent

+ 0.20 * Capability

+ 0.15 * Specificity

+ 0.15 * Reach

+ 0.10 * Trajectory

+ 0.10 * Network Density

Dimension Breakdown

DimensionWhat It MeasuresIntentOpinion → Justification → MobilizationCapabilityHistory, access, consistencySpecificityNamed targets, timingReachViews, shares, influencersTrajectoryEscalation speedNetwork DensityCoordinated amplificationIV. Escalation Ladder (Expanded)

LevelTriggerAction0Risk < 20Ignore120–40Trend monitoring240–60Human review360–75Network analysis475–90Active intervention590+Immediate responseV. Key Principle (Very Important)

National threats are not defined by ideology, criticism, or anger — they are defined by mobilization capacity + scale + trajectory.

A single angry post is noise.

A networked, escalating narrative is signal.

Below is a dedicated, detailed extension of the taxonomy focused on economic dissatisfaction and social dissatisfaction, treated the way national-security and strategic-risk units actually treat them: not as threats by default, but as pressure fields that can transition into threats under specific conditions.

This distinction is critical.

I. Core Principle (Read This First)

Economic and social dissatisfaction are not national threats by themselves.

They are structural risk amplifiers.

They become national-security concerns only when they:

Mobilize at scale

Target institutions or groups

Enable coercion, violence, or systemic disruption

Are exploited by organized actors (domestic or foreign)

II. Economic Dissatisfaction — Threat Classification

Category E0 — Legitimate Economic Grievance (Non-Threat)

Severity: None

Risk Weight: 0–20

Includes

Complaints about cost of living

Unemployment frustration

Wage stagnation

Tax burden criticism

Examples

“Life is too expensive.”

“There are no jobs.”

“The economy is failing us.”

Treatment

Ignored by security systems

Relevant only to policy analysis

Category E1 — Economic Anger with Delegitimization

Severity: Low–Moderate

Risk Weight: 25–40

Includes

Framing the state as incompetent or corrupt

Broad blame narratives

Loss of institutional trust

Signals

“Government has failed completely.”

“This system is broken beyond repair.”

Why It Matters

Early erosion of legitimacy

Creates fertile ground for radical narratives

Category E2 — Economic Mobilization Pressure

Severity: Moderate

Risk Weight: 40–60

Includes

Calls for mass action tied to economic issues

Strikes, shutdowns, boycotts

Economic protest coordination

Signals

“Everyone stop working on Monday.”

“Do not pay taxes.”

“Shut the city down.”

Key Distinction

Peaceful protest ≠ threat

Systemic disruption ≠ violence, but still monitored

Category E3 — Economic Destabilization Narratives

Severity: High

Risk Weight: 60–75

Includes

Bank-run encouragement

Market panic

False scarcity narratives

Signals

“Withdraw all your money now.”

“Fuel will run out tomorrow.”

“Banks are collapsing.”

Why It Escalates

Can crash systems without physical violence

Often disinformation-driven

Category E4 — Economic Sabotage / Financial Warfare

Severity: Critical

Risk Weight: 80–95

Includes

Coordinated attacks on financial systems

Targeting supply chains

Foreign-backed economic destabilization

Signals

Infrastructure targeting

Coordinated panic messaging

Insider or privileged knowledge claims

III. Social Dissatisfaction — Threat Classification

Category S0 — Normal Social Discontent (Non-Threat)

Severity: None

Risk Weight: 0–15

Includes

Complaints about inequality

Gender, youth, regional grievances

Cultural frustration

Examples

“Society is unfair.”

“Young people are ignored.”

Category S1 — Social Polarization Narratives

Severity: Low–Moderate

Risk Weight: 25–40

Includes

Us-vs-them framing

Identity-based resentment

Moral superiority narratives

Signals

“They have everything, we have nothing.”

“We are being replaced.”

Category S2 — Group-Based Mobilization

Severity: Moderate

Risk Weight: 45–65

Includes

Youth, ethnic, class-based mobilization

Protest escalation

Community targeting rhetoric

Signals

“Our people must rise.”

“They are the reason we suffer.”

Category S3 — Social Fracture & Violence Risk

Severity: High

Risk Weight: 65–80

Includes

Calls for collective punishment

Dehumanization

Vigilantism

Signals

“They deserve what’s coming.”

“We must deal with them ourselves.”

Category S4 — Societal Breakdown / Civil Conflict

Severity: Critical

Risk Weight: 85–100

Includes

Civil war rhetoric

Ethnic cleansing narratives

Armed communal violence

IV. How Dissatisfaction Translates Into National Threats

Dissatisfaction → Threat Transition Matrix

ConditionResultGrievance onlyNot a threatGrievance + MobilizationWatchMobilization + TargetingHigh riskTargeting + CoordinationThreatCoordination + ViolenceNational security crisisV. Early Warning Indicators (Very Important)

Security systems monitor patterns, not complaints.

Key Escalation Signals

Rapid narrative convergence

Shift from “I am suffering” → “They are to blame”

Shift from blame → calls for action

Shift from action → timing and coordination

Entry of external amplifiers (media, bots, foreign narratives)

VI. Why This Matters for Modeling (Critical Insight)

Economic and social dissatisfaction are:

Latent variables

Stress indicators

Fuel, not ignition

They are used to:

Predict where instability may emerge

Identify which narratives will resonate

Detect who may exploit grievances

VII. Final Ranking (Integrated)

RankCategoryThreat Level1Mass violence / terrorismExtreme2Infrastructure sabotageExtreme3Armed insurrectionVery High4Economic sabotageVery High5Social fracture → violenceHigh6Economic destabilizationHigh7Mobilized dissatisfactionMedium8Legitimate grievanceNone

1) Integrate satisfaction indices into the risk model

Treat **economic/social satisfaction** as **contextual stressors** that *modulate* threat risk. They should operate at **region/time/topic level**, not individual-level.

A. Define indices (normalized 0–1)

Let:

* **ESI** = Economic Satisfaction Index (higher = more satisfied)

  Inputs: cost-of-living sentiment, unemployment sentiment, wage sentiment, inflation expectations, service delivery sentiment.

* **SSI** = Social Satisfaction Index (higher = more satisfied)

  Inputs: institutional trust, social cohesion sentiment, perceived fairness/inequality, intergroup tolerance, safety/justice sentiment.

* **Shock** = Exogenous/internal shock indicator (0–1)

  Examples: abrupt price spikes, contested election period, major security incident, supply disruptions (measured as anomaly vs baseline).

* **Polarization** = polarization measure (0–1)

  From discourse: clustering/echo-chambers, cross-group hostility, “us-vs-them” prevalence.

Convert satisfaction to **dissatisfaction**:

* **ED = 1 − ESI**

* **SD = 1 − SSI**

B. Add a “Context Stress Multiplier” (CSM)

Base model (your earlier content/network scoring) produces **BaseRisk** (0–100) from intent, specificity, capability, reach, trajectory, network.

Add:

[

\textbf{CSM} = 1 + \alpha \cdot ED + \beta \cdot SD + \gamma \cdot Shock + \delta \cdot Polarization

]

Typical constraints:

* Choose small coefficients so context amplifies but does not dominate:

  * (\alpha,\beta \in [0.05, 0.20])

  * (\gamma \in [0.05, 0.25]) (shocks can amplify more)

  * (\delta \in [0.05, 0.20])

Then:

[

\textbf{AdjustedRisk} = \min(100, \textbf{BaseRisk} \times \textbf{CSM})

]

Interpretation: the *same post pattern* is treated as higher risk when dissatisfaction and shocks are high (because mobilization probability rises), but it still needs **BaseRisk evidence** (intent/specificity/coordination) to escalate.

C. Add a “Grievance Activation Prior” (GAP) for triage

Use satisfaction to adjust **monitoring intensity**, not enforcement.

[

\textbf{GAP} = w_1 ED + w_2 SD + w_3 Shock + w_4 Polarization

]

* Use GAP to decide sampling rates, analyst staffing, and thresholds for review.

Example policy:

* If **GAP high**, lower the *review threshold* for **Tier-2/Tier-3** content (incendiary mobilization / coordination) by a small margin (e.g., 5–10 points), but **never** for Tier-0 lawful speech.

D. A concrete scoring table you can implement

Keep your existing BaseRisk bands, then apply context:

Condition	Rule
BaseRisk < 40	Do not escalate solely due to dissatisfaction
40–60	If CSM ≥ 1.15, route to analyst review (B)
60–80	If CSM ≥ 1.10, route to active monitoring (C)
≥ 80	Escalate (D) regardless; CSM informs urgency

This prevents “people are unhappy” from becoming “people are threats.”

---

2) How dissatisfaction feeds radicalization pipelines (high-level)

Dissatisfaction typically functions as **fuel** that increases receptivity to narratives. The pipeline is a staged transformation from **grievance** to **mobilization**, often via identity and moral reframing.

A. Canonical pipeline stages (non-operational)

1. **Grievance accumulation**

   * Persistent economic strain (ED↑) and/or social distrust (SD↑).

   * Output: generalized frustration, hopelessness, “system failed.”

2. **Attribution & scapegoating**

   * Shift from “things are bad” → “*they* are doing this to us.”

   * Targets can be institutions, elites, or identity outgroups.

3. **Identity fusion**

* Individuals anchor to a group identity that offers belonging and explanation.

* “We are victims; we must defend ourselves.”

4. **Moral disengagement**

* Normalizes hostility; dehumanization and justifications appear.

* Violence becomes “regrettable but necessary” in rhetoric.

5. **Legitimation by narrative**

* Ideology provides a story that sanctifies action (revenge, purification, liberation).

* Often boosted by shocks and polarization.

6. **Mobilization & coordination**

* Transition from belief to action: calls for collective action, timing, roles.

* Network structure becomes the key predictor.

B. Where your satisfaction indices enter the pipeline

* **High ED (economic dissatisfaction)** increases:

* vulnerability to **scapegoating** (“someone stole your future”)

* responsiveness to **calls for disruptive action** (strikes, shutdowns)

* **High SD (social dissatisfaction)** increases:

  * susceptibility to **identity-based explanations**

  * distrust of institutions, making counter-messaging less effective

* **Shocks** accelerate stage transitions (especially 2→4 and 4→6).

* **Polarization** stabilizes radical beliefs (echo-chambers reduce correction).

C. Model signals by stage (safe, analytic framing)

These are *measurement concepts*, not “how to manipulate people.”

Stage	Observable signal family	What to measure (aggregate)
Grievance	economic hardship discourse	ED trend by region/topic
Attribution	blame concentration	“blame targets” entropy drops
Identity fusion	“we/us vs they/them”	pronoun polarity + group terms
Moral disengagement	dehumanization/justification	moral language + contempt markers
Legitimation	ideological certainty	absolutist framing, purity rhetoric
Mobilization	coordination language	time/place cues, collective imperatives
Network	amplification structure	dense clusters, repeated co-sharing

D. Practical integration into your risk engine

1. Compute **BaseRisk** from content + network indicators (your existing model).

2. Compute **CSM** from ED/SD/Shock/Polarization.

3. Produce:

   * **AdjustedRisk**

   * **Stage estimate** (1–6) as a separate label

4. Escalate only when:

   * Stage ≥ 4 *and* BaseRisk moderate/high, or

   * Stage ≥ 6 (coordination), independent of dissatisfaction.

Guardrails (important for correctness and governance)

* **Do not classify dissatisfaction as threat.** Classify it as **context**.

* Operate on **aggregate signals** (region/topic/time) unless there is clear, content-based escalation (coordination/credible threat).

* Track false positives separately for **protected speech** categories.

1) Psychological susceptibility layer (detailed)

Purpose

Estimate **how receptive a population/topic-space currently is** to radical frames, *without profiling individuals*. This layer modulates **pipeline transition speed** (grievance → scapegoating → mobilization), not “threat” by itself.

Unit of analysis

* **Topic × region × time window** (e.g., “cost of living” in Nairobi, past 7 days)

* Optionally **cluster-level** (a narrative cluster), but still aggregate.

What it measures

A **Susceptibility Index (SI)** capturing:

1. **Cognitive rigidity**: black/white reasoning and absolutism.

2. **Identity hunger**: belonging, “we/us” consolidation, “true people” framing.

3. **Humiliation & status loss**: dignity language, disrespect narratives.

4. **Conspiracy closure**: unfalsifiable explanations and “they control everything” loops.

5. **Moral disengagement readiness**: early normalization of harm (not yet action).

Signals (feature families)

* **Absolutist markers**: “always/never/only way/no other option”

* **Purity/contamination framing**: “cleanse/rot/poison/traitors”

* **Humiliation/dignity**: “disrespected, oppressed, laughed at, treated like dogs”

* **Conspiracy saturation**: claims with closed causal loops + rejection of evidence

* **Identity fusion**: “we must unite”, “our people”, “true citizens”

* **Catastrophizing**: “it’s over, nothing can be fixed”

Quantification

Compute SI (0–1) as a weighted blend of standardized sub-scores:

* **RigidityScore**: rate of absolutist + moral certainty phrases per 1k tokens

* **HumiliationScore**: dignity/status-loss lexicon frequency + co-occurrence with blame targets

* **ConspiracyScore**: proportion of posts with unfalsifiable claims and anti-evidence language

* **FusionScore**: group-identity pronoun ratio + in-group/out-group contrast

* **DisengagementReadiness**: justification language co-occurring with dehumanization (without explicit action)

Example:

[

SI = 0.25R + 0.20H + 0.20C + 0.20F + 0.15D

]

Integration rule

* SI does **not** raise escalation on its own.

* SI increases **review priority** when BaseRisk is in the middle band (40–70) and narratives persist.

* SI multiplies **transition likelihood** from stages 2→4 in your radicalization model.

Failure modes & guardrails

* **Political slogans** can mimic rigidity language → require persistence + network evidence.

* **Religious language** can look like “purity” rhetoric → require hostility/targeting co-signals.

* Avoid individual labeling entirely.

---

2) Leadership roles layer (detailed)

Purpose

Differentiate *who does what* in a threat ecosystem. Most harmful escalation depends on a small number of **mobilizers and brokers**, not the average participant.

Unit of analysis

* **Account role** (assigned probabilistically)

* **Cluster role composition** (counts/weights of roles)

Role taxonomy (operationally useful)

1. **Ideologue / Framer**

   * Produces coherent justification narratives, long threads, manifesto-style framing.

2. **Mobilizer**

   * Calls for action, coordinates participation, sets urgency (“now”, “tomorrow”, “on Friday”).

3. **Amplifier**

   * High-volume resharing, boosts reach, less original content.

4. **Broker / Bridge**

   * Connects communities; cross-posts narratives across otherwise separate clusters.

5. **Legitimizer**

   * Adds authority cues: “as a lawyer/pastor/veteran…”, institutional-sounding validation.

6. **Gatekeeper / Moderator**

   * Controls channels/groups; pins content; sets norms; can suppress or promote escalation.

7. **Operational signaler** (rare; handle conservatively)

   * Repeated, consistent coordination patterns; treat as “needs review,” not automatic.

Features to classify roles

* **Originality ratio**: original posts vs reposts

* **Call-to-action density**: imperatives + participation verbs

* **Temporal anchoring**: dates, countdowns, “tonight”

* **Cross-cluster posting**: entropy of audience/community spread

* **Centrality**: betweenness for brokers; in-degree for ideologues

* **Authority cues**: credential claims + formal language + “insider” phrasing

Cluster-level risk rules

Risk spikes when:

* Mobilizers exist **and** broker connectivity is high **and** narrative maturity is high.

Concrete gating:

* If **MobilizerProbability ≥ 0.7** for ≥2 nodes in cluster AND **BrokerPresence** high → escalate one level earlier for review.

* If only amplifiers are present without mobilizers/brokers → typically monitor, not escalate.

Failure modes & guardrails

* Influencers/journalists can look like brokers/amplifiers → require **hostile intent + coordination**.

* Satire accounts can mimic mobilization language → require multi-post confirmation and narrative consistency.

---

3) Offline coupling layer (detailed)

Purpose

Estimate whether online dynamics are likely to produce **near-term real-world impact** due to alignment with offline events, calendars, and opportunity windows.

Unit of analysis

* **Time window** (hour/day/week)

* **Country/region** (coarse)

* **Narrative topic**

Offline coupling components

1. **Event proximity**

   * Elections, court rulings, major protests, policy announcements, big sports/cultural gatherings.

2. **Resource constraints**

   * Fuel shortage, food price spikes, transport strikes.

3. **Security context**

   * Recent violence, raids, arrests, incidents (even rumors can be relevant if widespread).

4. **Institutional schedule**

   * Parliament sessions, budget day, exam periods, public holiday weekends.

Event Coupling Factor (ECF: 0–1)

Compute ECF from:

* **Proximity**: how close in time (e.g., within 24–72h)

* **Relevance**: topic alignment (e.g., “election fraud” near election day)

* **Mobilization readiness**: whether mobilizers are present in network

* **Constraint**: whether conditions make disruption easier (e.g., strikes already ongoing)

High-level formula:

[

ECF = f(\text{proximity}, \text{topic relevance}, \text{mobilizer presence}, \text{constraints})

]

Integration rule

* ECF should not label content “worse,” it labels it **more urgent**.

* ECF primarily affects **Time-to-Action** and **triage priority**, and modestly increases escalation when BaseRisk is already high.

Failure modes & guardrails

* Rumors about events can inflate ECF; require corroboration via multi-source signals (not necessarily external web, could be multi-cluster concordance).

* Keep location coarse to reduce privacy risk.

---

4) Legitimacy erosion layer (detailed)

Purpose

Separate:

* “Government policies are bad” (normal dissent)

  from:

* “State has no right to rule; ignore courts/police/elections” (system destabilization risk)

Unit of analysis

* Narrative/topic × time × region

* Cluster-level rhetoric patterns

Legitimacy Erosion Index (LEI: 0–1)

Subcomponents:

1. **Institution rejection**

   * Courts: “courts are fake; ignore judgments”

   * Elections: “voting is pointless; results are decided”

   * Police: “police are illegitimate; do not comply”

2. **Parallel authority endorsement**

   * “Only X movement is legitimate”

3. **Noncompliance advocacy**

   * “Do not pay taxes”, “stop obeying laws” (systemic, not specific protests)

4. **Dehumanization of institutions**

   * “State is a parasite”, “regime is cancer”

5. **Delegitimation with mobilization**

   * The above paired with “act now”, “take over”, “remove them”

Why LEI matters

LEI predicts shifts into:

* insurrection narratives,

* institutional sabotage,

* targeted hostility to officials.

Integration rule

* LEI increases the probability that a narrative should be classified into **Tier 2/3** categories (insurrection/election subversion), but only when coupled with mobilization/coordination evidence.

Guardrails

* Strong criticism ≠ LEI unless it includes **rejection + noncompliance + replacement** cues.

* Legal/political debates about reform can resemble delegitimation; require hostility markers and sustained pattern.

---

5) Narrative vs campaign maturation layer (detailed)

Purpose

Disentangle:

* **Rumor** (unstable, scattered)

* **Narrative** (stable frame, repeatable)

* **Campaign** (coordinated, goal-oriented, timed)

This layer is essential because national-level harm usually comes from **campaigns**, not isolated posts.

Definitions

* **Rumor**

  * High variance, short half-life, weak structure, no stable villain/goal.

* **Narrative**

  * Stable villains/victims, repeated slogans, consistent causal story, identity-linked.

* **Campaign**

  * Narrative plus coordination: timing, “what to do,” amplification strategy, repeated templated assets, cross-community propagation.

Maturation Score (MS: 0–100)

Compute MS using:

1. **Persistence** (days/weeks survival)

2. **Template similarity** (reused phrases, slogans, images)

3. **Target consistency** (same blamed entity/group)

4. **Goal clarity** (“boycott X”, “occupy Y”, “withdraw funds”)

5. **Coordination evidence** (synchrony, mobilizers, shared links)

6. **Cross-cluster spread** (broker activity)

Interpretation:

* MS < 35 → Rumor

* 35–70 → Narrative

* > 70 → Campaign

Integration rule

* Campaign status (>70) is a major gate for **active monitoring**.

* Narrative status triggers **analyst review** if BaseRisk moderate.

Failure modes & guardrails

* Organic memes can look templated; require **goal clarity + coordination** to classify as campaign.

* News cycles create temporary persistence; require **target consistency + action orientation**.

---

6) Resilience modeling layer (detailed)

Purpose

Model the environment’s ability to **absorb and correct** destabilizing narratives (reducing cascade risk and false positives).

Unit of analysis

* Topic × region × time

* Cluster-level counter-signal

Resilience Index (RI: 0–1)

Subcomponents:

1. **Counter-narrative presence**

   * credible voices disputing the claim

2. **Correction spread**

   * debunks travel as far/fast as claims

3. **Cross-group bridging**

   * evidence of dialogue across communities

4. **Institutional responsiveness**

   * visible, trusted responses reduce volatility (model as perception signal, not “government says”)

5. **Community moderation capacity**

   * effective de-escalation norms in major hubs

Integration rule

* RI **dampens** escalation (small effect) and is primarily used to:

  * reduce urgency when high,

  * allocate analyst time where resilience is low.

Failure modes

* “Counter-signal” can be hostile too (e.g., inflammatory rebuttals). Require that counter-signal reduces targeting/mobilization, not increases it.

---

7) Ethical guardrails layer (detailed)

Purpose

Prevent the system from becoming a “dissent detector,” and ensure it remains focused on **mobilization toward harm** and **coordinated destabilization**.

Non-negotiable guardrail rules (implement as hard constraints)

1. **Protected speech firewall**

   * Criticism, satire, peaceful protest advocacy: never auto-escalate beyond Level 1.

2. **No sentiment-only escalation**

   * Anger, profanity, insults do not imply threat.

3. **No group membership inference**

   * Do not infer protected characteristics; avoid identity labeling beyond what is necessary for hate/violence detection.

4. **Single-post limitation**

   * Do not produce Level 4/5 solely from a single post unless it’s an explicit, credible targeted threat; even then, require human review.

5. **Transparency logging**

   * Every escalation must store which signals triggered it (intent, coordination, maturity, specificity).

6. **Minimum evidence thresholds**

* Coordination-based escalation requires NCS evidence; campaign-based requires MS evidence.

7. **Human-in-the-loop for uncertainty**

* If adversarial adaptation is high or confidence low → route to review, not action.

Output requirement

Always include:

* **Confidence**

* **Which signal families fired**

* **Which guardrails applied**

* **Why this is not protected speech** (when escalating)

---

8) Structural priors layer (detailed)

Purpose

Capture long-horizon conditions that raise baseline volatility. Structural priors should **never trigger escalation**, only adjust baseline monitoring resources and context.

Unit of analysis

* Region × long time (months/years)

* Topic baseline

Structural Prior (SP: 0–1)

Inputs can include:

* youth unemployment trend

* inequality persistence

* historical election-cycle volatility

* chronic service delivery deficits

* past communal conflict indicators

* long-term trust trends

Integration rule

* SP affects:

  * sampling rates,

  * analyst staffing,

  * context multipliers modestly,

    but cannot override lack of content/network evidence.

Guardrail

SP cannot raise escalation above Level 1 without BaseRisk ≥ 40.

---

9) Adversarial adaptation layer (detailed)

Purpose

Detect when actors are **evading detection** (coded language, irony shields, vocabulary drift) while maintaining harmful intent/coordination patterns.

Unit of analysis

* Cluster-level changes over time

* Account-level drift patterns (aggregate)

Adversarial Adaptation Score (AA: 0–1)

Signals:

1. **Semantic drift with stable network**

   * language changes while amplification graph remains the same

2. **Codeword emergence**

   * new tokens correlated with previously flagged intents

3. **Irony masking**

   * sarcasm markers co-occurring with targeting/mobilization cues

4. **Platform migration**

   * sudden declines on one surface and synchronized rise elsewhere (if you ingest multiple sources)

5. **Compression**

   * shorter, more ambiguous posts paired with external links/images

Integration rule

* AA does **not** increase severity; it increases:

  * uncertainty,

  * need for human review,

  * need for adaptive lexicons/embeddings.

* If AA high and BaseRisk mid/high → escalate to **review**, not “critical.”

Guardrails

* Avoid expanding surveillance; focus on **model robustness**:

  * contextual embeddings,

  * drift detection,

  * analyst feedback loops.

---

10) Time-to-action layer (detailed)

Purpose

Estimate **when** risk could translate into real-world impact. This is a separate output from severity.

Unit of analysis

* Narrative/cluster level

* Rolling time windows

Outputs

* **Immediate (0–24h)**

* **Near-term (1–7d)**

* **Medium (1–4w)**

* **Chronic (1–6mo)**

Time-to-Action predictors

1. **Temporal anchoring**

   * explicit dates, “tomorrow,” “tonight,” countdowns

2. **Frequency compression**

   * rapid acceleration in posting/sharing

3. **Coordination density**

   * NCS high: synchronized, templated propagation

4. **Campaign maturity**

   * MS high: stable goals + repeated assets

5. **Offline coupling**

   * ECF high: event proximity

6. **Mobilizer presence**

   * roles actively instruct participation

7. **Legitimacy erosion + noncompliance cues**

  * when paired with action language, shortens TTA

Integration rule

* TTA sets **triage priority and monitoring cadence**, not severity.

* Example: two narratives both risk=70; one TTA=Immediate gets priority.

Guardrails

* Do not infer capability to act from demographics.

* Require at least two independent predictors (e.g., temporal anchoring + NCS) to label “Immediate.”

How these layers connect (minimal integration map)

Evidence vs Context vs Governance

* **Evidence**: CES/BES/NCS + MS (what is happening)

* **Context**: ED/SD/Shock/Polarization + SI + SP + ECF + RI (how likely it escalates, how fast)

* **Governance**: Guardrails + AA + Confidence (how carefully to treat outputs)

Practical escalation gating (recommended)

* Escalation above “review” requires **evidence** (coordination or campaign maturity or credible specificity).

* Context layers can only **prioritize**, not create a threat.

---

what do you think?
This is an exceptionally sophisticated, professional-grade design. It moves far beyond standard sentiment analysis into **behavioral dynamics** and **systemic risk modeling**.

It mirrors frameworks used by advanced intelligence units, Trust & Safety teams at major platforms, and conflict monitoring NGOs.

Here is a design review of your framework, highlighting what works brilliantly, where the implementation risks lie, and specific recommendations to tighten the architecture.

1. The Core Differentiator: Separating "Stress" from "Threat"

The Design Win:

Your distinction between **Dissatisfaction (Context)** and **Threat (Mobilization)** is the strongest part of this architecture. Most commercially available social listening tools fail here—they flag "high negative sentiment" as a crisis.

- By treating Economic/Social Dissatisfaction as a **Context Stress Multiplier (CSM)** rather than a direct threat signal, you drastically reduce false positives.
- **Why this works:** It acknowledges that a population can be furious (High Dissatisfaction) but passive. The threat only materializes when **Capability** and **Coordination** are introduced.

The Design Risk:

- **Double Counting:** Be careful that variables in your **BaseRisk** (e.g., Network Density/Polarization) do not overlap too heavily with variables in your **CSM** (e.g., Polarization). If "echo chambers" increase both the BaseRisk score *and* the CSM, you will artificially inflate the risk score for every heated political debate.
- **Recommendation:** Ensure "Polarization" in the CSM is measured by *sentiment/language* (us vs. them), while "Network Density" in BaseRisk is measured strictly by *graph topology* (interconnectivity). Keep semantic and structural metrics separate.

2. The Role Taxonomy & Leadership Layer

The Design Win:

Moving beyond "Influencers" to functional roles (**Mobilizer, Broker, Ideologue, Operational Signaler**) is critical.

- **Mobilizers** create the *time* (When).
- **Ideologues** create the *reason* (Why).
- **Brokers** create the *scale* (Reach).

The Refinement Opportunity:

You are missing one critical role: **The Unwitting Amplifier / "The Useful Idiot."**

- **Definition:** High-reach accounts (celebrities, mainstream news) that quote-tweet or share the *Ideologue's* content with a caption like "Can you believe this?"
- **Why it matters:** In almost every modern insurrection or riot, the jump from Tier 4 (Emerging) to Tier 2 (Severe) happens because a *Non-Threat* actor amplified a *Threat* actor.
- **Adjustment:** Add a "Cross-Tier Contagion" metric. If a Tier 0 account shares Tier 3 content, the **Reach** multiplier should spike, even if the Tier 0 account is not hostile.

3. The "Time-to-Action" (TTA) Layer

The Design Win:

Separating **Severity** (How bad is it?) from **Urgency** (When will it happen?) is excellent triage design. A "High Severity / Low Urgency" threat (e.g., a long-term separatist manifesto) requires a different response than a "Medium Severity / Immediate Urgency" threat (e.g., a flash mob looting call).

The Implementation Challenge:

- **Temporal Anchoring is noisy.** "Tomorrow" is easy to detect. But phrases like "When the time comes" or "Soon" are difficult to map to a TTA window.
- **Recommendation:** Weight **Offline Coupling (ECF)** heavily here. If the post says "We strike tomorrow," that is alarming. If the post says "We strike tomorrow" *and* tomorrow is Election Day (ECF High), that is critical. The text alone is unreliable without the calendar context.

4. Psychological Susceptibility & The NLP Gap

The Design Win:

Trying to measure **Cognitive Rigidity** and **Identity Fusion** is academically sound (based on Radicalization Psychology).

The Reality Check:

Detecting "Absolutist markers" or "Humiliation" via NLP on social media text is incredibly prone to error.

- *Sarcasm/Slang:* "I'm dead" (slang for laughing) vs. "I'm dead" (suicidal ideation) vs. "They are dead" (threat).

- *Cultural Nuance:* In some regions, aggressive posturing is a linguistic norm, not a mobilization signal.
- **Recommendation:** Do not rely on "keyword counting" for the **Susceptibility Index (SI)**. You need to look for **Linguistic Complexity drops**. Research shows that as violence approaches, the complexity of grammar drops, and sentence length shortens. Use "lexical diversity" (Type-Token Ratio) as a proxy for Cognitive Rigidity. Lower diversity = Higher rigidity.

5. Legitimacy Erosion (LEI) vs. Protected Speech

The Design Win:

The **LEI** is a vital "early warning" radar. It predicts the *collapse* of order before the *act* of violence.

The Guardrail Gap:

- **Scenario:** A legitimate whistleblower releases evidence of massive corruption in the judiciary. Public response: "The courts are fake," "Do not obey these judges."
- **The Conflict:** Your model flags this as High LEI (Institution Rejection + Noncompliance). However, this is arguably a democratic corrective process, not a national threat.
- **Refinement:** You need a **"Truth/Event Grounding"** check. If LEI spikes immediately following a *verified* news event (scandal), the system should tag it as **Reactive**, not **Structural**. Structural erosion (happening without a specific trigger) is the true threat.

6. The "Maturation" Concept

The Design Win:

Distinguishing Rumor → Narrative → Campaign is brilliant.

The Refinement:

- **Campaigns require Resources.** A true "Campaign" (Level 3 maturation) usually leaves a **financial footprint** or a **bot footprint**.
- **Add Signal:** Look for **"Asset Replication."** If the exact same image meme appears in 50 different clusters simultaneously, that is not organic sharing; that is a *Campaign*. If the text varies slightly but the *image hash* is identical, treat it as coordinated.

7. Missing Layer: Platform Migration & Dark Social

You mentioned "Platform Migration" under Adversarial Adaptation, but it deserves more weight.

- **The Dynamic:** Threats often **Recruit** on open platforms (Twitter/X, Facebook) but **Coordinate** on closed platforms (WhatsApp, Telegram, Signal).
- **The Signal:** The "Bridge to Dark" signal.
- **Logic:** If a high-risk cluster suddenly stops posting on the open web, *or* if they start posting "DM for details" or "Join the channel [Link]," the **Time-to-Action** should immediately shorten. The silence on the open web is often the loudest alarm.

Summary of Recommendations

1. **Add a "Bridge to Dark" Signal:** Monitor the transition from public grievance to private coordination links.
2. **Refine "Useful Idiot" Tracking:** Monitor when trusted/neutral nodes amplify radical nodes.
3. **Use Lexical Diversity for Rigidity:** Replace pure keyword counting with grammatical complexity analysis for the Susceptibility Index.
4. **Decouple Context from Structure:** Ensure CSM variables don't overlap with BaseRisk variables to avoid score inflation.
5. **Event Grounding:** Check if "Legitimacy Erosion" is reactive to a specific scandal (lower risk) or systemic (higher risk).

Verdict:

This is a robust, production-ready logical architecture. If you can solve the data engineering challenge of ingesting this much context (specifically the "Offline Coupling" and "Structural Priors"), the output will be significantly more valuable than standard threat intelligence feeds.