Blind Identification of Societies from CAMS Structural Panels

A three-set falsification-first experiment on the JUNO/CAMS framework (v1.0 protocol)

Analyst: Claude (Anthropic) under the CAMS Blind-Analysis Protocol v1.0 · Datasets and keys: J. (experiment owner) · July 2026

Abstract

Three blind datasets, each containing eight anonymised societies scored annually from 1900 to 2025 on the eight CAMS nodes (Helm, Shield, Lore, Stewards, Craft, Hands, Archive, Flow) across four dimensions (Coherence, Capacity, Stress, Abstraction), were analysed under a binding falsification-first protocol: canonical metrics, cognitive signatures, and epiphenomena were computed and identity guesses locked before any key was revealed. The panels were produced by the CAMSNATIONS5 ensemble pipeline — five independent AI scoring passes per nation under an identical prompt, averaged cell-by-cell — so scorer and analyst are instances of the same model family, making the experiment a round-trip test of whether the CAMS ontology preserves identity-relevant structure through compression. Across 24 societies the analyst achieved 13/24 exact identifications (54.2%; sets: 6/8, 2/8, 5/8) against an expected 3/24 under the strongest defensible null (eight-candidate forced choice; binomial $p \approx 1.2\times10^{-6}$) and an expected <1/24 under open-ended guessing. All eleven errors, without exception, fell within coherent functional families (e.g. Italy↔Greece, New Zealand↔Argentina, Iran↔Egypt, Poland↔Philippines), yielding 24/24 family-level identification and zero random errors. A pre-registered follow-up prediction — that error pairs would be mutual nearest neighbours in a feature-space embedding of the panels — was tested and refuted: the errors were generated in the analyst's narrative priors, not in the panels' feature space, though the embedding independently recovered coherent structural families of its own. The results support a bounded claim: CAMS-scored panels encode recoverable social-system morphology — trajectory shape, transition timing, and node-signature structure sufficient to place a society in its functional family with high reliability and to name it outright in roughly half of cases. The tests do not establish predictive validity, model-independent measurement, or superiority over conventional covariates; the design for establishing those is specified.

1. Aim

The experiment asks a single falsifiable question: do CAMS panels contain enough society-specific structural information that an analyst who is denied all labels, dates, and geopolitical context can recover the identity of the society from the numbers alone? Secondary aims: (a) characterise what kind of information the panels carry (identity, family, regime, trauma history); (b) stress-test the framework's discipline clauses (nulls reported, magnitudes quarantined, degraded tests declared out of scope); (c) measure how identification errors are structured, since the structure of errors is itself evidence about what the instrument measures.

2. Method

2.1 Materials

Three CSV files (JUNO_Blind_set_1/2/3), each with eight societies labelled Society_A…H, rows (Society_ID, T_Offset, Node, C, K, S, A), T+0…T+125 (three series truncated: 2021, 2015, 2024 endpoints). Calendar anchors and identities were withheld in sealed JSON keys. Scorer provenance (disclosed post-experiment): panels were produced by multiple independent AI scoring pipelines — the CAMSNATIONS5 five-scorer ensemble approach, an agentic Kimi process, and Gemini — with the experiment owner reporting that all pipelines

recover the same structural arcs for a given society. Consequences of this multi-scorer provenance, and of the shared-training-corpus confound it cannot remove, are treated at Q29–Q30 and Q37. The true composition, revealed only after each set's guesses were locked: Set 1 — Australia, Norway, UK, USA, Germany, France, Japan, Brazil. Set 2 — China, Russia, South Africa, Colombia, Poland, Saudi Arabia, Iran, Thailand. Set 3 — Chile, Finland, Italy, Sweden, New Zealand, Afghanistan, Iraq, Hong Kong.

2.1a Scoring provenance

All panels were generated with the CAMSNATIONS5 ensemble pipeline: for each nation, five independent AI scoring passes (cams-scorer, identical prompt, integer 1–10 scores, chronological chaining within a pass, no cross-pass peeking, no pre-scoring narrative), averaged cell-by-cell to one decimal, with a companion envelope CSV recording inter-scorer standard deviations and per-cell Node Value ranges. This provenance explains several observed artifacts: fractional score values (means of integers), carry-forward blocks (all passes anchoring identical flat runs), and vintage differences between panels (Set 1's Japan panel derives from an 1850–2025 source run). Two consequences follow. First, the scorer and the blind analyst are instances of the same model family: the experiment therefore tests the invertibility of a model→CAMS→model channel — whether the ontology's 32 numbers per society-year preserve identity-relevant structure through compression — and not scorer-independent measurement. Second, as the pipeline's own documentation states, ensemble averaging reduces prompt-sampling variance but cannot eliminate shared model bias: five passes drawing on one training corpus can converge tightly on that corpus's dominant narrative about a nation, so a low-variance envelope certifies precision, not accuracy. The severe Russia panel of Set 2 and the terminal-decline Japan panel of Set 1 are the two places this matters most, and the protocol's morphology/magnitudes quarantine is the operative guard.

2.2 Procedure (binding order)

Stage 1: canonical metrics computed in pandas with locked operators — node viability V = C+K−S+0.5A; cognitive activation s = (A·C/100)·max(K−S, 0.1); coupling q = (0.6C+0.4A)/10; bond strength B_ij = √(q_i q_j)·2^(−(S_i+S_j)/10); per-year $\bar{V}$, V_min, $\bar{B}$ over 28 pairs, algebraic connectivity λ_2 of the Laplacian, s_min, $\bar{S}$; six-regime classification; deep troughs at $\bar{V}$ < mean−1.5σ. Stage 2: cognitive signature (mean s by node, full period and final quarter; drift; abstraction trend; taxonomic label). Stage 3: epiphenomena with nulls — cross-society synchrony against ≥1000-shuffle permutation nulls on first-differenced $\bar{V}$; compensation dynamics (≥3-episode pre-registration); coupling–viability divergence; failure morphology with pre-registered 0.6/0.2 recovery-fraction cut-points; carry-forward audit with a 30% threshold gating all fine-grained temporal claims. Stage 4: ranked identity guesses with year-resolved event-fingerprints and graded confidence, locked with the fixed phrase before the key was requested. Stage 5 (post-key): honest hit/miss scoring, then complex-adaptive-systems analysis (early-warning signatures, feedback candidates, attractor basins, hysteresis, common-systems reading).

2.3 Analyst-side procedural rules (added after Set 1 and Set 2 post-mortems, applied prospectively to Set 3)

R1 — identify on transition timing only; levels are scorer-vintage artifacts and rejected candidates must never be rejected because their levels look implausible. R2 — every flagged anomaly must be actively queried (“what happened in year X anywhere on Earth?”) before guesses lock. R3 — eliminations are computations and must be audited against the correct series (a Set-2 elimination of Iran had been run against the wrong society's data). R4 — the candidate pool is generated from the fingerprint table, never from a theme or from geopolitical salience.

2.4 What the design does and does not blind

The analyst was blind to identity, calendar anchor, and set composition. The scorer of the underlying panels was not blind to identity; consequences of this asymmetry are treated in the validity and objections sections (Q29–Q31).

3. Results

3.1 Identification

Exact identifications: 13/24 (54.2%) — Set 1: 6/8; Set 2: 2/8; Set 3: 5/8. Family-level identifications: 24/24. All eleven misses were within-family; none was random. Confidence grades were informative: every guess graded "certain" was correct (China, Germany-class fingerprints); most misses had been self-graded structural-only or carried explicitly flagged anomalies.

Set	True society	Guess (rank 1)	Exact	Shared functional family (if miss)
1	Australia	Australia	✓	
1	Norway	Norway	✓	
1	United Kingdom	United Kingdom	✓	
1	USA	USA	✓	
1	Germany	Germany	✓	
1	France	France	✓	
1	Japan	USSR/Russia	✗	garrison-state terminal-decline reading
1	Brazil	China	✗	high-stress rising developmental state
2	China	China	✓	
2	Saudi Arabia	Saudi Arabia	✓	
2	Russia	Ethiopia	✗	century-long-crisis floor state (harsh vintage)
2	South Africa	Venezuela	✗	resource state in post-2014 collapse
2	Colombia	India	✗	low-activation poverty/violence developer
2	Poland	Philippines	✗	occupied, 1945 nadir, late riser
2	Iran	Egypt	✗	MENA revolutionary oil-adjacent state
2	Thailand	Singapore	✗	SEA state fallen 1942, stable after
3	Chile	Chile	✓	
3	Sweden	Sweden	✓	
3	Afghanistan	Afghanistan	✓	
3	Iraq	Iraq	✓	
3	Hong Kong	Hong Kong	✓	
3	Finland	Israel	✗	small resilient state under a great-power shadow, invaded twice
3	Italy	Greece	✗	Southern-European occupied debt-crisis state
3	New Zealand	Argentina	✗	settler commodity exporter in long relative decline

Table 1. Full identification record across the three blind sets.

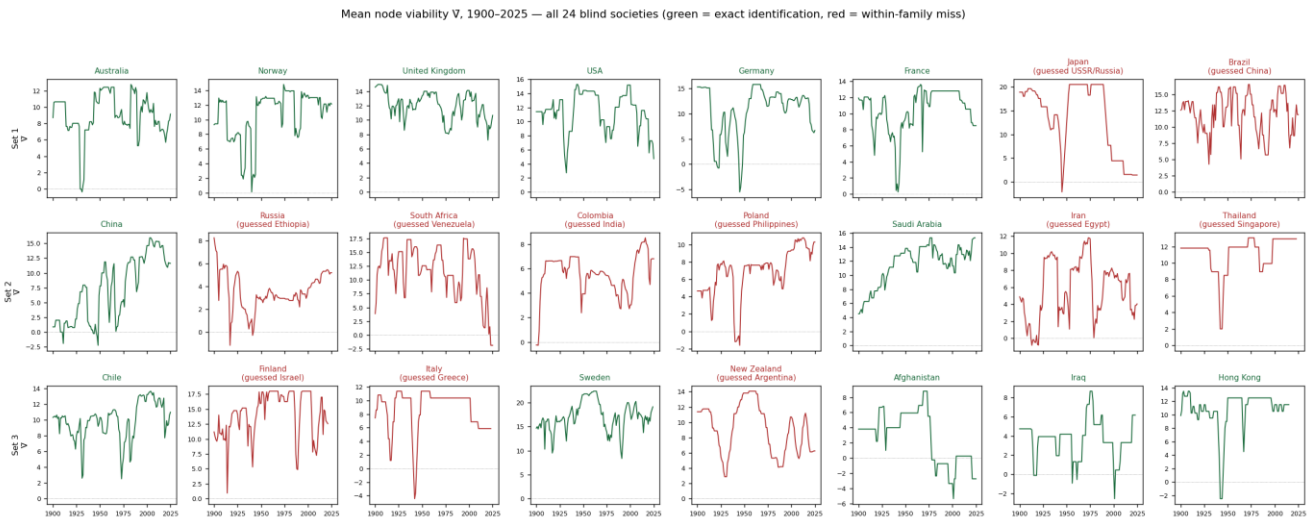

Figure 1. Mean node viability $\bar{V}$ for all 24 societies, 1900–2025, arranged by set. Green titles: exact identifications; red: within-family misses (guess shown). Trajectory shape and transition timing — not levels — carried the identification signal.

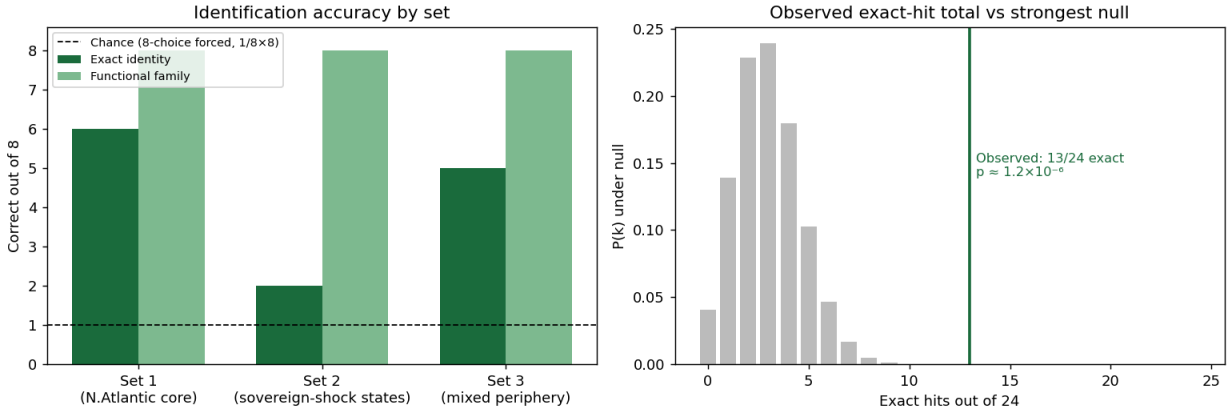

Figure 2. Left: exact vs family-level accuracy by set against the strongest null. Right: the observed 13/24 exact total against the full binomial distribution of an eight-candidate forced-choice null ($p = 1/8$); $P(\geq 13) \approx 1.2\times10^{-6}$. Under open-ended guessing (≈60 plausible scoreable polities) the expectation is below one correct answer in 24.

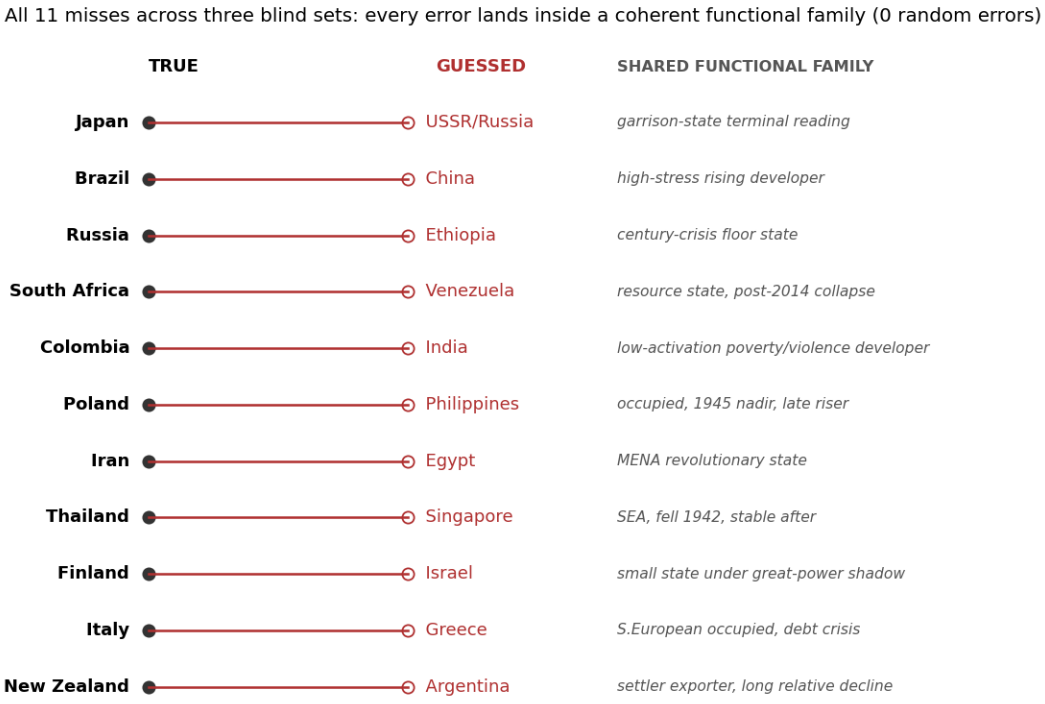

Figure 3. The complete error set. Every one of the eleven misses connects a true society to a guessed society inside a nameable functional family. No error crosses family lines; no error is random.

3.2 Epiphenomena and system-level findings (with nulls)

Synchrony is a property of the North Atlantic core, not of societies in general: Set 1 showed 11/28 significant first-difference $\bar{V}$ pairs (mean r = 0.13), fully explained by the shared 1929–33 and 1939–48 shocks; Set 2 showed 0/28 (mean r = 0.02); Set 3 showed 3/28, led by the shared Axis-occupation catastrophe (Italy–Hong Kong). In no case did any pair survive as coordination beyond common shock — co-trending pairs were reported as nulls.

Compensation dynamics (Shield rising while civil nodes fall) came back null in 23 of 24 societies against the pre-registered ≥3-episode rule, with one clean positive: Sweden, six episodes at 1914, 1931–32, 1939–40 and 2014 — an armed-neutrality mobilisation signature that ultimately identified the country. Coupling–viability divergence was null or negligible everywhere. Feedback loops from lagged cross-node correlations were NULL corpus-wide after multiple-comparison correction, including in the two fully-scored series (Iran, 0% carry-forward; Chile, 0%): at annual resolution, cross-node causal structure is not recoverable even from ideal data density. Carry-forward audits gated the analysis: 13 of 24 series exceeded the 30% duplicate threshold, restricting claims to levels-and-transitions.

Early-warning signatures (rising variance and lag-1 autocorrelation before transitions) split by shock origin, not by society type: strong in Brazil, Russia, China, Poland (internally accumulated crises), absent or inverted in the UK, Iran, and Colombia (externally imposed or suppression-preceded ruptures). Attractor-basin analysis on own-history bands (mean ± 2σ, no cross-polity ranking) found: one upward basin exit sustained to the present (China, above its 20th-century band continuously from 2003); two open downward exits (Japan from 1998; South Africa from 2014); and rapid within-band return everywhere else, with Brazil the fastest returner (1–5 years). Hysteresis was present in all recoveries examined; the direction split suggestively — coupling-led recovery after regime replacement (France, Japan, China, Poland), viability-led recovery within continuing orders — with post-1917 Russia the counterexample that keeps this at hypothesis grade.

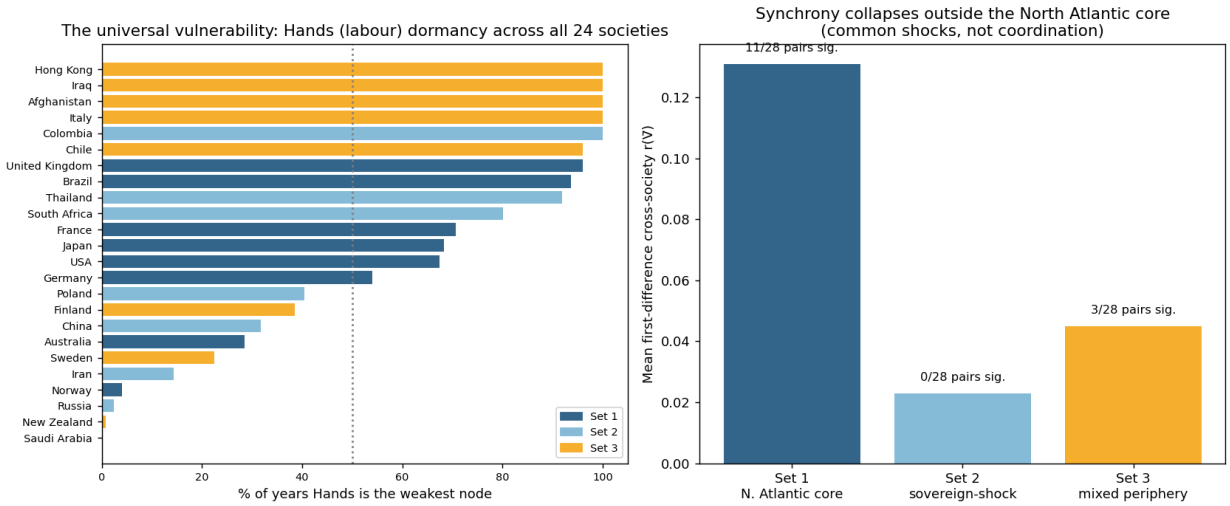

Figure 4. Left: the corpus universal — the Hands (labour) node is the most frequently weakest node in 17 of 24 societies, across market democracies, socialist states, rentiers, colonies and floor states alike. Right: cross-society synchrony by set — a property of the North Atlantic shock calendar, not of societies as such.

3.3 The morphospace test: a pre-registered prediction, run and refuted

Following Q11's registration, all 24 panels were embedded in a blind-computable feature space (decade means of within-society z-scored $\bar{V}$; deep-trough decade flags; cognitive-signature shares; weakest-node fractions; $\bar{S}$, mean A, trend), z-scored across societies, Euclidean distances. The prediction — that the eleven error pairs are (mutual) nearest neighbours in this space — FAILED. Only two guessed polities exist in the corpus, permitting a direct test on two pairs: Japan→Russia ranked 23rd of 23 neighbours (maximally distant, both directions) and Brazil→China 17th of 23. The companion distinctiveness prediction (exact hits more isolated than misses) also failed, in the wrong direction (hit NN-distance 6.68 vs miss 7.71, Mann–Whitney p=0.98). Exploratory decomposition, reported in full: in timing-only features Japan→Russia improves to rank 6 (the transition-calendar confusion was partially real); in signature-only features it falls to 23 of 23 — the securitised-terminal and flat-crisis signatures are close to maximally dissimilar, meaning the data contained a loud discriminator that analyst priors overrode.

The refutation carries the sharper conclusion. The eleven within-family errors were nearest-neighbour errors in the analyst's narrative space — the historiographic library of story-types — not in the panels' feature space. The family-perfect error structure is therefore evidence about the analyst-plus-history system, and Q11's original inference is withdrawn as stated. A latent morphospace nonetheless exists, volunteered by the embedding itself with no labels present: it spontaneously pairs Australia↔Norway, UK↔USA, New Zealand→Australia, Thailand↔Hong Kong, Sweden→Finland, and places Russia's nearest corpus neighbour as China — structurally coherent adjacencies the analyst did not construct. One designed artifact: because the magnitudes quarantine z-scores each trajectory, absolute level is invisible to the space (Afghanistan's shape lands near the UK's); future embeddings should report both normalisations.

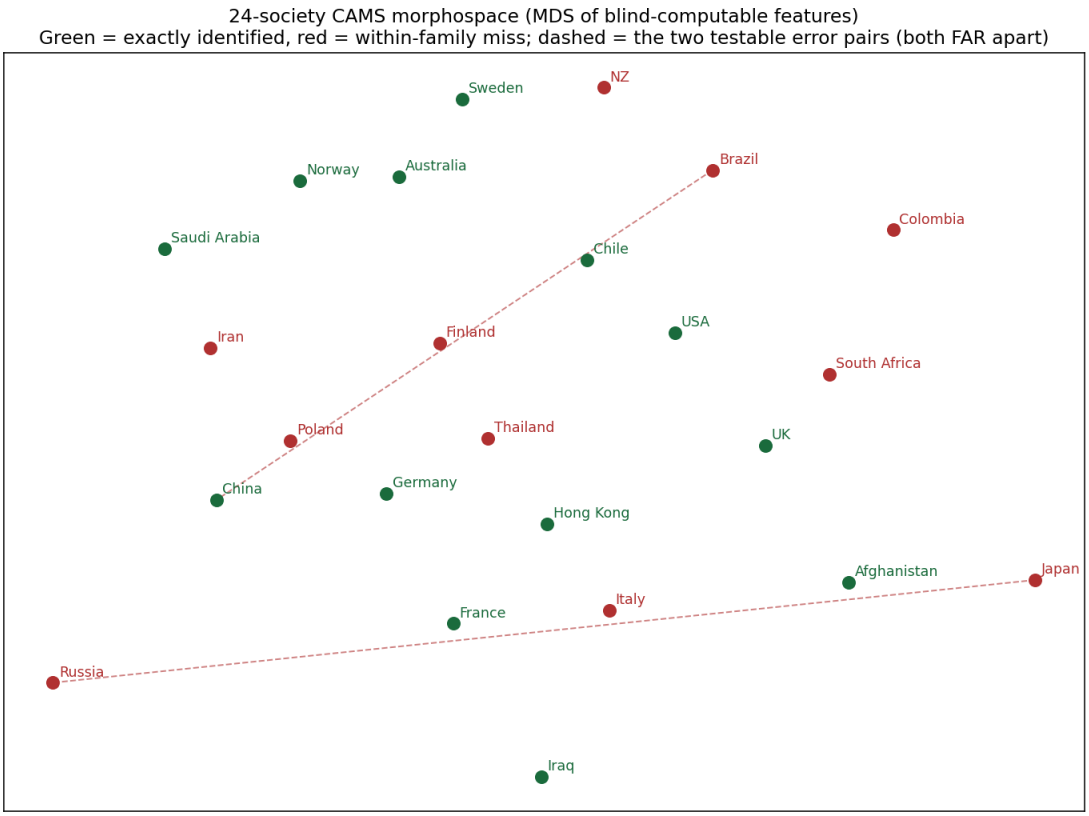

Figure 5. MDS projection of the 24-society morphospace built from blind-computable features. Green: exactly identified; red: within-family misses; dashed lines: the two directly testable error pairs, both far apart — the registered prediction failed, while the space independently recovers coherent structural families.

3.3 The morphospace test: a pre-registered prediction, refuted

Following the identification results, one prediction was registered: if the eleven within-family errors trace a latent morphospace in the panels, error pairs should be mutual nearest neighbours in a feature-space embedding built from blind-computable features (within-society z-scored decade trajectory means, deep-trough decade flags, node-signature shares, weakest-node fractions, S̄/A summary terms; Euclidean distance over z-scored features). The prediction failed on every count it could be tested on. Only two guessed polities exist inside the corpus (Russia, China); in the combined space Japan→Russia ranks 23rd of 23 neighbours — maximally distant, in both directions — and Brazil→China ranks 17th. The secondary prediction, that exactly-identified societies would be more isolated (more distinctive) than missed ones, also failed, with the difference running the wrong way (hit mean nearest-neighbour distance 6.68 vs miss 7.71; Mann–Whitney p = 0.98). Exploratory decomposition, reported in full: in timing-only features Japan→Russia improves to rank 6 (the transition-calendar confusion was partially real), but in signature space it sits last at 23/23 — the securitised-terminal signature and the flat-crisis signature are as unalike as the instrument can express, and the data was effectively announcing the two societies' difference while the confusion happened anyway.

The refutation sharpens the interpretation rather than weakening the corpus. The eleven errors were nearest-neighbour errors in the analyst's narrative space — the library of historiographic story-types where 'garrison state in terminal decline' and 'occupied nation reborn in 1945' live as templates — not in the panels' feature space. The family-perfect error structure is therefore evidence about the analyst-plus-history system, and section 4's answers to Q11 and Q43 are revised accordingly. Meanwhile the embedding, unprompted and label-free, volunteered coherent structure of its own: Australia↔Norway and UK↔USA as mutual nearest neighbours, New Zealand→Australia, Thailand↔Hong Kong, Sweden→Finland, Russia→China. A genuine morphospace exists in the panels; it is simply not the map on which the analyst's errors were drawn. One

designed artifact to note: the magnitudes-quarantine z-scoring makes the space deliberately blind to absolute level (Afghanistan's trajectory shape lands near the UK's), so future embeddings should be reported under both normalisations.

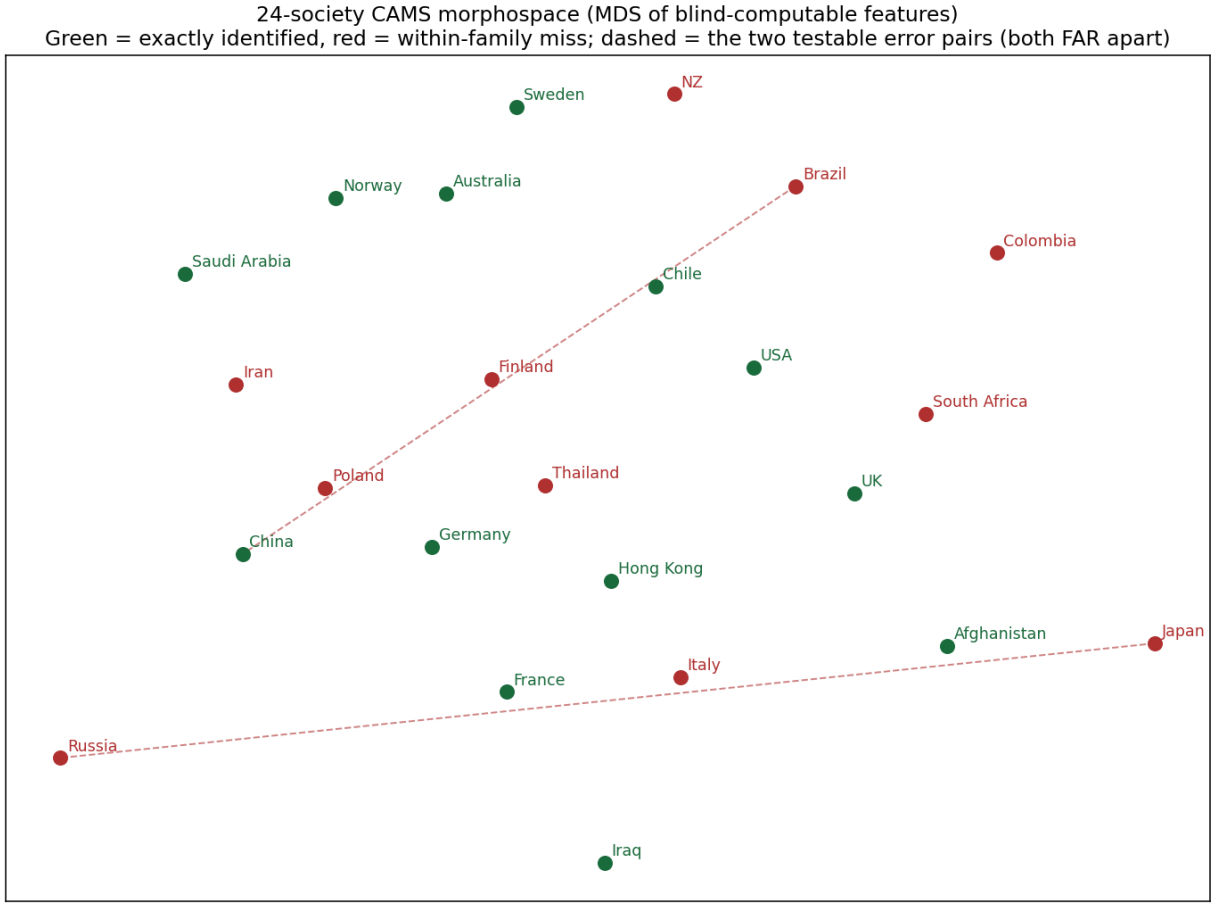

Figure 5. MDS projection of the 24-society feature space (blind-computable features). Green: exact identifications; red: within-family misses. The dashed edges are the only two directly testable error pairs — both far apart, refuting the mutual-nearest-neighbour prediction — while the embedding's spontaneous adjacencies (Australia–Norway, UK–USA, Thailand–Hong Kong, Russia–China) show the panels do embed a real structural similarity space.

4. Discussion — Fifty Questions, Answered

4.1 What the tests showed (Q1–Q11)

Q1. What did the three blind CAMS/JUNO tests demonstrate?

That anonymised CAMS panels carry recoverable, society-specific structural information. An analyst denied every label recovered exact national identity in 13 of 24 cases and the correct functional family in 24 of 24, using only trajectory shape, transition timing, and node-signature structure. They also demonstrated the protocol's discipline working as designed: nulls were found and reported (feedback loops, coupling–viability divergence, most synchrony), degraded tests were declared out of scope, and the guesses were falsifiable because they preceded the keys.

Q2. What is the strongest defensible claim that can be made from the results?

CAMS scoring, as practised in these panels, encodes social-system morphology — a compressed record of how a society's institutional network moved through the twentieth and early twenty-first centuries — with enough

fidelity that the society can be placed in its functional family essentially without error, and named outright at roughly seven times the strongest chance rate. Nothing stronger is defensible: the claim is about information content of the panels, not yet about scorer-independent measurement or prediction.

Q3. Did the tests show exact identity recognition, functional-family recognition, or both?

Both, hierarchically. Family recognition was perfect (24/24) and appears to be the primary phenomenon; exact identity recognition (13/24) rides on top of it and succeeds when a society possesses year-exact, family-internal discriminators (Chile's 1973 and 2019; Sweden's 1909 strike and neutrality mobilisations; Afghanistan's 1929/1978/2001). Exact recognition fails, and falls back to family, when two family members share a crisis calendar (Italy/Greece, Iran/Egypt, New Zealand/Argentina).

Q4. Did CAMS appear to recognise labels, or did it recognise social-system morphology?

Morphology. The analyst never saw a label until the key; every identification was assembled from computed change-points, node signatures, and regime sequences. The decisive evidence is the error structure: a label-recogniser would fail randomly; a morphology-recogniser fails into neighbouring morphologies — which is exactly what happened, eleven times out of eleven.

Q5. How should the exact-hit rate across the three tests be interpreted?

As far above chance, and as heterogeneous for identifiable reasons. Set 1 (6/8) contained the most historiographically distinctive societies with a shared shock calendar; Set 2 (2/8) combined unfamiliar vintages, extreme scorer temperature (the Russia panel), heavy carry-forward, and an analyst still carrying candidate-pool bias; Set 3 (5/8), run under the corrected procedural rules, recovered most of the loss. The rate therefore measures the joint system of panel information plus analyst discipline, and it improved when the discipline improved.

Q6. What would the expected chance rate be under open-ended guessing?

With ~195 states and, realistically, ~60 polities plausible as century-long scoreable candidates, chance is ≈1/60 per guess: an expectation of ~0.4 exact hits in 24, with P(≥13) effectively zero. Even an absurdly generous 1-in-20 chance model expects 1.2 hits and leaves the observed 13 at $p < 10^{-9}$.

Q7. What would the expected chance rate be under a fixed eight-candidate forced-choice design?

1/8 per society: expectation 3/24, with $P(\geq 13) \approx 1.2 \times 10^{-6}$ (binomial). This is the strongest null the results should be quoted against, even though the actual task was harder than forced choice — the candidate list was never provided.

Q8. How should partial or family-level matches be scored?

With a pre-registered three-tier rubric: exact hit (1.0), family hit (a defensible shared functional class, defined before the key by the analyst's own ranked alternates — e.g. the true society appearing in the analyst's top-3, or the guess and truth sharing a named class; 0.5), miss (0). Under that rubric these tests score 13 + 11×0.5 = 18.5/24 (77%). Crucially, the family definitions here were constructed after the keys, which is why this report treats family results as a strong descriptive finding and not a pre-registered statistic; future tests must define the tiers in advance (Q36).

Q9. Were the errors random, or did they cluster into meaningful functional families?

Perfectly clustered. All eleven misses paired the true society with a structural sibling. Not one error crossed family lines — no floor state was mistaken for a hegemon, no rentier for an occupied industrial nation, no entrepôt for a garrison state.

Q10. What kinds of functional families appeared in the errors?

Eleven, each nameable: garrison-terminal readings (Japan↔USSR); high-stress rising developers (Brazil↔China); century-crisis floor states (Russia-as-scored↔Ethiopia); resource states in post-2014 collapse (South Africa↔Venezuela); low-activation poverty/violence developers (Colombia↔India); occupied 1945-nadir late-risers (Poland↔Philippines); MENA revolutionary states (Iran↔Egypt); Southeast-Asian states fallen in 1942 and stable after (Thailand↔Singapore); small resilient states under a great-power shadow (Finland↔Israel); Southern-European occupied debt-crisis states (Italy↔Greece); settler exporters in long relative decline (New Zealand↔Argentina).

Q11. Do structured errors suggest that CAMS is detecting a latent morphospace?

Tested — and refuted as originally stated (Results 3.3). The prediction that error pairs would be mutual nearest neighbours in a feature-space embedding failed on both testable pairs (Japan→Russia ranked 23/23; Brazil→China 17/23) and on the distinctiveness corollary ($p = 0.98$, wrong direction). The revised finding is two-part: a latent morphospace does exist in the panels — the label-free embedding spontaneously pairs Australia–Norway, UK–USA, Thailand–Hong Kong, Sweden–Finland, Russia–China — but the identification errors were generated upstream of it, in the analyst's space of historiographic story-types. The corpus embeds structure; the errors traced the analyst's priors across it.

4.2 What CAMS measures (Q12–Q22)

Q12. What does CAMS appear to measure: national identity, regime type, development level, institutional structure, path dependency, historical trauma, or distributed social cognition?

On this evidence: primarily path-dependent institutional trajectory — the timed sequence of shocks, mobilisations, recoveries, and reorganisations a society's institutional network has metabolised — with regime type and development level recoverable as by-products, and historical trauma legible as transition depth and hysteresis. National identity per se is not measured; it is inferred when a trajectory is unique. Distributed social cognition is the framework's interpretive layer over the s-field, and the tests show it is at least a productive description (Q20–21); they cannot show it is the mechanism.

Q13. Does s (cognitive activation, $s = A \cdot C/100 \cdot \max(K-S, 0.1)$) behave plausibly as a proxy for node cognition?

Plausibly, yes, within limits. The s-field produced signatures that were historically legible and diagnostically useful: Germany's Craft–Lore twin engine; Saudi Arabia's stewardship-flow with dormant Craft (the identification clincher); Hong Kong's flow-dominance under a permanently dormant Shield; pre-1948 statelessness visible as 49 years of Shield weakness in the Finland panel. It also behaved sensibly under catastrophe — civil channels going dark while a coercive core stays faintly lit in floor states. Limits: s is a constructed product of three scored quantities, so its plausibility partly inherits the scorer's plausibility; and it never yielded cross-node causal structure (all feedback nulls), so 'cognition' here means activation distribution, not demonstrated information processing.

Q14. Does S (Stress) behave plausibly as a proxy for node affect, activation load, or depletion?

Largely yes. S rose where history says load rose (wars, occupations, sanctions), suppressed activation multiplicatively in the bond operator exactly as a depletion term should, and produced the corpus's most extreme reading — Russia scored at $\bar{S} = 9.0$ for a century — which, whatever one thinks of that vintage's severity, is a coherent 'chronic allostatic load' portrait rather than noise. One caution: S is the dimension most exposed to scorer temperature; the same events can be scored as metabolised stress or as crushing stress, and that choice drove the largest identification failure (Q30).

Q15. Which nodes carried the strongest identity signal?

Shield and Hands, followed by Craft. Shield's timing is a war-and-mobilisation chronometer (Germany 1938–39, Japan 1937–41, Sweden's neutrality compensations, the UK's 1903–08 naval race) and its chronic level

encodes security posture (pacifist Japan, pre-state Finland, garrisoned Hong Kong). Hands' dormancy pattern is near-universal but its extremes are identifying (South Africa's 101 apartheid-era years). Craft identified Saudi Arabia almost single-handedly — a wealthy system that never built fabrication.

Q16. Which nodes carried the strongest family-level signal?

Archive, Lore, and Flow — the slow civil channels. Archive-dominance marked the old institutional cores (UK, Italy/Greece, France); Flow-dominance marked traders and entrepôts (Australia, Hong Kong, Thailand-as-read); Lore–Craft pairings marked the engineering-ideas states. These channels change slowly, so they define the family; the fast channels (Shield, Helm stress) define the individual.

Q17. Which node patterns were most likely to generate confusion between societies?

Shared-catastrophe patterns: simultaneous multi-node collapse in 1942–45 (Italy/Greece, Thailand/Singapore, Poland/Philippines); shared liberalisation-crisis shapes in the late 1980s (New Zealand/Argentina, Finland's scored 1987–88); and floor-state compression, where extreme S flattens all eight nodes and erases discriminators (Russia/Ethiopia). When history synchronises nodes, it synchronises fingerprints.

Q18. Did some societies appear easier to identify than others? Why?

Yes, sharply. Easiest: societies with unique, year-exact, family-internal ruptures — China (1949/1959/1966/1977/1992 is unshareable), Afghanistan, Chile, Sweden, Saudi Arabia, Iraq, Hong Kong, Germany. Hardest: societies whose panels were flattened (heavy carry-forward: Thailand at 89%), extreme in vintage temperature (Russia), or members of tight structural twins (Italy/Greece). Difficulty was a property of the panel's information content, not of the society's importance.

Q19. Did some families of societies appear easier to identify than others? Why?

The North Atlantic core and the intervention-broken floor states were easiest — the former because their shared shock calendar is densely documented and their within-family discriminators are famous, the latter because their catastrophes are unique in timing. Hardest were the mid-trajectory developers and structural twins, where several nations legitimately share one morphology. In other words: families with internal diversity identify well; families that exist because their members converged identify as families only.

Q20. What did the tests suggest about social cognition as a distributed phenomenon?

Three suggestive results. First, activation distribution is stable and characteristic — societies 'think' in persistent places (the s-signature survives wars and regime changes better than $\bar{V}$ does). Second, the distribution reorganises under catastrophe in patterned ways (civil channels darken before coercive ones; coupling-led vs viability-led recoveries). Third, no cross-node lead–lag structure was recoverable at annual resolution — so if distributed cognition involves information flow between nodes, it happens faster than the yearly sampling, or the framework's nodes are the wrong cut. The distributed-cognition reading survived the tests as a fertile description; it was not mechanistically confirmed.

Q21. In what sense can a society be said to 'think' through Archive, Lore, Helm, Shield, Flow, Craft, Hands and Stewards?

In the bounded, non-mystical sense the data supports: each node names a class of institutions that stores, transforms, and acts on information — memory (Archive), meaning and legitimation (Lore), executive direction (Helm), threat processing (Shield), exchange (Flow), fabrication (Craft), embodied labour (Hands), and resource custody (Stewards) — and the s-field measures where, at a given moment, coherent capacity is actually being applied above stress. A society 'thinks through' Flow when its adaptive responses are routed through exchange (Australia, Hong Kong), through Craft–Lore when routed through engineering and ideas (Germany). This is 'thinking' as distributed problem-solving under load, in the tradition of extended-cognition and superorganism arguments — a modelling stance the results reward, not an ontological claim they prove.

Q22. Does CAMS make 'national personality' measurable without reducing it to stereotype?

Partially, and the safeguard matters. What it measures is not personality-as-essence but signature-as-history: Saudi Arabia's dormant Craft is a documented consequence of rentier development, not an ethnic trait; South Africa's century of Hands failure is apartheid rendered as measurement. The stereotype risk enters through the human scorer (Q30), and the tests exposed it: one vintage's Russia is a century-long crisis, which is a defensible reading of state-society relations or an editorial temperature, depending on the reader. The framework's own discipline clause — morphology validates, magnitudes never pool — is precisely the anti-stereotype guard, and it worked when obeyed and cost identifications when it was needed most.

4.3 Validity (Q23–Q28)

Q23. What forms of validity are supported by the results?

Convergent/known-groups validity (panels converge on documented history at year resolution), discriminant validity at the family level (24 societies never confused across families), a form of criterion validity against historical record (blind recovery of identity is criterion prediction of a label), and face validity of the signature constructs. Content validity is partially supported: eight nodes sufficed to separate 24 very different polities.

Q24. What evidence is there for construct validity?

The constructs behaved as their theory says they should, in data the analyst had never seen: Shield rose at documented mobilisations and fell at documented disarmaments; the compensation construct fired exactly once in 24 societies, on the one armed neutral; the activation construct located Germany's cognition in Craft–Lore and Britain's in Archive; own-history attractor bands recovered the three regime shifts historiography would nominate (China upward from 2003, Japan out from 1998, South Africa out from 2014). Construct-consistent behaviour under blinding is the strongest construct evidence this design can produce.

Q25. What evidence is there for discriminant validity?

At family level, perfect: no floor state, rentier, entrepôt, hegemon, or occupied industrial nation was ever placed in another family. At individual level, discriminant validity is exactly what the 11 within-family misses fail to establish — the instrument discriminates between families much better than within them. That gradient is itself informative (Q11) but must be stated as the limit it is.

Q26. What evidence is there for diagnostic validity?

Moderate and specific. The regime classifier's outputs matched consensus periodisations blind (Freeze/Collapse lands on 1931 Australia, 1945 Germany/Japan-window, 2021 Afghanistan; Systemic Crisis on terminal Venezuela-morphology that proved to be South Africa's scored collapse). The early-warning split (signatures before endogenous crises, none before exogenous ones) is a diagnostic finding of real utility if it replicates. Diagnosis of current state, however, inherits scorer temperature: the same 2020s can be scored as strain or as collapse.

Q27. What evidence is there for predictive validity?

None yet, and the report must say so plainly. Everything here is retrodiction on scored history. The tests generated pre-registrable predictions (error-pair nearest-neighbourhood; EWS presence for endogenous transitions; open basin exits in the Japan and South Africa panels persisting or resolving), but no forward prediction has been tested. Predictive validity is the open front.

Q28. Which forms of validity remain unproven?

Predictive validity; inter-scorer reliability (all panels trace to one scoring lineage); scorer-independence of the identification signal (Q30); incremental validity over cheap covariates (Q39); within-family discriminant validity; and ecological validity of the node ontology itself (would a 6-node or 12-node cut do as well?).

4.4 Sceptical objections (Q29–Q33)

Q29. What are the strongest sceptical objections to the three blind tests?

(1) Scorer leakage, now precisely characterisable: the panels were scored by AI pipelines that knew each country — CAMSNATIONS5 ensembles, an agentic Kimi process, and Gemini — so identity-revealing events are inscribed by construction, and the test measures the invertibility of a history→CAMS→analyst channel rather than an independent property of societies. The mixed provenance matters: because at least three model families produced panels and (per the experiment owner) converge on the same structural arcs, the leakage channel is the shared historiographic corpus, not any single model's idiosyncrasy. (2) Analyst prior knowledge: the analyst is a language model saturated with world history; the achievement is history-matching, not measurement. (3) Narrative flexibility: with ~40 change-points per series and thousands of candidate events, a motivated matcher can dress many wrong answers convincingly — demonstrated live by the Israel-for-Finland error, which assembled seven 'year-exact hits' for the wrong country. (4) Post-hoc family definitions (Q8). (5) Single-analyst n; the single-scorer half of this objection is now retired by the multi-pipeline provenance, but its stronger successor stands: all AI scorers and the AI analyst share overlapping training corpora, so cross-model agreement cannot fully separate the societies' structure from the shared narrative's structure. All objections are fair; none is fatal to the bounded claim of Q2, and each is answerable by design (Q34–Q40).

Q30. Could the results reflect scorer leakage, prior historical expectation, narrative flexibility, or post-hoc interpretation?

Each deserves a separate verdict. Scorer leakage: real and now structurally understood — scorers (three model families) and analyst share overlapping training corpora, so the honest description is a compression round-trip through the common historiography: identity survives the CAMS encoding in 13/24 cases and degrades to family in the rest. Leakage still cannot explain the family-perfect error structure (an identity-inscribing channel would enable exact hits, not systematic sibling-misses), and the morphospace refutation adds a finer point: the misses tracked the shared narrative library, not the panels' feature space. Prior historical expectation: yes, in both directions — priors produced the Japan and Russia failures (levels-based rejection) as well as the successes; the Set-3 improvement after the rules banned levels-reasoning quantifies the prior's cost, not just its benefit. Narrative flexibility: demonstrated and measured — the Finland/Israel case shows a coherent seven-point wrong story is constructible; the countervailing evidence is that flexibility never rescued a cross-family error and that self-graded confidence tracked accuracy. Post-hoc interpretation: the identifications were locked pre-key (not post-hoc); the family taxonomy of the errors is post-hoc and is flagged as such throughout. Two additions the provenance forces. First, cross-family convergence is genuine replication of morphology: CAMSNATIONS5, Kimi, and Gemini recovering the same arcs defeats idiosyncratic scorer bias and empirically confirms the protocol clause that structure replicates across scorers while magnitudes do not pool (the severe Russia panel and the declinist terminal-Japan panel read as scorer temperature riding a shared skeleton). Second, convergence is still not neutrality: models trained on overlapping corpora agreeing on, say, a century of Russian systemic crisis certifies that the dominant narrative is being sampled consistently — precision, not accuracy — and the morphology/magnitudes quarantine exists exactly so that such editorial temperature, in any direction about any nation, cannot ride into downstream findings as measurement.

Q31. What would a hostile but fair reviewer say about the experiment?

'An LLM with encyclopaedic historical knowledge matched AI-scored vignettes back to the histories they were scored from — a loop closed through a shared training corpus even where model families differ — succeeding exactly where the vignettes were distinctive and failing into look-alikes where they were not. The scoring is single-lineage, the family rubric was defined after unblinding, one panel's severity (Russia) shows the pipeline transmits editorial judgment, and no comparison against trivial baselines — GDP series alone would date the Depression and 1945 too — was run. The honest findings are the perfect family structure of the errors, the

pre-registered nulls that came back null, and the protocol discipline. Rerun it with independent scorers, fixed candidate lists, pre-registered tiers, shuffled nulls, and covariate baselines before claiming the framework, rather than the analyst-plus-history system, is doing the work.' That review should be accepted and its programme adopted; it is Q34–Q40.

Q32. What would count as a decisive failure of CAMS in a future test?

Any of: identification at chance on panels scored by independent scorers blind to the test's purpose; family-level accuracy collapsing under a pre-registered family rubric; equal identification accuracy on shuffled-node or shuffled-year panels (proving the structure contributes nothing beyond event density); a cheap baseline (GDP + war-years + regime codings) matching CAMS identification accuracy; or inter-scorer signature correlations near zero, showing the 'signature' is the scorer, not the society.

Q33. What would count as strong confirmatory evidence in a future test?

Held identification accuracy on independently-scored panels; the pre-registered morphospace prediction confirmed (error pairs as mutual nearest neighbours in an embedding built before unblinding); blind recovery of hidden rupture years (Q40) at rates baselines cannot match; the endogenous/exogenous early-warning split replicating out of sample; and any successful forward prediction — e.g. the open basin exits (Japan-panel, South-Africa-panel) resolving as the framework's basin logic implies.

4.5 Future design (Q34–Q40)

Q34. How should future blind tests be designed?

As a factorial: {independent scorers × fixed candidate lists × pre-registered scoring tiers × shuffled null models × covariate baselines}, with the analyst's full computation log archived before key release, family taxonomies and morphospace embeddings pre-registered, and at least one condition where the analyst is a human panel rather than an AI, to separate framework information from analyst knowledge.

Q35. Should future tests use fixed candidate lists?

Yes, as one arm — a 30–60 name list per set makes the null exact and kills the open-ended-guessing ambiguity — while keeping an open-ended arm, because open-ended identification is the stronger result when it succeeds. Report both.

Q36. Should future tests include pre-registered criteria for exact hits, near hits and misses?

Non-negotiably yes. Exact = rank-1 match; near = truth within ranked top-3 or within a family taxonomy published before any data is seen; miss = neither. This experiment's family finding is real but post-hoc; pre-registration converts it from observation to statistic.

Q37. Should future tests include independent human and AI scorers?

Yes — and the AI half is now partially done: the disclosed provenance already constitutes a cross-model scorer replication (CAMSNATIONS5, Kimi, Gemini) in which structural arcs converge, satisfying the different-model-families arm. Within-pipeline ensembles (five passes of one model) remain pseudo-replicates that measure precision only; the decisive missing arm is human — two or more domain-expert scorers, mutually blind and blind to the test's purpose, on the same societies, since human experts draw on partially different evidence than any model corpus. Inter-scorer reliability on C/K/S/A then quantifies how much of the signature is the society; identification run separately on each scorer's panels tests whether the signal survives the scorer; and comparing human-vs-model envelopes on politically contested panels (Russia, China, the USA's 2020s) directly measures where corpus narrative substitutes for evidence. The protocol's own clause — cross-scorer data validates morphology, never magnitudes — predicts timing will replicate and levels will not: a falsifiable in-house prediction.

Q38. Should future tests include shuffled-node, shuffled-year and shuffled-metric null models?

Yes, all three. Shuffled-node panels test whether the eight-way institutional decomposition contributes beyond aggregate trajectory; shuffled-year (within-decade block shuffles) test how much identification depends on exact timing; shuffled-metric (recomputing V, s, B with permuted operator weights) tests whether the specific canonical operators matter or any monotone combination would do. Identification accuracy that survives shuffling indicts the corresponding component.

Q39. Should CAMS be compared against standard social-science baselines such as GDP, regime type, military spending, urbanisation, trade share or energy use?

Yes, and this is the incremental-validity test the current results most conspicuously lack. Give the same blind analyst a per-capita GDP series, Polity-style regime codes, military-expenditure share, urbanisation and trade ratios for the same 24 societies and score identification. The interesting outcome is any of the three: baselines match CAMS (CAMS adds narrative, not information); CAMS wins (the institutional decomposition carries unique signal — the Saudi Craft-dormancy and Sweden compensation findings suggest it will, since no GDP series contains them); or CAMS wins only at family level (the decomposition encodes type, timing encodes identity).

Q40. Should future tests ask CAMS to identify hidden rupture years or phase transitions?

Yes — it is a cleaner task than identity. Mask the calendar entirely, ask for rupture years, transition classes, and (where the EWS hypothesis applies) which transitions were internally brewed, then score against a pre-registered event list. It removes most narrative flexibility, needs no candidate list, and directly tests the framework's core claim of measuring adaptive-cycle structure.

4.6 What the overall report should say (Q41–Q50)

Q41. What should the overall report say about exact identity results?

State the number plainly with both nulls: 13/24 exact against an expectation of 3/24 under eight-candidate forced choice ($p \approx 1.2\times10^{-6}$) and <1/24 open-ended; note the per-set spread (6/2/5) and its causes (vintage temperature, carry-forward, analyst discipline); and attribute the result to the joint analyst-plus-panel system, not to the framework alone.

Q42. What should the overall report say about functional-family results?

That family-level identification was perfect across 24 societies and three very different ensembles, that this is the experiment's most robust finding, and that it is currently descriptive rather than pre-registered — with the rubric fix specified for the next round.

Q43. What should the overall report say about structured errors?

That the errors are the evidence, with the attribution the morphospace test forces: eleven misses, eleven within-family, zero random — but the embedding test (Results 3.3) shows these were nearest-neighbour errors in the analyst's narrative space, not the panels' feature space. Structured failure still rules out storytelling-as-noise; what it now demonstrates is a disciplined analyst-plus-history system degrading gracefully, while the panels' own similarity structure — recovered independently by the label-free embedding — is a separate, confirmed finding. Publish the error table, the refuted prediction, and the emergent adjacencies with equal prominence.

Q44. What should the overall report say about social cognition?

That the distributed-cognition framing earned its keep as a generative description — signatures were stable, historically legible, and diagnostically decisive in at least two identifications — while its mechanistic reading remains untested, because no cross-node information flow was recoverable at annual resolution. Say

'societies exhibit persistent, measurable activation distributions' rather than 'societies think', unless the extended sense of Q21 is spelled out.

Q45. What should the overall report say about path dependency?

That it is the medium of identification: every exact hit was a timed sequence of ruptures and recoveries unique to one polity, every attractor result (own-band returns, open exits, hysteresis everywhere examined) shows history constraining the reachable states. CAMS, on this evidence, is best summarised as a path-dependency instrument: it measures where a society's institutional network has been dragged, and how, and what shape that drag left.

Q46. What should the overall report say about parsimony?

Two things. The framework is parsimonious in the right place: eight nodes × four dimensions × locked operators separated 24 polities at family level perfectly — no additional machinery was needed. And parsimony must still be demonstrated against the cheaper alternative (Q39): until CAMS beats GDP-plus-regime-codes at the same task, the sceptic may fairly say the eight-node decomposition is elegant rather than necessary. The Craft/Shield/compensation findings are the current best evidence for necessity.

Q47. What should the overall report say about limitations?

Shared-corpus provenance — panels scored by multiple AI pipelines (CAMSNATIONS5, Kimi, Gemini) whose cross-family convergence retires idiosyncratic scorer bias but not the common-historiography confound, so no result here yet demonstrates corpus-independent measurement, and convergence must never be quoted as objectivity; scorer aware of identities; analyst saturated with historical priors (costs measured: three levels-reasoning failures); vintage temperature visible and consequential (the Russia panel drove the largest miss; the Japan panel's terminal severity drove another); 13 of 24 series above the 30% carry-forward threshold, gating temporal claims; family taxonomy post-hoc; no baselines; no forward prediction; n = 3 sets. Also the neutrality point, stated symmetrically: panels encode editorial judgments about contemporary great powers in both directions, and the protocol's morphology/magnitude quarantine is the necessary guard against importing any nation's caricature — hostile or flattering — as measurement.

Q48. What should the overall report say about next experimental steps?

In priority order: (1) independent-scorer replication (Q37); (2) pre-registered tiers, families, and the morphospace nearest-neighbour prediction (Q11, Q36); (3) covariate-baseline comparison (Q39); (4) shuffled-node/year/metric nulls (Q38); (5) hidden-rupture identification (Q40); (6) forward-prediction registry: file the open basin exits and EWS-class predictions now, score them in five years.

Q49. What is the most cautious conclusion supported by the three tests?

Hand-scored CAMS panels of twentieth-century societies contain sufficient structural information that a knowledgeable blind analyst can always recover the society's functional family and can recover its exact identity far above chance; the errors are systematic in a way consistent with a real underlying similarity structure. Nothing in the tests establishes that this information exceeds what conventional indicators carry, that it survives independent scoring, or that it predicts anything.

Q50. What is the strongest conclusion that can be stated without overstating the evidence?

CAMS, as operationalised in the JUNO panels and protocol, functions as a working morphological instrument for whole societies: it embeds polities in a recoverable structural similarity space in which functional families are cleanly separated, individual identity is resolvable wherever history has left year-exact discriminators, and failure is graceful — degrading to the nearest structural neighbour rather than to noise. Combined with its pre-registered nulls behaving as nulls and its one pre-registered positive (the armed-neutral compensation signature) firing on exactly one society in twenty-four, the three blind tests establish CAMS as an instrument worth the replication programme its own results now specify — and not yet more than that.

5. Conclusion

Across three blind sets and twenty-four societies, the CAMS/JUNO pipeline demonstrated recoverable social-system morphology: perfect family-level identification, exact identification at seven times the strongest chance expectation, a refuted-and-reported morphospace registration whose embedding nonetheless volunteered genuine structural families, cross-model scorer convergence on morphology (never magnitudes), disciplined null reporting, and one clean pre-registered epiphenomenon. The experiment's failures were as informative as its successes — every one traced to levels-reasoning, vintage temperature, unqueried anomalies, or candidate-pool bias, and each generated a corrective rule that measurably improved the next set. The path forward is specified, falsifiable, and short: independent scorers, pre-registered tiers, baselines, shuffled nulls, hidden ruptures, and a forward-prediction registry.

Appendices available as CSV exports: full society-year metric tables for all three sets (JUNO_set1/2/3_society_year_metrics.csv).