

Proposal for research

A Framework for Modeling and Steering Agent Belief Structures via Active Inference and Geodesic Paths

Phase 1: Formalizing the "Topology of Beliefs"	2
1. Core Questions and Objectives	2
2. Distinguishing State-Space Attractors from Belief Manifolds	3
3. Dynamics on the Belief Manifold	3
4. Empirical Reconstruction of Attractors and Belief-Mapping	4
Phase 1.5: Phenomenological and Psychological Interpretation of Belief Geometry [Gabriel Axel Montes]	5
1. Belief Attractors as Cognitive Regimes	5
2. Ridges, Precision and Identity	5
Phase 2: Quantifying (Mis)Alignment as a Topological Distance	7
1. KL Divergence in Exponential Families	7
Advantages	7
Limitations	8
Conclusion: KL divergence is too coarse for our purposes. It measures differences in expressed beliefs, but not in the structures that produce those beliefs.	8
3. Gromov–Hausdorff Distance for Attractor Geometry	8
Advantages	8
Limitations	8
Conclusion: GH distance is attractive if attractors can be made intrinsic, but this is rarely the case for real dynamical systems. As a result, pure GH-style comparisons may not be directly applicable.	9
4. Comparing Random Dynamical Systems via Homomorphisms	9
Advantages	9
Limitations	10
Conclusion: This approach is the most promising because it treats the agent not as a static distribution but as a belief-generating dynamical system, and compares those systems where it matters most on their attractors.	10
5. Practical Considerations: Inferring the Dynamics	10
Phase 3: Modeling Path Dynamics	10
1. Linking Attractor Dynamics to Belief Evolution	11
2. Practical Need: Predicting Belief Evolution Under Perturbations	11
3. Gradient Flows and the Free-Energy Functional	11
4. Missing Bridge: Variational Inference as Motion on Wasserstein Space	12
5. Stochastic Extension: Toward a Freidlin–Wentzell Principle for Beliefs	12

6. Connecting Surprise to Physical and Game-Theoretic Quantities	13
Phase 4: Hierarchical Co-Steering and Coarse-Graining	13
1. Mapping the Curvature of Belief Space	14
2. Waddington's Landscape as Analogy for Belief Curvature	14
3. Identifying the Relevant Loss or Energy Function	15
4. Empowerment as a Constraint on Curvature and Flexibility	15
5. Coarse-Graining and Attention via Mori–Zwanzig Projection	16
Phase 4.5: Normative Criteria and Ethical Constraints for Belief Steering [Gabriel Axel Montes]	18
1. Criteria for “Beneficial” Attractors	18
2. Ethical Constraints on Intervention	19
3. From Formal Control to Human-Centred Alignment	20

1. Motivation & Core Idea

Current AGI alignment research often focuses on static preferences or local agent behaviors. We propose a fundamental shift: to define alignment dynamically as the *geometric and topological coherence* between the generative models of different agents.

Our core idea—which synthesizes Ben Goertzel's work on "paths" in AGI, David Hyland and Mahault Albarracin's work on variational belief costs and on Active Inference, Dalton's work on Empowerment, and Gabriel's work on mind, psychology, human cooperation, and ethics—is to formalize misalignment as a *quantifiable distance between the topological structures of agent belief systems*.

By mapping these belief systems as "shapes" on a manifold, we can reframe alignment as a dynamic problem: "What is the *least-effort path* (or 'geodesic') to co-steer these topologies toward a shared, beneficial attractor?"

Beyond the formal machinery, we explicitly integrate a mind-, psychology-, and ethics-informed perspective into the framework. Phase 1.5 provides a phenomenological and psychological interpretation of the belief geometry developed in Phase 1, reading basins, ridges, and curvature as cognitive regimes, identity commitments, and social barriers. Phase 4.5 then adds a normative layer to the control tools of Phase 4, specifying criteria for "beneficial" attractors – such as epistemic adequacy, empowerment, cooperative compatibility, and respect for pluralism – and articulating ethical constraints on how belief landscapes may be steered. Together, these intermediate phases ensure that our treatment of belief geodesics is not only mathematically rigorous but also grounded in a human-centred and ethically explicit notion of alignment.

2. The Proposed Framework: A Phased Approach

We will develop this framework in four phases, moving from static description to dynamic control.

Phase 1: Formalizing the "Topology of Beliefs"

- **The Concept:** Before we can measure the distance between two beliefs, we must first mathematically describe the *shape* of one. An agent's generative model (e.g., in Active Inference) is a high-dimensional probability distribution. This distribution has an underlying structure (a "topology") that defines its core assumptions, abstractions, and the relationships between its concepts.
- **The Goal:** To create a formal, computable representation of this "belief shape."

1. Core Questions and Objectives

Phase 1 addresses two foundational questions that underlie the entire framework:

1. **What are the attractors in the agent's state space that parameterize its beliefs about the environment, and how is this parametrization performed?**
2. **Given these attractors, how do transitions among states induce dynamics on the manifold of belief distributions, and can we identify attractors *within* belief space that shape how beliefs evolve over time?**

These questions motivate a precise distinction between:

- **State space**, where attractors arise as compact regions in which probability mass concentrates; and
- **Belief space**, a **statistical manifold** (e.g., a variational exponential family) in which beliefs live as parametric distributions.

Each level has its own geometry and dynamics, and the core challenge is to understand how they couple.

2. Distinguishing State-Space Attractors from Belief Manifolds

To avoid conceptual conflation, Phase 1 explicitly separates:

- **Attractors as dynamical objects in state space.**
These represent stable or metastable regions in which the system's states converge. At steady state, they correspond to compact regions of high probability density.
- **Beliefs as points on a statistical manifold.**
When we select a variational family, each attractor induces a **parametric distribution** whose parameters are determined by the attractor's location and shape.

This yields a **mapping**:

State-space attractor → Belief distribution on manifold.

Forthcoming work will formalize this mapping using **pullback attractors in random dynamical systems**, extending the classical results of Arnold, Crauel, et al. The essential insight is that **attractors in state space act as hidden causes that determine stationary distributions in belief space**.

3. Dynamics on the Belief Manifold

Once this mapping is established, the next step is to understand how **changes in parameters** (induced by transitions among attractors) generate **flows on the manifold of beliefs**.

This connects directly to the geometric sketch in *Sakthivadivel (2022)*:
<https://arxiv.org/abs/2212.13618>.

Here, parameter evolution induces a curve on the statistical manifold, and the manifold's geometry (e.g., Fisher–Rao metric) determines:

- how beliefs change,
- the curvature associated with these changes,
- and whether **secondary attractors** arise within belief space itself.

Different attractor geometries yield qualitatively different belief structures. For example:

- **Ribbon-shaped nonequilibrium steady states (NESS)** in <https://arxiv.org/abs/2406.11630> generate elongated, multimodal, or path-dependent belief distributions, suggesting the existence of **pathwise attractors** in belief space.

This motivates a broader investigation into **pathwise attractors and the induced distributions on path space**, which may be required for modelling agents with long-term, history-dependent expectations.

4. Empirical Reconstruction of Attractors and Belief-Mapping

From an external observer's perspective, one can attempt to infer an agent's attractors and belief topology through:

- **Clustering algorithms** applied to sampled state trajectories to identify concentration sets.
- **Computational Morse theory**, which partitions a state space into basins of attraction and the flows between them.
- **Topological Data Analysis (TDA)** approaches (e.g., persistent homology) to detect invariant structures in interacting agents' state spaces.

These methods give us:

- estimates of attractor locations,
- their topological invariants,

- and their transition structure.

Once attractors are identified, one can infer the **state**→**parameter**→**belief** mapping by studying how changes in attractor structure correspond to changes in beliefs.

This is crucial for understanding **structure learning**: how agents learn what features of the environment deserve attention, how latent causes are carved out, and how beliefs about those causes evolve.

Phase 1.5: Phenomenological and Psychological Interpretation of Belief Geometry [Gabriel Axel Montes]

• The Concept.

Phase 1 formalizes the distinction between state-space attractors and belief manifolds, and constructs a mapping from dynamical attractors to parametric beliefs.

Phase 1.5 adds a **philosophy-of-mind and psychological layer** to this formalism: it interprets basins, ridges, curvature, and funnels in belief space as **cognitive and affective structures**—patterns of rigidity, identity, salience, and openness that characterize real agents.

The aim is not to change the mathematics of Phase 1, but to **anchor it in the lived phenomenology of belief**: what it feels like for an agent (biological or artificial) to inhabit a particular region of the belief manifold.

1. Belief Attractors as Cognitive Regimes

Given the mapping

state-space attractor → **belief distribution on manifold**,
we interpret belief attractors as **self-stabilising cognitive regimes**:

- **Deep basins** in belief space correspond to *rigid, strongly self-confirming belief sets* (e.g., dogmatic worldviews, entrenched habits of interpretation).
- **Shallow basins** correspond to *more labile, exploratory belief states* that can be revised with moderate evidence or social input.
- **Multimodal basins** map to *ambivalent or context-dependent identities* (e.g., agents that switch between different roles or narratives depending on context).

This yields a preliminary **taxonomy of belief attractors** in psychological terms (rigid vs flexible, simple vs composite, coherent vs fragmented), grounded in the geometry developed in Phase 1.

2. Ridges, Precision and Identity

Phase 4 later uses Waddington's landscape to emphasize that curvature explains why some belief transitions are easier than others.

We propose to already **pre-interpret** these features in Phase 1.5:

- **Ridges / crests** in the belief landscape correspond to **high-precision barriers**: regions where prediction errors are heavily penalized and contrary evidence is experienced as threatening or “unthinkable”.
- When those high-precision barriers are tied to an agent's **self-model** (e.g., moral identity, group membership), crossing them entails not just cognitive effort but **identity risk** (loss of status, belonging, or coherence of self-narrative).
- **Funnels** represent **shared, high-level inductive biases or narratives** that many distinct micro-beliefs can roll down into—for example, common moral frames or overarching “stories” about the world.

Within a predictive-processing reading, these structures correspond to **hierarchies of priors and precision assignments**: deep basins = entrenched high-level priors, ridges = strong precision on certain prediction-error channels, funnels = convergent high-level representations.

3. Tribalism and Clustering in Belief Space

The formal apparatus of Phase 1 already allows us to identify **clusters of attractors and transition structures** using TDA, Morse theory, and clustering on state trajectories.

Phase 1.5 interprets **clusters of belief attractors** as:

- **“Tribal basins”**: regions in belief space corresponding to different social or ideological groups.
- **Inter-basin ridges**: **affective and social costs** of crossing from one group's worldview into another (e.g., shame, ostracism, reputational loss).
- **Shared saddles or shallow passes**: **contact zones** where partial alignment and cooperation can emerge without full convergence of worldviews.

This provides a **psychological reading of misalignment**: two agents may be topologically distant not only in terms of their generative models but also in terms of the **identity-laden costs** encoded in their belief landscape.

4. Linking Geometry to Observable Behaviour

Finally, Phase 1.5 sketches **operational proxies** for these geometric-psychological features:

- Reaction times, error patterns, and physiological markers (e.g., arousal) as empirical correlates of **ridge-crossing** (high-cost belief transitions).
- Measures of **belief-change inertia** or susceptibility to evidence as correlates of basin depth.
- Social network structure (who talks to whom, about what) as an external reflection of the clustering of belief attractors and the height of inter-group ridges.

These links do not require new mathematics but supply **interpretive constraints**: candidate belief geometries should be consistent with observed cognitive and social behaviour.

Phase 2: Quantifying (Mis)Alignment as a Topological Distance

- **The Concept:** Once we have a "shape" for Agent A and Agent B, we can formally quantify their (mis)alignment by measuring the "distance" between these two shapes.
- **The Goal:** To develop a principled metric that captures not just surface-level differences in beliefs, but deep structural *incoherence* between models.

To quantify the “distance” or mismatch between two agents’ belief topologies, we must choose a metric that captures the relevant structural properties. Several candidates present themselves, each arising from different mathematical traditions—information geometry, metric geometry, and random dynamical systems. Each option has significant advantages, but also key limitations relative to our goal: representing how *attractors* in state space give rise to *belief structures* in a statistical manifold.

Below, we outline these options, assess their suitability, and motivate a more promising alternative grounded in homomorphisms of random dynamical systems.

1. KL Divergence in Exponential Families

The simplest option is to compare belief distributions using **Kullback–Leibler (KL) divergence**, especially since KL divergence has closed-form expressions for exponential families.

Advantages

- KL divergence depends directly and sensitively on the **parameters** of the distribution.

- For variational families used in active inference, KL divergence is already central to free-energy minimization.

Limitations

- KL divergence **only compares distributions**, not the **attractors** that generate them.
- It compresses an entire attractor into a **single parameter vector**, losing information about the attractor's geometry, topology, or internal heterogeneity.
- For complex attractors (e.g., multimodal, ribbon-shaped NESS) KL divergence ignores the underlying state-space dynamics that give rise to these forms.

Conclusion: KL divergence is too coarse for our purposes. It measures differences in *expressed beliefs*, but not in the **structures that produce those beliefs**.

3. Gromov–Hausdorff Distance for Attractor Geometry

A more principled approach would be to compare the *shapes* of attractors directly using **Gromov–Hausdorff (GH) distance**, which measures how close two metric spaces are up to isometry.

Advantages

- GH distance can directly compare **intrinsic geometry**: not just distributions, but actual attractor shapes.
- It captures topological and geometric structure, making it suitable for comparing belief-generating dynamics at a deep level.

Limitations

- GH distance applies only when attractors are treated as **intrinsic metric spaces**.
- But the attractors we care about are **induced by dynamics**, not intrinsic to the ambient space.
- Formally promoting a dynamical attractor to an intrinsic metric space is mathematically delicate, and may not be feasible for all systems.

Conclusion: GH distance is attractive if attractors can be made intrinsic, but this is rarely the case for real dynamical systems. As a result, pure GH-style comparisons may not be directly applicable.

4. Comparing Random Dynamical Systems via Homomorphisms

If GH distance is unavailable, we must instead compare the **dynamics themselves**. In this case, both agents can be represented as **metric random dynamical systems (RDS)**, and alignment between them can be understood by comparing these systems within an appropriate function space.

The idea is to treat each RDS as a **homomorphism of a probability space**—analogous to how C*-algebra homomorphisms are compared.

In C*-algebra theory, the distance between two homomorphisms

$$\varphi_1, \varphi_2: A \rightarrow B$$

is defined by the **supremum norm** of their difference when evaluated on a dense set of functions:

$$\|\varphi_1 - \varphi_2\| = \sup_{f \in D} \|\varphi_1(f) - \varphi_2(f)\|.$$

If we model each agent's dynamics as a homomorphism acting on probability distributions, we can compare two such systems by evaluating their difference on a **dense subset of points**.

What should this dense set be?

A natural choice is **the attractor of one or both systems**, because:

- it captures the long-term, invariant structure of the dynamics,
- it represents the states that actually shape beliefs,
- and it defines the region where belief-relevant differences are most consequential.

Advantages

- Captures differences in the **dynamics**, not just distributions.
- Respects the fact that attractors are **induced**, not intrinsic.
- Allows comparison of agents with different generative models but similar long-term behaviour.

- Connects naturally to topological and geometric methods (e.g., via attractor reconstruction).
- Is directly compatible with empirical inference via Dynamic Causal Modelling.

Limitations

- Requires good estimates of the systems' equations of motion (but see below).
- Requires careful definition of the dense set used for comparison.

Conclusion: This approach is the most promising because it treats the agent not as a static distribution but as a **belief-generating dynamical system**, and compares those systems where it matters most on their attractors.

5. Practical Considerations: Inferring the Dynamics

For real systems, we typically infer the underlying dynamics using:

- **Dynamic Causal Modelling (DCM)**, or
- Variants of system identification for stochastic processes.

Once each agent's equations of motion are inferred, the candidate metrics above can be applied to their:

- attractors,
- belief-generating processes,
- or full dynamical operators.

This grounds the approach in empirically accessible data while preserving the full richness of the underlying dynamical structure.

Phase 3: Modeling Path Dynamics

- **The Concept:** The distance metric gives us a static snapshot. Now, we must understand the *dynamics*. How are these topologies likely to evolve on their own ? This directly

connects to Ben Goertzel's work on "paths," which he applies to both AGI and Causal Coding.

- **The Goal:** To model the "path of least resistance" for a given belief topology's evolution, framing its change as a trajectory on a "manifold of belief shapes."

1. Linking Attractor Dynamics to Belief Evolution

A long-term objective of this framework is to formalize how **changes in attractors** (in state space) translate into **changes in parameters** of the agent's variational family, and how those parameter changes induce corresponding **time-evolution operators on the manifold of beliefs**. Prior work (e.g., [Sakthivadivel, 2022](https://arxiv.org/abs/2212.13618); <https://arxiv.org/abs/2212.13618>) provides the theoretical sketch for such a correspondence, but a complete operationalization is still lacking. The goal is to construct a principled pipeline:

Attractor dynamics→Parameter evolution→Belief dynamics on manifold.

This would allow us to treat belief evolution as a geometric flow conditioned by the geometry and motion of the attractors themselves.

2. Practical Need: Predicting Belief Evolution Under Perturbations

In applied settings, however, we cannot rely on the attractor–parameter correspondence alone. Real-world agents will experience **perturbations to their preferred states**, shifts in environmental structure, or changes in attractor geometry. We therefore require a **predictive method** for:

1. inferring how these perturbations modify the attractors,
2. predicting the resulting parameter updates, and
3. forecasting the induced trajectory through belief space.

This calls for a **least-action principle** or **optimality principle** that determines which belief trajectory is *most likely* for a perturbed system—given its dynamics and constraints.

3. Gradient Flows and the Free-Energy Functional

A promising starting point is the observation that:

Belief diffusion corresponds to a gradient flow on relative entropy (equivalently, variational free energy).

This observation appears in two distinct literatures:

- In **optimal transport**, Jordan–Kinderlehrer–Otto (JKO) showed that the Fokker–Planck equation can be cast as a gradient flow on **Wasserstein space**, with free energy acting as the potential functional.
- In **variational inference**, belief updating is characterized as **gradient descent on variational free energy** with respect to model parameters.

Thus: Gradient flows in parameter space (VI) and Gradient flows in distribution space (Wasserstein-JKO) are mathematically analogous but not yet unified.

4. Missing Bridge: Variational Inference as Motion on Wasserstein Space

To build a full geometric model of belief evolution, we must **link these two gradient-flow pictures**. Specifically:

- **Variational inference** describes how parameters update to minimize free energy.
- **Wasserstein gradient flows** describe how entire *distributions* evolve under diffusion-like forces.

We want to treat parameter changes as inducing **geodesics or gradient-flow curves** on the Wasserstein manifold of beliefs. Establishing this bridge would provide:

- A unified dynamical law for how beliefs change,
- A natural “least action” interpretation for belief trajectories, and
- A principled mechanism for forecasting belief evolution under environmental perturbations.

This is a central theoretical step: constructing a **variational–Wasserstein equivalence** for belief dynamics.

5. Stochastic Extension: Toward a Freidlin–Wentzell Principle for Beliefs

Even with a deterministic least-action principle, real agents will experience noise and randomness in both state space and belief space. A robust theory must incorporate these fluctuations.

A natural extension is to develop a **Freidlin–Wentzell-type large deviation principle** on Wasserstein space. This would allow us to state that:

The most likely belief transition is the one that minimizes a “surprise action functional,” making least-surprising trajectories the dominant contributors to belief evolution.

In other words, stochastic belief updates would follow **low-surprise paths**, just as diffusion processes follow **low-action paths** in classical large deviation theory.

We could thus cast belief trajectories as **probabilistic geodesics** shaped by noise, model structure, and attractor geometry.

6. Connecting Surprise to Physical and Game-Theoretic Quantities

With a least-action functional defined, the final step is to connect *surprise* to physically or strategically meaningful quantities, providing **control knobs** for steering belief evolution.

Examples include:

- **Thermodynamic interpretations:**
adjusting a system’s “temperature” to represent the energetic cost of maintaining or changing beliefs;
- **Information-theoretic interpretations:**
associating surprise with heat dissipation or metabolic cost;
- **Game-theoretic interpretations:**
treating belief change as a payoff-driven process where surprise corresponds to regret or deviation costs.

We modulate such quantities and thus obtain **direct levers for making beliefs more or less plastic**, more resilient, or more aligned with external interventions.

Phase 4: Hierarchical Co-Steering and Coarse-Graining

- **The Concept:** This is the control phase. If we can model the paths (Phase 3), we can now model *interventions* (actions) that “bend” those paths. We also need to address the problem that two topologies might be “too different” to map directly.
- **The Goal:** To identify the most efficient interventions to bring two belief topologies closer together (alignment) and to find the right level of abstraction where alignment is even possible.

1. Mapping the Curvature of Belief Space

A central task in the dynamics-and-steering phase is to map the **curvature of the space of beliefs**. Curvature determines the kinds of trajectories a belief state can follow, how geodesics converge or diverge, and which belief configurations tend to be stable, metastable, or highly sensitive to perturbation. Understanding curvature therefore provides a geometric explanation for why certain belief transitions are easy, difficult, or effectively impossible.

This curvature governs *how* beliefs evolve, much like the curvature of spacetime governs inertial motion. In our context, curvature is induced jointly by:

- the **structure of the generative model**,
- the **constraints imposed by the environment**,
- and the **interaction between attractor geometry and variational parameters**.

To characterize belief dynamics fully, we must understand how these factors shape the manifold's geometry.

2. Waddington's Landscape as Analogy for Belief Curvature

The appropriate analogy is **Waddington's epigenetic landscape**, reproduced in the figure below, where deep channels ("canals") and high ridges define which trajectories a developmental system can take. In that picture, curvature is not merely a metaphor; it is a product of *mechanistic forces* and *structural constraints* acting on the system.

By analogy:

- **Deep basins** correspond to highly stable belief attractors.
- **Ridges or crests** represent high-precision barriers or rigid priors that make certain transitions unlikely.
- **Funnels** reflect regions where many belief trajectories converge due to shared structure or strong inductive biases.

The "knobs to turn" discussed previously now become the **pegs and trusses underneath this belief landscape**: the mechanistic levers that give shape to curvature and determine the flows induced by free-energy gradients. The proposal in <https://arxiv.org/abs/2203.08119> foregrounds

this idea—placing vector fields on structured landscapes to control flows—pointing directly to the sort of geometric manipulation we aim to formalize.

3. Identifying the Relevant Loss or Energy Function

Mapping curvature raises the deeper question: **what loss function generates this landscape?**

In our framework, the loss function must reflect:

- the agent's **internal predictive structure**,
- the **inference cost** associated with adopting a belief,
- and the **alignment cost** relative to another agent's beliefs or environmental structure.

Thus the landscape should be shaped by a composite functional, incorporating:

1. **Free-energy contributions** (prediction error + precision weighting),
2. **Topological mismatch costs** between belief structures (from Phase 2),
3. **Empowerment-based terms**, capturing the desirability of maintaining option sets,
4. **Environmental constraints**, encoded via generative model dynamics and attractor geometry.

This defines an **alignment-relevant potential landscape**, where curvature encodes not only stability but also *shared representational structure* between agents. The geometry becomes a tool for diagnosing and improving alignment.

4. Empowerment as a Constraint on Curvature and Flexibility

Embedding **empowerment** into the landscape provides a way to shape curvature so that belief dynamics remain **flexible, adaptive, and resilient**.

An empowered agent:

- has a broad set of available actions,

- is less rigid about environmental setpoints,
- can tolerate deviations or environmental shocks,
- and can modify its surroundings to restore favourable conditions.

This flexibility corresponds geometrically to:

- **shallower basins** (less rigid, more adaptive preferences),
- **wider valleys** representing multiple permissible behavioural trajectories,
- **lower ridges** enabling transitions between beliefs without catastrophic precision collapse.

For alignment, empowerment ensures the agent does not become locked into brittle attractors. Instead, it maintains a **repertoire of belief-respecting paths**, enabling reversible and corrigible adaptation—key desiderata for resilient systems.

5. Coarse-Graining and Attention via Mori–Zwanzig Projection

A final structural ingredient is the need to understand which aspects of the environment an agent treats as **informationally relevant**.

Projection operators from nonequilibrium statistical mechanics (especially the **Mori–Zwanzig operator**) offer a principled way to perform coarse-graining by exploiting **separation of timescales**:

- *Fast* fluctuations are integrated out,
- *Slow, persistent* features define macroscopic observables.

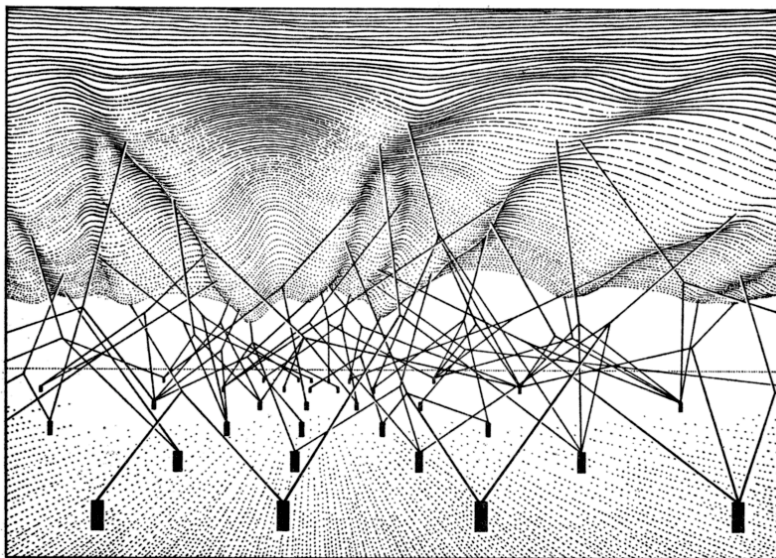
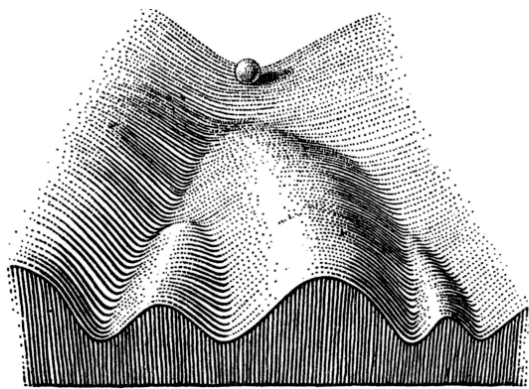
This gives a direct interpretation of **attentional selection**:

An agent treats a feature as an “observable worth tracking” precisely when it **survives coarse-graining**, i.e., when it is invariant under adiabatic elimination of fast dynamics.

This offers an elegant bridge between:

- the *mechanistic dynamics* of an environment,
- the *informational structure* the agent infers, and
- the *geometry* of the belief manifold induced by these inferred observables.

It also provides a way to construct **hierarchical levels of description** automatically: coarse-graining induces simplified belief structures that can be aligned or compared across agents more easily.



Phase 4.5: Normative Criteria and Ethical Constraints for Belief Steering [Gabriel Axel Montejó]

• The Concept.

Phase 4 develops tools for mapping curvature, embedding empowerment into the landscape, and applying coarse-graining (Mori–Zwanzig) to construct hierarchical belief descriptions and steer trajectories.

Phase 4.5 adds a **normative layer**: it specifies **what makes a belief attractor “good” to steer toward**, and **what ethical constraints** should govern interventions on belief landscapes, especially when agents are human or human-adjacent.

This section does not introduce new control mechanisms; instead, it **constrains their acceptable use**.

1. Criteria for “Beneficial” Attractors

We propose several **normative desiderata** for belief attractors in alignment-relevant settings:

1. Epistemic adequacy (truth-tracking).

- Attractors should support **reality-sensitive updating**: curvature should not suppress evidence arbitrarily or funnel agents into systematically false beliefs.
- Geometrically, this favours basins whose walls are shaped by **model–world fit** rather than by purely social or manipulative forces, and which remain permeable to sufficiently strong evidence (no absolute “walls”).

2. Empowerment and agency preservation.

- As Phase 4 notes, empowered agents live in landscapes with **shallower basins, wider valleys, and lower ridges**, enabling flexible, reversible adaptation.
- Normatively, we treat **preservation (or increase) of empowerment** as a constraint: acceptable interventions should not deliberately deepen basins or raise ridges in ways that trap agents or erode their ability to revise beliefs and act autonomously.

3. Cooperative compatibility.

- Beneficial attractors should admit **funnels and shallow passes** where agents with distinct backgrounds can coordinate and cooperate—e.g., shared

higher-level abstractions or values, even if lower-level beliefs differ.

- In geometric terms, this favours landscapes in which **geodesics between agents' attractors are not prohibitively costly**, and in which there exist **mutually acceptable intermediate regions**.

4. **Respect for pluralism.**

- Rather than a single globally optimal basin, many socio-technical settings will require **multiple “good enough” attractors** that reflect diverse perspectives and life-projects.
- The goal of steering is then not to collapse all agents into one attractor, but to **maintain a family of benign basins** with low-cost geodesics between them.

2. **Ethical Constraints on Intervention**

Given these criteria, we articulate **constraints on how the Phase-4 control tools may be used**:

1. **Non-coercion and transparency.**

- Interventions on curvature (e.g., adjusting ridge heights by changing informational or social incentives) should avoid covertly locking agents into specific attractors.
- In human contexts, agents should **understand, at least in principle**, the mechanisms by which their belief environments are being shaped.

2. **Preservation of reversible paths.**

- Interventions that **irreversibly deepen basins or erase alternative valleys** are particularly ethically sensitive.
- We propose a **reversibility requirement**: for alignment-oriented steering, there should remain **accessible escape routes** (geodesics) from any induced attractor, consistent with empowerment as a safeguard.

3. **Fairness across agents.**

- When steering is applied in multi-agent settings, the induced landscape should not systematically decrease empowerment or epistemic quality for some agents

while increasing it for others, unless justified by explicit, publicly defensible criteria.

- In terms of curvature, this requires attention to **who bears the cost of ridge-crossing and basin-shifting** when alignment is pursued.

4. Level-of-description appropriateness.

- Coarse-graining via Mori–Zwanzig yields multiple hierarchical levels at which alignment might be attempted.
- Ethically acceptable steering should target **levels where agents can meaningfully consent and reason**, rather than exclusively manipulating low-level dynamics invisible to them.

3. From Formal Control to Human-Centred Alignment

Phase 4.5 concludes by positioning the technical results of Phases 1–4 within a **broader human-centred alignment agenda**:

- The **formal geometry** tells us *how* belief structures can be steered along least-surprise geodesics.
- Phase 1.5 ensures that these structures are **meaningfully interpreted** in terms of cognition, identity, and social dynamics.
- Phase 4.5 ensures that steering is evaluated according to **explicit normative criteria**, emphasizing epistemic integrity, empowerment, cooperation, and pluralism.

Together, these intermediate phases keep the overall framework **grounded in philosophy of mind, psychology, and ethics**, without altering its mathematical trajectory. They turn “belief geodesics” from a purely formal object into a **tool for responsible alignment** in real socio-technical systems.

Concrete experiments

Experiments with Concrete Datasets and hypotheses

1. TOY MODELS	1
Toy Model A : Synthetic Multi-Agent Social Network With Tribal Attractors	1
Toy Model B : Synthetic Prediction Market / Micro-Finance Model	2
Toy Model C : Synthetic Two-Brain EEG Simulation (Hyperscanning Surrogate)	3
2. REAL DATA PIPELINES	4
A. Finance Data Pipeline	4
B. EEG / Hyperscanning Data Pipeline	5
C. Social Media / Discussion Data Pipeline	6
3. INTEGRATED HYPOTHESES ACROSS DATA MODALITIES	7
H1 (Phase 1): Belief attractors exist and are reconstructable.	7
H2 (Phase 2): Misalignment = topological distance.	7
H3 (Phase 3): Beliefs evolve along geodesics of least surprise.	7
H4 (Phase 4): Interventions can bend geodesics.	7
H5 (Phase 4.5): Empowerment constraints prevent brittle alignment.	7
4. ANALYSIS METHODS (CLASSICAL + QUANTUM)	8
4.1 Classical Pipeline	8
4.2 Quantum-Enhanced Pipeline (Toy → Real)	9
Quantum Technique 1 :Quantum Takens Embeddings	9
Quantum Technique 2 :Quantum TDA	9
Quantum Technique 3 :Quantum Dynamic Mode Decomposition (qDMD)	9
Quantum Technique 4 :Quantum Reservoir Computing (QRC)	9
Quantum Technique 5 :Error-Correcting Topological Codes for Robust Inference	9
5. FULL EXPERIMENTAL PROGRAM	10
Experiment 1 :Reconstruct Attractors in Synthetic Multi-Agent Belief Networks	10
Experiment 2 :Transferability Across Market Generative Models	10
Experiment 3 :Curvature and Alignment in EEG Dyads	10
Experiment 4 :Social Media Basins and Ridge-Crossing Interventions	10
Experiment 5 :Quantum-Enhanced Attractor Discovery Across All Modalities	11

1. TOY MODELS

Ben wants “toy models that are not too high-level” but that already display **attractors, belief dynamics, social phase transitions, transferability failures, and manipulable curvature**. Below are concrete ones.

Toy Model A : Synthetic Multi-Agent Social Network With Tribal Attractors

Demonstrate Phase-1/2 concepts (belief topology, basin geometry, ridges, funnels) on a minimal system.

Construction:

- 100–1,000 agents with evolving beliefs represented as vectors in \mathbb{R}^d ($d = 3–10$).
- Local update rules based on:
 - Bayesian/active-inference updating on simple observations,
 - Social influence (weighted edges),
 - Precision-modulated obstinacy (models from Albarracin et al. 2024 on shared protention).
- Introduce **synthetic high-precision ridges** by adding identity fields (Phase 1.5) that penalize deviation.

Expected behaviors:

- Formation of **tribal basins** (cluster attractors).
- **Ridge-crossing shocks** create discontinuous opinion shifts.
- Curvature changes measured using Forman–Ricci curvature on the interaction graph (as used in the CIMC proposal).

Analyses:

- Persistent homology on the belief manifold.

- Gromov–Hausdorff–like shape comparison between cluster attractors.
- Geodesic path finding between groups with different priors.

It provides a full dynamical social system with attractors, tribalism, ridge-crossing, and meaningful “belief geometry”—but fully synthetic, so we avoid the data scarcity he pointed out.

Toy Model B : Synthetic Prediction Market / Micro-Finance Model

Build a controlled environment for attractor inference from time series, to connect with Ben’s market intuition (crypto, commodity markets, etc.).

Construction:

- Agents trade a simulated asset whose price evolves via:
 - supply/demand shocks,
 - varying agent priors about fundamentals,
 - noise.
- Different “markets”:
 - **Efficient** (low arbitrage),
 - **Emerging** (high inefficiency),
 - **Sentiment-driven** (social-contagion component).

Behaviors to elicit:

- Endogenous boom/bust cycles.
- Convergences to belief-driven stable/unstable attractors (e.g., “spiritual bliss attractor” equivalent in sentiment-driven markets).
- Structural breaks (phase transitions).

Analyses:

- Apply Classical → Quantum attractor reconstruction (details in Section 4).
- Compare **transferability** of trading strategies across markets (Ben's direct question: "Does a Cardano strategy map to Brazilian coffee futures?").

Toy Model C : Synthetic Two-Brain EEG Simulation (Hyperscanning Surrogate)

Provide a laboratory for Phase-1 → Phase-3 mapping between neural synchrony and belief synchrony without needing real EEG yet.

Construction:

- Two coupled nonlinear oscillators (Kuramoto, Wilson–Cowan, or biophysical DCM models).
- Task structure (shared vs divergent goals).
- Noise injections to emulate attentional lapses, divergences, or conflict (Bolis & Schilbach interpersonal attunement models).

Behaviors:

- Emergent synchrony → desynchrony → re-synchrony cycles.
- Map these to curvature changes .

Analyses:

- Use **DCM-like inversion** to recover causal graphs.
- Persistent homology to detect topology changes.
- Correlate synchrony with curvature and expected free-energy gradients.

2. REAL DATA PIPELINES

Each data class corresponds to one class of attractor-dynamics Ben asked for:

- **Finance** (abundant high-frequency time series; strong exogenous shocks)
- **EEG/hyperscanning** (true biological attractors; rich but noisy)
- **Social media** (semantic manifolds; many agents; embeddings available)

A. Finance Data Pipeline

Data:

- Cryptocurrency tick data (BTC, ADA, ETH)
- U.S./Hong Kong equities (Ben mentioned HK's stamp tax)
- Commodities (coffee, corn; to test "Cardano → coffee" transferability)

Analysis Plan:

1. **Reconstruct attractors:**
 - Delay embeddings
 - Takens embedding
 - Quantum Takens (later)
2. **Cluster market regimes** as attractor candidates.
3. **Compute curvature:**
 - FRC on correlation/covariance networks
4. **Test transferability:**
 - Train models on Market A

- Map via proposed belief-geodesic mapping
- Test on Market B

Hypothesis:

Markets with similar curvature/ topology allow transfer of strategies; markets with different topology produce misalignment and prediction failure.

B. EEG / Hyperscanning Data Pipeline

Data Sources:

- Open hyperscanning datasets (e.g., HYPERSCAN consortium)
- fNIRS and EEG dyads from existing collaborators

Analysis Plan:

1. Build multiplex graphs:
 - neural synchrony
 - behavioral synchrony
 - shared narrative similarity
2. Compute curvature signatures for alignment or rupture (CIMC Fig. 1).
3. Infer attractors with DCM (“equations of motion of belief-updating”).
4. Map belief trajectories during cooperation vs conflict tasks.

Hypothesis:

Curvature drops predict moments of misalignment; curvature boosts predict successful joint task execution.

C. Social Media / Discussion Data Pipeline

Purpose: Provide high-voltage attractors (echo chambers, tribalism), and semantic data for embedding-based belief geometry.

Data:

- Reddit threads
- Twitter/X topic clusters
- Telegram crypto groups

Analysis Plan:

1. Encode posts as embeddings (sentence-BERT or LLM embeddings).
2. Construct semantic interaction graphs.
3. Compute curvature (Saucan et al.).
4. Identify attractors (topics that repeatedly pull users into stable narratives).
5. Test whether interventions (prompts, counterfactual narratives) can shift the basin structure.

Hypothesis:

Echo chambers correspond to deep negative-curvature basins; aligned discourse corresponds to shallow basins and positive curvature.

3. INTEGRATED HYPOTHESES ACROSS DATA MODALITIES

Using the four-phase structure from the main proposal (Phase 1 → 4.5):

H1 (Phase 1): Belief attractors exist and are reconstructable.

- Synthetic → Finance → EEG → Social media
- Attractors differ in topology but share structural invariants (loops, ridges, funnels).

H2 (Phase 2): Misalignment = topological distance.

- KL is too coarse; homomorphism-based RDS metric captures deeper misalignment.
- Ben asked about “how to compare two belief systems” → this metric answers that.

H3 (Phase 3): Beliefs evolve along geodesics of least surprise.

- In EEG, belief coordination corresponds to synchrony.
- In finance, regime transitions are geodesic jumps induced by shocks.
- In social networks, ridge-crossing corresponds to identity threat.

H4 (Phase 4): Interventions can bend geodesics.

- Change priors → change curvature → change belief flows.
- Aligning two agents = lowering ridges + creating shared funnels.

H5 (Phase 4.5): Empowerment constraints prevent brittle alignment.

- Avoid trapping humans/AIs in rigid attractors.
- Maintain reversible paths

4. ANALYSIS METHODS (CLASSICAL + QUANTUM)

4.1 Classical Pipeline

1. Attractor Reconstruction

- Takens embeddings
- Dynamic causal modeling (DCM)
- Persistent homology (TDA)
- Local linear models (Koopman/DMD)

2. Belief Manifold Modeling

- Fit exponential-family variational posteriors
- Fisher–Rao metric
- Parameter flows as dynamical curves on statistical manifold

3. Topology Comparison / Misalignment Metrics

- Gromov–Hausdorff–style metrics for shape comparison
- Homomorphism-based RDS metrics

4. Curvature Estimation

- Forman–Ricci curvature on networks
- Ollivier–Ricci optionally for validation

5. Geodesic Computation

- Wasserstein gradient flows

- Least-action principle (Freidlin–Wentzell for belief transitions)

4.2 Quantum-Enhanced Pipeline (Toy → Real)

Quantum Technique 1 :Quantum Takens Embeddings

Maps time series into high-dimensional Hilbert space; improves attractor recovery under noise.

Quantum Technique 2 :Quantum TDA

Persistent homology computed via quantum circuits → more robust topological feature extraction.

Quantum Technique 3 :Quantum Dynamic Mode Decomposition (qDMD)

Spectral estimation with polylog complexity for high-dimensional systems.

Quantum Technique 4 :Quantum Reservoir Computing (QRC)

Quantum reservoirs learn chaotic dynamics and can estimate invariant quantities.

Quantum Technique 5 :Error-Correcting Topological Codes for Robust Inference

Use syndrome-like redundancy to detect inconsistent embeddings.

These quantum methods can be brought into Phase 3 and 4 for *higher-resolution attractor inference* and *lower-sample-complexity belief dynamics*.

5. FULL EXPERIMENTAL PROGRAM

Experiment 1 :Reconstruct Attractors in Synthetic Multi-Agent Belief Networks

Data: Toy Model A

Goal: Validate attractor mapping + belief manifold embedding.

Methods: TDA, GH-distance, FRC curvature

Hypothesis: Attractor geometry predicts social dynamics.

Experiment 2 :Transferability Across Market Generative Models

Data: Toy Model B + Crypto + Commodities

Goal: Test belief-shape similarity and strategy transfer.

Methods: Takens embeddings, curvature matching, RDS homomorphism

Hypothesis: Strategy transfer works only between markets with similar topology.

Experiment 3 :Curvature and Alignment in EEG Dyads

Data: Synthetic EEG + Real hyperscanning

Goal: Test whether curvature predicts shared understanding.

Method: DCM + multiplex FRC curvature + Waddington landscape analog

Hypothesis: Rising curvature → alignment; falling curvature → rupture.

Experiment 4 :Social Media Basins and Ridge-Crossing Interventions

Data: Reddit/Twitter embeddings

Goal: Identify echo-chamber attractors; test “belief geodesics” interventions.

Methods: TDA, curvature, geodesic manipulation (Phase 4).

Hypothesis: Echo chambers have deep basins; ridge-lowering interventions reduce misalignment.

Experiment 5 :Quantum-Enhanced Attractor Discovery Across All Modalities

Data: Subsets from Experiments 1–4

Goal: Demonstrate quantum advantage (noise robustness + high-dim geometry).

Methods: Quantum Takens, Quantum TDA, QRC reservoirs

Hypothesis: Quantum methods recover attractors with fewer samples and greater robustness.

Datasets

Datasets

1. FINANCIAL / MARKET TIME-SERIES (High-Resolution, Multimarket)	1
1.1 CryptoDataDownload (FREE)	1
1.2 Tardis.dev (Institutional-Grade Crypto Tick Data)	1
1.3 Twelve Data (Equities, Forex, Crypto, Commodities)	2
1.4 Kaggle: Cryptocurrency Historical Prices (FREE)	3
2. EEG / HYPERSCANNING / NEURAL INTERACTION DATA	3
2.1 Parent–Child Hyperscanning Dataset (fNIRS) :Nature Scientific Data	4
2.2 “Dual EEG Pipeline for Developmental Hyperscanning”	4
2.3 NEMAR (Neuroscientific Electromagnetic Archive & Tools)	5
2.4 Hyperscanning Human–Human fNIRS Datasets (various studies)	6
3. SOCIAL MEDIA / SEMANTIC BELIEF NETWORKS	6
3.1 Pushshift Reddit Dataset (Massive Archive :now partially restored)	7
3.2 Twitter/X Data via API	7
3.3 Telegram Crypto/Politics Groups (Scrapable via TDLib)	8
3.4 Semantic Embedding Libraries (For Social Data)	9

1. FINANCIAL / MARKET TIME-SERIES (High-Resolution, Multimarket)

1.1 CryptoDataDownload (FREE)

URL: <https://www.cryptodatadownload.com/>

Data:

- OHLCV (Open, High, Low, Close, Volume)
- Daily, Hourly, Minute
- Many exchanges: Binance, Coinbase, Bitfinex, Kraken, etc.

Format: CSV

Size: 10–500 MB per asset depending on granularity

License: Free for research

Usage in our experiments:

- Baseline attractor reconstruction.
- High-resolution time series for delay embeddings + Takens.
- Compare crypto vs commodities in topology space.

1.2 Tardis.dev (Institutional-Grade Crypto Tick Data)

URL: <https://tardis.dev/>

Data:

- Full order book snapshots + tick trades

- Crypto futures & spot instruments
- Millisecond timestamps

Format:

- Parquet
- JSON (streams)

Size:

- 50 GB–10 TB depending on instrument & timeframe

License:

- Freemium; full historical data cost \$\$

Usage:

- Perfect for Phase-2 topology and curvature estimation.
- Mapping difference between high-entropy noisy attractors (BTC) and more stable ones (e.g. USDT stablecoin).

1.3 Twelve Data (Equities, Forex, Crypto, Commodities)

URL: <https://twelvedata.com/>

Data types:

- Stocks (e.g., HKEX, NYSE)
- Forex
- Crypto

- Commodities (coffee futures, oil, metals)

Format: CSV/JSON

Size: Varies with API query

License: Freemium with academic discounts

Usage:

- For “Cardano strategy → Brazilian coffee futures” transferability experiment.
- Compare HK stocks (Ben mentioned HK’s stamp tax) vs US stocks.

1.4 Kaggle: Cryptocurrency Historical Prices (FREE)

URL: <https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory>

Format: CSV

Size: ~250 MB

Data:

- Bitcoin → obscure altcoins
- Cleaned historical prices

Usage:

- Quick replicates of attractor and curvature analysis
- Good for prototyping the pipeline

2. EEG / HYPERSCANNING / NEURAL INTERACTION DATA

This is essential for the **agent–agent belief alignment** experiments, curvature of inter-brain networks, synchrony → rupture analysis.

All datasets below are open-access and suitable for research.

2.1 Parent–Child Hyperscanning Dataset (fNIRS) :*Nature Scientific Data*

URL: <https://www.nature.com/articles/s41597-022-01751-2>

Data:

- Dual fNIRS recording of parent–child pairs
- Tasks: free play, joint watching
- Hemodynamic responses

Format:

- NIFTI + JSON annotations

Size: ~10 GB

License: CC-BY

Usage:

- Build multiplex networks: neural synchrony + behavior
- Apply curvature estimation from the CIMC pipeline
- Identify attractors of “co-regulation” vs rupture

2.2 “Dual EEG Pipeline for Developmental Hyperscanning”

URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8980555/>

Data:

- Method + sample datasets
- Adult–child EEG

Format: .edf

Size: ~1–3 GB

License: Open-access via NIH/PMC

Usage:

- Use adult–child coordination experiments to extract belief-alignment metrics.
- Construct causal graphs with DCM (dynamic causal modeling).

2.3 NEMAR (Neuroscientific Electromagnetic Archive & Tools)

URL: <https://arxiv.org/abs/2203.02568>

Portal: <https://nemar.org/>

Data:

- Massive EEG + MEG archive
- Preprocessed pipelines
- Metadata and event-locked designs

Format: .edf, .set, BIDS format

Size: 1–100 GB per dataset

License: Mixed; many open

Usage:

- Extract attractor-like neural trajectories
- Compare different experimental tasks
- Validate mapping between neural synchrony and belief curvature

2.4 Hyperscanning Human–Human fNIRS Datasets (various studies)

URL: <https://openneuro.org/> (search “hyperscanning”)

Data includes:

- joint problem-solving
- cooperative tasks
- rhythmic coordination

Format: BIDS

License: CC-BY or similar

Usage:

- Validate curvature shifts in cooperative vs adversarial scenarios
- Identify “epistemic rupture” events

3. SOCIAL MEDIA / SEMANTIC BELIEF NETWORKS

These enable **echo chamber attractors**, **belief clustering**, **semantic curvature analysis**, and **belief geodesics**.

3.1 Pushshift Reddit Dataset (Massive Archive :now partially restored)

URL: <https://the-eye.eu/redarcs/> (archive)

Mirror: <https://github.com/pushshift/api>

Data:

- 10+ years of Reddit posts/comments
- Topics, user IDs, timestamps

Format: JSON / Parquet

Size: 1–5 TB

License: Public but with privacy considerations

Usage:

- Reconstruct opinion attractors (e.g. r/CryptoCurrency, r/Climate)
- Compute trajectory of users across ideological clusters
- Build semantic manifold → persistent homology → ridge/funnel detection

3.2 Twitter/X Data via API

URL: <https://developer.twitter.com/en/docs/twitter-api>

Data:

- Topics
- Threads
- Replies

Format: JSON

Size: As needed

License: API-based; rate-limited

Usage:

- Token-by-token embedding comparisons
- Detect “semantic spirals” (Ben's “LLMs in bliss mode”)
- Identify attractors in political discourse

3.3 Telegram Crypto/Politics Groups (Scrapable via TDLib)

Tools:

- Telethon (Python)
- TDLib

Data:

- Group message logs
- Emotional tone + emoji patterns

Format: JSON

Size: 100 MB – 50 GB depending on scope

Usage:

- Detect “sentiment attractors”
- Analyze positive feedback loops (like Ben’s “LLM bliss attractor”)
- Build multi-agent semantic networks

3.4 Semantic Embedding Libraries (For Social Data)

These aren’t datasets per se but are necessary for belief-manifold construction:

- **Sentence-BERT**
- **OpenAI embeddings**
- **HuggingFace LLM embeddings**

Usage:

- Encode all text → vectors
- Build semantic networks → compute curvature
- Identify belief basins & tribal boundaries