# Finding a Frame through Strategic Abstraction

**Sandy Tanwisuth, Niklas Lauffer**

**Keywords:** strategic abstraction, multi-agent coordination, soft best response, contrastive learning, bounded rationality, epistemic reasoning, general-sum games, representation learning under partial observability

## Summary

Coordination in real-world, mixed-motive environments often fails not because goals are misaligned, but because agents lack a shared understanding of which distinctions in others' behavior are relevant to their own decisions. This paper introduces a theory of agentic coordination in which agents do not model others in full. Instead, they learn to refine *abstraction frames* defined as structured filters that isolate behavior patterns with strategic influence. We operationalize this theory through a three-layer architecture. The **Influence Filter** determines whether a partner's actions consistently reshape the ego agent's soft best response. If so, **Short-horizon Compatibility** evaluates divergence between the resulting best response landscapes across candidate co-policies. A third layer, **Long-horizon Compatibility**, is proposed to capture long-term convergence in inferred incentives across extended trajectories. Together, these layers define soft decision boundaries over relevance, approximating the Value of Information and Value of Opportunity decomposition from Bayes-Adaptive MDPs. Rather than maintaining beliefs over full co-policies or latent states, agents maintain distributions over abstraction frames which are compact representations of the distinctions their own behavior must condition on to coordinate. We formalize this refinement process with **Strategic InfoNCE**, a contrastive learning objective that embeds co-policies based on their induced incentive gradients. A convergence signal, denoted by $\Delta\mathrm{SEC}(t)$, tracks entropy reduction over abstraction disagreement and marks epistemic readiness for coordination. In this view, coordination does not depend on behavioral mimicry, identity modeling, or shared intentions. It emerges through alignment over what distinctions are strategically relevant. *Strategic Abstractions* offers a tractable, cognitively inspired framework for finding the frame, enabling agents to learn how to act together by resolving what matters.

## Contribution(s)

1. We reframe coordination as abstraction refinement over strategic influence.
   **Context:** Agents evaluate which distinctions in co-agent behavior meaningfully alter their own incentives, inspired by Strategic Equivalence Relations Lauffer et al. (2023) under bounded rationality.
2. We propose a three-layered filtering framework: Influence Filter, Short-horizon Compatibility, and a speculative Long-horizon Compatibility extension.
   **Context:** Each layer defines a soft decision boundary over co-policy space, supporting decentralized epistemic refinement under partial observability.
3. We introduce **Strategic InfoNCE**, a contrastive learning objective that aligns representations with soft best response similarity.
   **Context:** This grounds abstraction in influence, not identity, enabling scalable reasoning over co-agent relevance.
4. We reinterpret Bayes-Adaptive value decomposition Lidayan et al. (2025) as inspiration for abstraction-layer boundaries.
   **Context:** This provides a principled mechanism for strategic generalization in mixed-motive settings without full observability.

# Finding a Frame through Strategic Abstraction

**Sandy Tanwisuth**[1,2,3,†], **Niklas Lauffer**[2,3,4,†]
sandy314@duck.com, nlauffer@berkeley.edu

[1]**Independent Researcher**
[2]**Center for Human Compatible AI (CHAI)**
[3]**Machine Learning and Alignment Theory and Scholars (MATS) Program**
[4]**Department of Electrical Engineering and Computing Science, University of California, Berkeley**

† Both authors are funded by Cooperative AI Foundation Early Career Research Grant and PhD Fellowship respectively.

## Abstract

Real-world coordination begins not with complete knowledge, but with the search for the right frame: a way of interpreting others' behavior that reveals what matters for one's own decisions. Biological agents rely on minimal cues that signal when another's actions reshape their own incentives. We formalize this intuition with a theory of agentic coordination in which agents refine abstraction frames, filters over strategic relevance, rather than model co-agents in full. The framework introduces a layered process: the *Influence Filter* detects deformation of the ego agent's soft best response; *Short-horizon Compatibility* evaluates divergence in best response landscapes; and *Long-horizon Compatibility* (a proposed extension) captures long-term convergence in inferred incentives. These layers define soft decision boundaries that approximate the Value of Information and Value of Opportunity decomposition from Bayes-Adaptive MDPs. Agents maintain distributions over *Strategic Equivalence Classes* and refine them via *Strategic InfoNCE*, a contrastive objective over induced incentive gradients. A convergence signal, $\Delta\mathrm{SEC}(t)$, tracks entropy reduction over abstraction disagreement. In this view, coordination emerges from epistemic alignment over relevance, not identity.

## 1 Introduction

Coordination begins not with understanding, but with the search for relevance. In interactive, mixed-motive environments, agents must not only act under uncertainty but decide which distinctions in others' behavior are strategically relevant for shaping their own decisions. This paper invites the reader into that search. It is not only a framework, but a minimal scaffold for resolving epistemic ambiguity in coordination.

Our central claim is that coordination is not merely a computational problem but an epistemic one. In multi-agent reinforcement learning (MARL), most existing approaches treat coordination as either equilibrium computation Brown (1951); Hu & Wellman (2003) or policy inference Rabinowitz et al. (2018), often requiring identity modeling, full trajectory reconstruction, or centralized training. These methods can be brittle or inefficient under bounded rationality, deep uncertainty, and partial observability Leibo et al. (2017); Hughes et al. (2018). They presuppose a solved model of the other agent when in reality, even deciding what to model may be ambiguous.

What if the more foundational question is not how to act given a model, but how to recognize which distinctions matter? This becomes especially urgent when incentives are entangled, observed actions obscure latent intentions, and influence is distributed across time and context. Rather than solving others, we propose that agents must iteratively refine agreement over what is strategically relevant. Instead of reconstructing partner policies in full, agents learn to compress interaction histories into *strategic frames*: abstraction filters that preserve only distinctions that deform their own decision boundaries.

This framing builds on insights from cognitive science and developmental learning, where humans construct simplified models sufficient for aligning behavior rather than exhaustive theories of mind Baker et al. (2011); Jara-Ettinger (2019). Infants coordinate with caregivers through minimal signals like gaze or joint attention Gergely & Csibra (2003); Scaife & Bruner (1975), and social animals rely on simple, robust cues to coordinate effectively without modeling others' goals in detail Tomasello et al. (2005); Barrett (2011). In all of these cases, agents act as relevance-seeking learners, not omniscient predictors.

Bounded-rational models of strategic reasoning offer similar support. Cognitive hierarchy theory assumes agents respond to simplified levels of reasoning rather than full models of others Chong et al. (2005). Classical methods like fictitious play track empirical action frequencies instead of identity or intent Brown (1951). These approaches reflect a core insight: strategic coordination can emerge from approximating influence under epistemic constraints.

In MARL, Strategic Equivalence Relations (SER) formalize this idea by clustering co-policies based on the soft best responses they induce, not their surface behaviors Lauffer et al. (2023). But computing these relations typically requires access to partner policies or reward functions, resources unavailable to bounded agents. This opens deeper questions: is abstraction alone sufficient for coordination? When can distinctions be safely compressed, and when must epistemic ambiguity be resolved? If refinement is necessary, what exactly should agents infer and how?

We frame coordination as a layered process of *strategic abstraction refinement*. Rather than predicting identity or reconstructing full co-policies, agents update structured representations of how others reshape their own incentives. This unfolds in three stages. First, the **Influence Filter** asks whether partner behavior consistently deforms the agent's own reward landscape. If influence is negligible or noisy, abstraction is deferred. Second, under clear influence, the agent estimates **Short-horizon Compatibility** by comparing the soft best responses induced by different co-policies. Finally, the agent assesses **Long-horizon Compatibility** to determine whether short-term influence generalizes across trajectories, guarding against premature convergence in mixed-motive settings.

This structure builds on the decomposition introduced in Bayes-Adaptive MDPs (BAMDPs), where each belief state $b_t$, a posterior over latent dynamics or reward functions, defines the agent's epistemic state Bellman & Kalaba (1959); Martin (1965); Lidayan et al. (2025). The value of a belief decomposes into two components: the *Value of Opportunity* (VOO), which quantifies expected return under current uncertainty, and the *Value of Information* (VOI), which captures the expected benefit of future belief refinement. Ross et al. extended this framework to the partially observable case through Bayes-Adaptive POMDPs (BAPOMDPs), formalizing belief updates over both latent states and model parameters and offering a principled way to plan under joint epistemic uncertainty Ross et al. (2007).

We adapt this decomposition to a decentralized, multi-agent setting where agents lack access to global latent states and cannot explicitly model all partner behaviors. Instead, agents maintain beliefs over abstraction-layer *agreement*—shared inferences about which distinctions in partner behavior are strategically relevant. In this framing, the **Influence Filter** serves as a VOI signal, gating when partner behavior provides sufficient deformation of the agent's incentive landscape to merit abstraction refinement. **Short-horizon Compatibility** approximates a VOO term, estimating the actionable value of treating partners as strategically similar. **Long-horizon Compatibility**, as a proposed extension, reintroduces VOI over time by tracking the consistency of influence across trajectories. In contrast to classical BAMDP and BAPOMDP approaches, which operate over latent

variables or model parameters, our formulation treats *strategic abstraction* itself as the epistemic object: refining not the world model, but the relevance model.

This abstraction-based filtering explains generalization failures in coordination. In Coin Game Raileanu et al. (2018), two distinct behaviors (hovering passively vs. aggressively collecting red coins) may induce the same best response, signaling non-cooperation, despite surface-level differences. In Overcooked Carroll et al. (2020), strategically equivalent partners can differ in speed or path yet remain interchangeable in layouts like Coordination Ring. But in layouts like Locked-In, small variations in timing force incompatible responses, requiring finer abstraction. These failures arise not from observational noise, but from overcommitting to irrelevant distinctions or underrepresenting relevant ones. They reflect deeper risks documented in recent work on multi-agent system failures, including miscoordination, incentive mismatch, and abstraction collapse Hammond et al. (2025).

While prior works such as BAD (Foerster et al., 2018), SAD (Hu et al., 2021b), and OBL (Hu et al., 2021a) promote coordination through population-level training or Bayesian modeling, they often assume centralized data or full cooperation. Our approach treats coordination as an epistemic problem of abstraction refinement under bounded rationality and partial observability. Rather than predicting behavior or inferring goals, agents update soft best responses based on which distinctions in partner behavior remain unresolved.

To operationalize this, we introduce **Strategic InfoNCE**, a contrastive objective adapted from contrastive predictive coding Oord et al. (2019). While CPC aligns representations over observation sequences, Strategic InfoNCE embeds co-policies based on how they deform the ego agent's soft best response distribution. Co-policies are considered similar when they induce similar incentive shifts. The agent samples soft responses under each co-policy, compares them to negatives from others, and uses a contrastive loss to align embeddings by influence—not identity or transition similarity. This forms a representation space grounded in strategic relevance.

We offer this not as a full theory, but as a minimal and testable reframing. Strategic Abstractions do not require full observability, centralized data, or explicit meta-games (Howard, 2003). Instead of constructing empirical payoff tables as in EGTA (Wellman et al., 2025), agents learn relevance through soft best response deformation. And while top-down approaches like Guaranteed Safe AI (GSAI) ("davidad" Dalrymple et al., 2024) offer verifiability through world models, our approach enables safety through adaptability: agents learn what matters by interacting, not assuming it in advance.

In this light, abstraction is not a shortcut but a necessity. Coordination does not require solving others. It requires resolving ambiguity over influence: strategically, incrementally, and under uncertainty. Strategic Abstractions offers a general framework for doing just that: retaining only the distinctions that deform incentives, and refining them as agents learn to converge.

## 2 Preliminaries: Abstraction and Epistemic Uncertainty

Coordination in mixed-motive, partially observable environments is fundamentally an epistemic problem. Agents are not simply uncertain about hidden states or other agents' actions; they are uncertain about which distinctions in observed behavior are strategically *relevant* for their own decision-making. This view builds on insights from Sequential Social Dilemmas (SSDs) (Leibo et al., 2017), which show that strategic behavior arises from temporally extended policy interactions, not atomic actions. We extend this further: rather than classifying partner behavior as cooperative or defecting, we ask which aspects of that behavior remain relevant for one's own incentives. Our framework reframes coordination as a layered process of abstraction refinement: instead of recovering full policies, agents learn to compress and update representations of influence, defined as what others' actions imply for their own incentives.

We adopt the formal structure of a general-sum Decentralized Partially Observable Markov Decision Process (Dec-POMDP) (Oliehoek et al., 2006) with agent-specific rewards:

$$\mathcal{G} = (I, S, \{A_i\}_{i \in I}, T, \{R_i\}_{i \in I}, \{\Omega_i\}_{i \in I}, O, \gamma),$$

where:

- $I = \{1, \ldots, m\}$ is the finite set of agents;
- $S$ is the set of latent environment states;
- $A_i$ is the action set for agent $i$, with $A = \prod_{i \in I} A_i$ the joint action space;
- $T : S \times A \to \Delta(S)$ is the transition function;
- $R_i : S \times A \to \mathbb{R}$ is the reward function for agent $i$;
- $\Omega_i$ is the observation space for agent $i$, with $\Omega = \prod_{i \in I} \Omega_i$ the joint observation space;
- $O : S \times A \to \Delta(\Omega)$ is the observation function;
- $\gamma \in [0, 1)$ is the discount factor.

Each agent $i$ observes a local trajectory $h_i^t = (o_i^1, a_i^1, \ldots, o_i^t)$ and selects actions via a stochastic policy $\pi_i : H_i \to \Delta(A_i)$. Agents do not observe the environment state $s_t$, nor their co-players' policies $\pi_{-i}$ or rewards $R_{-i}$.

We interpret this formalism through an epistemic lens: histories are not merely data streams, but evolving records of interaction from which agents infer what distinctions remain decision-relevant. Under bounded rationality, agents refine compact strategic abstractions $\phi_t^i$ over time, learning not who their partners are, but which distinctions still matter for coordination.

This perspective deviates from full Bayesian modeling approaches such as the Bayes-Adaptive POMDPs (BA-POMDPs) Ross et al. (2007) and its hypothetical decentralized extension (Dec-BAPOMDP) which represent uncertainty over latent dynamics or rewards via belief states and treat planning as inference in belief space. While these models formally encode memory by embedding history into a Markovian structure over beliefs, they are often intractable in decentralized settings and require assumptions about the generative model of others. In contrast, our framework avoids full posterior tracking by reframing the decision problem as one of epistemic compression. Agents do not model latent structure directly; they update abstraction layers over history that preserve only incentive-relevant distinctions. This effectively breaks the memoryless assumption of standard MDPs, but in a principled way: history becomes epistemically meaningful not for reconstructing hidden states, but for filtering ambiguity about strategic influence.

This foundational view enables us to reinterpret Dec-POMDP inference not as full state reconstruction, but as *relevance refinement*: a layered compression of uncertainty aligned with epistemic needs for acting well.

### 2.1 Strategic Equivalence and Ambiguity

Coordination in general-sum, partially observable environments is not merely a problem of inferring partner actions; it is a problem of identifying which distinctions in those actions are *strategically relevant*. Following (Lauffer et al., 2023), we frame this relevance through the lens of **Strategic Equivalence**: agents should only differentiate between partner behaviors that meaningfully deform their own incentive structure.

**Definition 1 (Strategic Equivalence (Lauffer et al., 2023))** *Two co-agent policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ are strategically equivalent for agent $i$ if they induce the same best response set:*

$$\pi_{-i} \sim_i \pi'_{-i} \quad \textit{iff} \quad \mathrm{BR}_i(\pi_{-i}) = \mathrm{BR}_i(\pi'_{-i})$$

This induces a partition of the co-policy space into **Strategic Equivalence Classes (SECs)**: minimal behavioral clusters that preserve incentive-relevant distinctions. However, exact computation of

SECs assumes access to full co-policies and unbounded rationality, conditions that are rarely met in practice.

Instead, our framework treats these partitions as soft, behaviorally grounded approximations. At each timestep, the agent maintains a belief over which co-policy distinctions are still strategically ambiguous i.e., whether observed behaviors could lead to different actions if left unresolved.

**Definition 2 (Strategic Ambiguity)** *A co-policy set $\Pi_{-i}$ is strategically ambiguous for agent $i$ if it contains policies that induce distinct best responses:*

$$\exists\, \pi_{-i}, \pi'_{-i} \in \Pi_{-i} \quad \textit{s.t.} \quad \mathrm{BR}_i(\pi_{-i}) \neq \mathrm{BR}_i(\pi'_{-i})$$

The agent's objective becomes resolving this ambiguity: not by recovering identities, but by refining a soft abstraction filter $\phi_t^i$ that encodes influence-relevant structure. This abstraction acts as a decision boundary over co-policy space: collapsing strategically equivalent partners and preserving distinctions only when they meaningfully shift the agent's soft best response.

Later sections (Section 3.1–3.5) formalize how soft best response distributions define a continuous relaxation of strategic equivalence, and how $\Delta\mathrm{SEC}(t)$, the entropy drop over soft SEC belief, measures epistemic convergence. These concepts operationalize abstraction refinement as tractable filtering over what matters. It is not about who the partner is, but about how their behavior reshapes decision boundaries.

### 2.2 Epistemic Value Decomposition

We adapt a central insight from Bayes-Adaptive Markov Decision Processes (BAMDPs) (Lidayan et al., 2025), where the agent maintains a posterior belief $b(\theta)$ over latent MDP parameters $\theta$. The value of a belief state reflects two epistemic contributions: the immediate utility of acting under the current belief and the expected improvement from future refinement. This leads to the canonical decomposition:

$$V^{\pi_i}(b) = \underbrace{\mathbb{E}_{\theta\sim b}\left[V^{\pi_i}(M_\theta)\right]}_{\text{Value of Opportunity (VOO)}} + \underbrace{\mathbb{E}_{\theta\sim b}\left[\Delta V^{\pi_i} \mid \text{future observations}\right]}_{\text{Value of Information (VOI)}}$$

This BAMDP formulation motivates agents to act not only to maximize return but also to resolve uncertainty that could deform future decisions. However, in mixed-motive multi-agent settings, uncertainty does not solely arise from latent environmental parameters—it emerges from ambiguous influence: the uncertain impact of co-agent behavior on one's own incentives.

#### 2.2.1 Abstraction as Epistemic Surrogate.

Instead of maintaining a posterior over latent MDPs, each agent $i$ in our framework maintains a soft abstraction $\phi_t^i := f(h_t^i)$ over its own interaction history $h_t^i$. This abstraction summarizes which distinctions in co-agent behavior are still relevant for deforming the ego agent's soft best response distribution. These filters act as epistemic surrogates for belief: encoding distinctions worth preserving, collapsing those that do not alter decision geometry, and deferring those that may become relevant in extended horizons.

#### 2.2.2 Reinterpreting VOO and VOI over Strategic Influence.

We reinterpret the BAMDP decomposition over this abstraction space, aligning with the structural formulation in Section 3.3. Each abstraction layer maps to a distinct component of epistemic value:

- **Influence Filter** $\leftrightarrow$ *VOI:* Captures whether distinctions in co-agent behavior induce significant shifts in the ego agent's soft best response. High mutual information $I(a_i; \pi_{-i})$ implies that further abstraction refinement is epistemically valuable.

- **Short-horizon Compatibility** $\leftrightarrow$ *VOO:* Measures whether currently preserved distinctions suffice to support effective behavior. It tests local substitutability across co-policies within soft strategic equivalence regions, quantified via divergence in influence embeddings.
- **Long-horizon Compatibility** $\leftrightarrow$ *Deferred VOI (speculative):* Encodes whether behaviorally indistinct co-policies may diverge over outcome trajectories. Though currently unimplemented, it frames value refinement over latent long-run influence and supports future extension.

#### 2.2.3 Two Key Departures.

This reinterpretation introduces two principled shifts from BAMDP:

1. **From belief over dynamics to belief over influence relevance:** Rather than tracking hidden $\theta$, we maintain $\phi_t^i$ as a latent abstraction encoding relevance for incentive deformation. The agent updates $\phi_t^i$ by minimizing $\Delta\mathrm{SEC}(t)$ which is a proxy for how much ambiguity remains about which distinctions in $\pi_{-i}$ still deform the agent's response.
2. **From model-based planning to behaviorally grounded surrogates:** We avoid posterior sampling or planning over transition models. Instead, epistemic gains are estimated from divergence in soft best responses, operationalized through contrastive learning (see Section 3.4). This enables tractable coordination in decentralized, bounded-rational regimes.

Strategic abstraction filters instantiate a form of epistemic reasoning under bounded capacity: each layer approximates the BAMDP decomposition not over world models but over influence-relevant distinctions. This aligns with our reinterpretation of coordination as resolving *strategic ambiguity*, where the object of uncertainty is not the world, but the incentive effects of others' behavior.

For formal operationalization of these abstraction filters as decision boundaries over belief space, see Section 3.3.

### 2.3 Epistemic Compression and Contrastive Learning

Coordination under bounded rationality requires agents to resolve which behavioral distinctions matter; not who their partner is, but how the partner's behavior alters their own decision boundary. We formalize abstraction as the epistemic act of compressing distinctions in partner behavior into a representation sufficient to guide best-response adaptation. This view reframes representation learning as strategic influence filtering.

#### 2.3.1 Compression via Mutual Information.

Our approach builds on Contrastive Predictive Coding (CPC) (Oord et al., 2019), which formalizes abstraction as information-preserving compression. Given inputs $x$, $x^+$, and negatives $\{x_j^-\}$, CPC optimizes the InfoNCE objective:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\mathrm{sim}(f(x), f(x^+))}{\mathrm{sim}(f(x), f(x^+)) + \sum_j \mathrm{sim}(f(x), f(x_j^-))}\right]$$

where $f$ is a learned encoder and $\mathrm{sim}$ is a similarity function. Minimizing this loss bounds the mutual information $I(f(x); f(x^+))$, preserving features that align with positive samples while collapsing irrelevant variation.

### 2.4 Strategic InfoNCE: Strategic Compression.

To learn *strategic-relevant* abstractions, we replace predictive targets with behavioral gradients. Following the SER framework (Lauffer et al., 2023), we define the epistemic relevance of a co-policy $\pi_{-i}$ by how it deforms the ego agent's soft best response $\mathrm{BR}_\tau^i(\pi_{-i})$. Co-policies that induce indistinguishable soft BRs are treated as strategically equivalent.

Rather than recover partner identity or latent state, we train an encoder $f(\pi_{-i})$ that maps co-policies to embeddings consistent with their incentive impact. This is formalized through the *Strategic InfoNCE loss* (Definition 6 in Section 3.2):

$$\mathcal{L}_{\text{Strategic InfoNCE}} = -\mathbb{E}\left[\log \frac{\text{sim}(f(\pi_{-i}), f(a^+))}{\text{sim}(f(\pi_{-i}), f(a^+)) + \sum_j \text{sim}(f(\pi_{-i}), f(a_j^-))}\right]$$

where $a^+ \sim \text{BR}_\tau^i(\pi_{-i})$ and $a_j^- \sim \text{BR}_\tau^i(\pi'_{-i})$ for strategically distinct $\pi'_{-i}$.

#### 2.4.1 Abstracting over Strategic Equivalence.

This contrastive formulation operationalizes a soft approximation of Strategic Equivalence Classes (SECs). The learned abstraction $\phi_t^i := f(\pi_{-i})$ clusters co-policies by their influence effect, collapsing distinctions that do not induce meaningful changes in the ego agent's response. The encoder thereby implements a surrogate sufficient statistic for co-agent relevance.

#### 2.4.2 Influence Filter as Epistemic Gate.

We interpret the mutual information $I(a_i; \pi_{-i})$ as a proxy for epistemic legibility. When this quantity is low, the co-agent's policy induces no reliable deformation in the ego agent's preferences, and further modeling offers little epistemic gain. This underpins the *Influence Filter* layer (see Section 3.2), which gates downstream refinement. High information signals relevance; low information implies safe collapse.

#### 2.4.3 Abstraction as Filtering, Not Prediction.

This perspective departs from standard predictive contrastive learning. The goal is not to infer latent dynamics or reconstruct trajectories, but to retain distinctions in $\pi_{-i}$ that matter for action selection. Strategic InfoNCE implements this compression faithfully: it encodes relevance through behaviorally grounded soft best response traces, not through imitation or state reconstruction.

By embedding only distinctions that deform the agent's decision geometry, this contrastive method implements epistemic filtering over influence. It supports tractable coordination under bounded memory, enabling agents to learn what matters without full posterior inference.

### 2.5 Belief over Abstraction frames

We model each agent's epistemic state not as a belief over latent environment parameters, but as a belief over which distinctions in co-agent behavior remain strategically relevant. This reformulation preserves the epistemic architecture of Bayes-Adaptive MDPs (BAMDPs) (Lidayan et al., 2025), while shifting the object of inference: from dynamics-driven latent variables $\theta$ to influence-driven abstraction filters $\phi_t^i$.

#### 2.5.1 Layered Posterior Structure

Formally, we define the abstraction filter $\phi_t^i$ as a soft posterior over a layered space of epistemic decision boundaries. Each layer corresponds to a distinct abstraction over co-agent influence. Following the layered structure introduced in the method section (see Section 3.2), we approximate the agent's posterior over co-policies as:

$$b_t^i(\pi_{-i}) \approx b_t^{IF}(\pi_{-i}) \cdot b_t^{SC}(\pi_{-i}) \cdot b_t^{LC}(\pi_{-i})$$

Each term represents a structured belief over regions of the co-policy space that approximate Strategic Equivalence Classes (SECs) (Lauffer et al., 2023) under bounded rationality:

- $b_t^{IF}(\pi_{-i})$ quantifies whether distinctions in $\pi_{-i}$ deform the ego agent's soft best response i.e., whether influence is epistemically legible.
- $b_t^{SC}(\pi_{-i})$ captures whether the current abstraction supports robust, stable best responses i.e., whether distinctions in $\pi_{-i}$ are functionally irrelevant in the short term.
- $b_t^{LC}(\pi_{-i})$, when included, estimates whether subtle differences in $\pi_{-i}$ may accrue strategic value over extended interaction. We treat this layer as a speculative extension, not operationalized in our current implementation.

This factorization generalizes the BAMDP principle that beliefs serve as sufficient statistics for planning. Whereas BAMDP maintains $b(\theta)$ over latent dynamics, our framework maintains $\phi_t^i$, a posterior over influence-relevant distinctions that gate soft best response adaptation.

#### 2.5.2 Update via Behavioral Divergence

Rather than using likelihood-based Bayesian updates, we refine $\phi_t^i$ through contrastive behavioral signals. At each timestep, the agent minimizes abstraction-level disagreement via:

$$\Delta\mathrm{SEC}(t) = H[\phi_t^i] - H[\phi_{t+1}^i]$$

where $H[\cdot]$ denotes entropy over the abstraction-layer posterior. This entropy reduction quantifies how much epistemic ambiguity remains about whether distinctions in $\pi_{-i}$ induce divergent soft best responses. As $\Delta\mathrm{SEC}(t) \to 0$, the abstraction filter approaches epistemic sufficiency.

#### 2.5.3 Grounding in Strategic Equivalence

Each abstraction layer implements a soft approximation of Strategic Equivalence Classes. Formally, two co-agent policies are strategically equivalent if they induce the same best response:

$$\pi_{-i} \sim_i \pi'_{-i} \quad \text{iff} \quad \mathrm{BR}^i(\pi_{-i}) = \mathrm{BR}^i(\pi'_{-i})$$

Under bounded rationality, we relax this definition via a divergence threshold:

$$\mathrm{SD}_\tau(\pi_{-i}, \pi'_{-i}) := \mathrm{KL}\left[\mathrm{BR}_\tau^i(\pi_{-i}) \,\|\, \mathrm{BR}_\tau^i(\pi'_{-i})\right] \approx 0$$

This strategic divergence $\mathrm{SD}_\tau$ defines the soft geometry of each abstraction layer. Co-policies that fall within low-divergence regions are collapsed into a shared representation, while distinctions across high-divergence regions are retained.

#### 2.5.4 Operationalization via Contrastive Learning

The Strategic InfoNCE loss introduced in Section 3 provides the operational mechanism for belief refinement. Rather than reconstruct co-agent policies, the encoder $f$ learns to preserve only distinctions that alter the ego agent's soft best response. In effect, $f$ compresses interaction history into influence-relevant embeddings and updates $\phi_t^i$ by tracking shifts in SEC boundary structure.

**Note on Practical Estimation.** While this framework grounds abstraction learning in epistemic structure, estimation of these frames in practice may be constrained by limited observations or capacity. We consider extensions for scalable approximation in future work.

## 3 Layered Reasoning under Strategic Ambiguity

Coordination under uncertainty is not hard because partner behavior is hidden, it is hard because it is ambiguous. Agents must decide not just what others might do, but which differences in their

behavior matter for their own decision-making. Our method reframes coordination as an epistemic filtering process: agents learn to abstract away irrelevant distinctions while preserving those that deform their own incentive structure.

We formalize this in the partially observable setting of POSGs, adopting Dec-POMDP notation (Section 2) for clarity. However, our framework does not rely on centralized supervision. Rather than inferring the full partner policy, each agent learns a compressed abstraction $\phi_t^i$ of co-agent influence that capturing how past interaction reshapes its own soft best response distribution.

This abstraction is trained through contrastive self-supervision: agents observe the behavioral consequences of interaction and align embeddings based on the gradients they induce in ego incentives. No access to partner identity or ground truth policies is required.

For efficiency, our planned implementation optionally uses a Centralized Training and Decentralized Execution (CTDE) setup Amato (2024), where joint trajectories can accelerate learning. But fundamentally, Strategic InfoNCE can operate from decentralized experience alone: it learns what matters to coordinate; not by modeling others, but by identifying how their behavior shifts the agent's own strategy space.

### 3.1 Soft Best Response as Epistemic Primitive

Coordination under uncertainty is not merely about predicting what a partner will do. It is about determining which aspects of their behavior reshape the agent's own incentives. Rather than solving for game outcomes, we solve for *epistemic sufficiency under interaction*: what distinctions must be retained to act well, given bounded observability and memory.

**Definition 3 (Soft Best Response)** *We formalize this influence through the soft best response distribution:*

$$BR_i^\tau(a_i \mid \pi_{-i}) := \frac{\exp(Q_i^{\pi_i,\pi_{-i}}(a_i)/\tau)}{\sum_{a'} \exp(Q_i^{\pi_i,\pi_{-i}}(a')/\tau)}$$

*where $Q_i^{\pi_i,\pi_{-i}}(a_i)$ denotes the expected return to agent $i$ for taking action $a_i$, assuming ego policy $\pi_i$ and co-agent policy $\pi_{-i}$, and where $\tau > 0$ modulates bounded rationality. This distribution captures the agent's stochastic preference structure in response to a particular co-agent.*

We formalize this influence through the *soft best response* distribution:

$$\mathrm{BR}_i^\tau(a_i \mid \pi_{-i}) := \frac{\exp(Q_i^{\pi_i,\pi_{-i}}(a_i)/\tau)}{\sum_{a'} \exp(Q_i^{\pi_i,\pi_{-i}}(a')/\tau)}$$

where $Q_i^{\pi_i,\pi_{-i}}(a_i)$ denotes the expected return to agent $i$ for taking action $a_i$, assuming ego policy $\pi_i$ and co-agent policy $\pi_{-i}$, and where $\tau > 0$ modulates bounded rationality. This distribution captures the agent's stochastic preference structure in response to a particular co-agent.

**Interpretation.** The soft best response defines a functional mapping:

$$\mathrm{BR}_i^\tau : \Pi_{-i} \to \Delta(A_i)$$

which associates each co-agent policy with a probability distribution over ego actions, representing how the partner's behavior deforms the agent's incentive landscape. Rather than inferring *who* the partner is, the agent infers how interaction affects its own strategic boundary.

We interpret $\mathrm{BR}_i^\tau(\pi_{-i}) \in \Delta(A_i)$ as a *preference signature*: a minimal, behaviorally grounded trace of how the co-agent influences the ego's choices. What matters is not reconstructing the hidden state of $\pi_{-i}$, but preserving distinctions in how it shifts incentives.

This motivates a divergence measure over co-policies based on their strategic effect:

**Definition 4 (Strategic Divergence)**

$$SD_\tau(\pi_{-i}, \pi'_{-i}) := D_{KL}(BR_i^\tau(\pi_{-i}) \,\|\, BR_i^\tau(\pi'_{-i}))$$

*Two co-policies are considered* soft strategically equivalent *if this divergence is zero—i.e., they are indistinguishable in their influence on the agent's preferences.*

**Theorem 1 (Zero Strategic Divergence Implies Strategic Equivalence)** *Let $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ be co-agent policies such that:*

$$D_{\mathrm{KL}}\left(BR_i^\tau(\pi_{-i}) \,\|\, BR_i^\tau(\pi'_{-i})\right) = 0$$

*Then they induce the same hard best response set:*

$$BR_i(\pi_{-i}) = BR_i(\pi'_{-i}) \quad \Rightarrow \quad \pi_{-i} \sim_i \pi'_{-i}$$

**Proof 1** *If $BR_i^\tau(\pi_{-i}) = BR_i^\tau(\pi'_{-i})$, then their softmax distributions over Q-values are equal. This implies the Q-values differ only by a constant shift across actions, preserving the argmax structure. Therefore, the resulting hard best response sets are identical: $\pi_{-i} \sim_i \pi'_{-i}$.*

**Remark.** This theorem grounds our use of soft best response as an epistemic primitive. Agents do not need to recover partner identities; they only need to track how behavior modulates their own decision geometry. Strategic equivalence, in this framing, becomes a question of indistinguishability in influence.

From a Bayes-Adaptive perspective, the agent's history $h_t$ encodes all epistemically relevant information. Our abstraction $\phi_t^i := f(h_t)$ serves as a compressed belief statistic: preserving distinctions in $\mathrm{BR}_i^\tau$ sufficient for epistemic value computation. Rather than computing full Bayes-optimal value, the agent filters $h_t$ through learned influence embeddings that approximate which distinctions matter to act.

All abstraction frames (Influence Filter, Short-horizon Compatibility, and Long-horizon Compatibility) can be viewed as epistemic filters over this influence space. Each acts as a lossy compression gate on $h_t$, retaining only those aspects of interaction that deform the soft BR. In doing so, they approximate a decomposition of value into relevance (VOO) and refinement (VOI), aligning coordination with belief compression.

This reframes the coordination problem not as inference over hidden types, but as abstraction alignment over epistemic relevance: learning which parts of the world and the partner relevant to preserve in order to act well.

### 3.2 Filtering Influence-Equivalence via Mutual Information

Having introduced soft best response as the epistemic primitive for capturing strategic influence, we now define a series of filters that implement tractable approximations over co-policy space. These filters compress the agent's belief over co-agent influence into a latent abstraction $\phi_t^i = f(h_t)$, retaining only distinctions relevant to the ego agent's decision boundary.

Each component below acts as a surrogate sufficient statistic over history $h_t$, approximating Bayes-optimal belief refinement under bounded rationality. This aligns with the BAMDP framing: rather than tracking latent states directly, the agent maintains soft posteriors over Strategic Equivalence Class (SEC) membership.

**Relaxed Influence Equivalence.** We begin by softening the notion of strategic equivalence into a fuzzy, tolerance-based decision boundary.

**Definition 5 ($\varepsilon$-Soft Strategic Equivalence)** *Two co-policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ are said to be* $\varepsilon$-soft strategically equivalent *if:*

$$SD_i^\tau(\pi_{-i}, \pi'_{-i}) := D_{\mathrm{KL}}\left(BR_i^\tau(\pi_{-i}) \,\|\, BR_i^\tau(\pi'_{-i})\right) \leq \varepsilon$$

*This defines a belief region in which co-policies induce indistinguishable influence on the ego agent's soft best response.*

**Learning Embedding-Based Compression.** We next define a contrastive loss over co-policies that clusters them according to their effect on ego preferences.

**Definition 6 (Strategic InfoNCE Loss)** *To learn embeddings that reflect influence-relevant distinctions, we define:*

$$\mathcal{L}_{\textit{InfoNCE}} = -\mathbb{E}\left[\log \frac{f(a^+, h(\pi_{-i}))}{\sum_j f(a_j^-, h(\pi_{-i}))}\right]$$

*where $a^+ \sim BR_i^\tau(\pi_{-i})$ is a positive action, $a_j^- \sim BR_i^\tau(\pi_j^-)$ are negative samples from unrelated co-policies, and $f$ is a similarity function (e.g., dot product or cosine).*

**Lemma 1 (InfoNCE Minimization Implies Embedding Consistency)** *Let $h : \Pi_{-i} \to \mathbb{R}^d$ be trained to minimize $\mathcal{L}_{\textit{InfoNCE}}$. Then:*

$$SD_i^\tau(\pi_{-i}, \pi'_{-i}) \leq \varepsilon \quad \Rightarrow \quad \|h(\pi_{-i}) - h(\pi'_{-i})\| \text{ is small}$$

*Thus, the learned embedding space approximates a soft clustering over strategic equivalence.*

**Theorem 2 (InfoNCE Bounds Strategic Mutual Information)** *Minimizing $\mathcal{L}_{\textit{InfoNCE}}$ maximizes a lower bound on the mutual information between co-policy and ego response:*

$$I_{BR} := I(a_i; \pi_{-i}) = \mathbb{E}_{\pi_{-i}}\left[D_{\mathrm{KL}}(BR_i^\tau(\pi_{-i}) \| p(a_i))\right]$$

*Thus, InfoNCE approximates an embedding of $\Pi_{-i}$ that preserves influence-relevant distinctions for soft best response behavior.*

**Influence Filter as VOI Gate.** To gate abstraction refinement, we introduce a mutual information measure that filters co-policies based on epistemic legibility.

**Definition 7 (Influence Filter via Mutual Information)** *Given a baseline response prior $p(a_i)$, define:*

$$IF(\pi_{-i}) := I(a_i; \pi_{-i}) = \mathbb{E}_{\pi_{-i}}\left[D_{\mathrm{KL}}(p(a_i \mid \pi_{-i}) \| p(a_i))\right]$$

*This quantifies how much the co-policy reduces uncertainty in ego response, enabling influence-gated abstraction.*

**Definition 8 (Contrastive Estimation of Influence Filter)** *Influence Filter can also be approximated via a contrastive loss:*

$$\mathcal{L}_{IF} = -\mathbb{E}\left[\log \frac{f(a^+, h_{IF}(\pi_{-i}))}{\sum_j f(a_j^-, h_{IF}(\pi_{-i}))}\right]$$

*where $h_{IF}$ is a learned embedding function.*

**Definition 9 (Influence Filtering)** *The agent restricts attention to co-policies with high epistemic legibility:*

$$\mathcal{P}_{\textit{intent}} := \{\pi_{-i} \in \Pi_{-i} \mid IF(\pi_{-i}) \geq \theta_{IF}\}$$

*This defines a VOI-based posterior region over SECs, avoiding noise-induced refinement.*

**Remark (Bayes-Adaptive Interpretation).** These constructs implement belief refinement over influence-relevant features:

- **Soft SEC** defines a fuzzy posterior on substitutable co-policies.
- **Strategic InfoNCE** learns a surrogate sufficient statistic over $h_t$ for influence-preserving embedding.
- **Influence Filter** acts as an epistemic filter, gating abstraction by VOI thresholds.

Together, they approximate a decomposed posterior belief over co-agent policy space:

$$b_t^i(\pi_{-i}) \approx b_{\text{IF}}^i \cdot b_{\text{InfoNCE}}^i \cdot b_{\text{softSEC}}^i$$

where each factor corresponds to a filter layer. This belief refinement avoids full posterior inference while retaining the distinctions necessary for acting well under uncertainty.

### 3.3 Abstraction Filters as Decision Boundaries Over Belief Space

We reinterpret Bayes-Adaptive belief refinement for multi-agent coordination by introducing *abstraction filters* as soft decision boundaries over the agent's posterior beliefs about Strategic Equivalence Class (SEC) membership. A SEC groups co-policies that are interchangeable in their influence on the ego agent's decision frontier. Rather than maintaining beliefs over latent environmental dynamics as in BAMDPs, we maintain posteriors over *co-policy neighborhoods*, defined by their effect on ego-agent incentives.

Formally, we define a latent abstraction variable $\phi_t^i = f(h_t)$, where $h_t$ is the interaction history. This variable represents a compressed belief over strategically relevant distinctions. Each abstraction filter acts as a surrogate sufficient statistic for approximating soft best responses, enabling coordination under bounded rationality and decentralized observability.

**Compression Principle.** Each abstraction layer implements an epistemic compression:

$$h_t \longmapsto \phi_t^i = f(h_t)$$

retaining only distinctions that matter for recovering the ego agent's soft best response. This defines a functional posterior over influence-relevant regions in co-policy space.

**Influence Filter: VOI-Gated Influence Legibility.** Influence Filter quantifies whether a co-policy $\pi_{-i}$ induces sufficiently determinate preferences in the ego agent. This serves as a Value of Information (VOI) gate over belief refinement.

**Definition 10 (Influence Filter via Entropy)** *Let $BR_i^\tau(\pi_{-i}) \in \Delta(A_i)$ be the soft best response. Then:*

$$IF(\pi_{-i}) := \mathcal{H}\left[BR_i^\tau(\pi_{-i})\right]$$

*Low entropy implies high legibility: the partner's influence on the ego's preferences is determinate.*

**Definition 11 (Mutual Information Formulation)** *If a baseline response prior $p(a_i)$ is known, we alternatively define:*

$$IF(\pi_{-i}) := I(a_i; \pi_{-i}) = \mathbb{E}_{\pi_{-i}}\left[D_{\text{KL}}(p(a_i \mid \pi_{-i}) \,\|\, p(a_i))\right]$$

*This quantifies epistemic gain about the ego agent's actions.*

**Definition 12 (Influence Filtering)** *Define the intent region:*

$$\mathcal{P}_{intent} := \{\pi_{-i} \in \Pi_{-i} \mid IF(\pi_{-i}) \geq \theta_{IF}\}$$

*Only legible co-policies are passed to downstream abstraction frames.*

**Short-horizon Compatibility (SC): VOO-Grounded Predictive Alignment.** SC tests whether two co-policies induce similar short-horizon deformations of ego preferences, operationalized through a learned abstraction space.

**Definition 13 (Short-horizon Compatibility)** *Let $h_t$ and $h'_t$ be short interaction histories with co-policies $\pi_{-i}$ and $\pi'_{-i}$, respectively. Let:*

$$\phi_t^i = f(h_t), \quad \phi_t'^i = f(h'_t)$$

*be learned embeddings from a contrastive model trained to preserve soft BR gradients. Define:*

$$SC(\pi_{-i}, \pi'_{-i}) := D_{\mathrm{SC}}(\phi_t^i, \phi_t'^i)$$

*where $D_{\mathrm{SC}}$ is a divergence (e.g., cosine or Euclidean). If SC $\leq \delta_{SC}$, the co-policies are considered compatible over short horizons.*

**Remark 1** *SC approximates **Value of Opportunity (VOO)**. Refinement is triggered only when partner influence induces nontrivial deviation in ego policy. This generalizes SER by learning equivalence-relevant geometry.*

**Definition 14 (Filtered Belief Approximation)** *This defines a soft posterior:*

$$b_t^i(\pi_{-i}) \propto p(h_t \mid \pi_{-i}) \cdot p(\pi_{-i})$$

*with $f(h_t)$ acting as a surrogate sufficient statistic.*

**Long-horizon Compatibility (LC): Deferred VOI via Outcome Divergence.** LC identifies co-policies that diverge in long-horizon influence, even if locally aligned.

**Definition 15 (Long-horizon Compatibility)** *Let $\phi(s, a) \in \mathbb{R}^d$ be semantically meaningful features. Define the ego agent's successor representation under co-policy $\pi_{-i}$ as:*

$$\xi^{\pi_i, \pi_{-i}} := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \,\middle|\, \pi_i, \pi_{-i}\right]$$

*Then:*

$$LC(\pi_{-i}, \pi'_{-i}) := D_{\mathrm{LC}}\left(\xi^{\pi_i, \pi_{-i}} \,\|\, \xi^{\pi_i, \pi'_{-i}}\right)$$

**Remark 2** *LC approximates long-horizon VOI: although co-policies appear locally equivalent, their cumulative effects diverge. Refinement is warranted when this divergence becomes actionable.*

**Assumption 1 (Short-Term Epistemic Commitment)** *During inference at time $t$, agents treat their recent past policy $\pi^i_{t-\tau:t}$ as fixed, enabling co-policy inference under bounded memory and stable short-term assumptions. This does not preclude long-term adaptation or re-optimization.*

**Unified Mapping to BAMDP.** Each abstraction layer corresponds to a specific component of Bayes-Adaptive decision refinement:

- **Influence Filter** $\leftrightarrow$ VOI (Legibility): Skip refinement when co-policy influence is already determinate.
- **SC** $\leftrightarrow$ VOO (Local Discrimination): Cluster co-policies with similar immediate effect.
- **LC** $\leftrightarrow$ VOI (Deferred): Trigger belief refinement only when long-term divergence is behaviorally relevant.

**Epistemic Summary.** Abstraction filters act as soft decision boundaries over posterior belief in co-agent influence space. Each filter defines a region over which co-policies are behaviorally substitutable under bounded observability. Formally, we approximate:

$$b_t^i(\pi_{-i}) \approx b_{\text{IF}}^i \cdot b_{\text{SC}}^i \cdot b_{\text{LC}}^i$$

Each term corresponds to a filter layer: legibility, predictive alignment (SC), and deferred outcome divergence (LC). Coordination is thereby reframed as epistemic compression over influence-relevant distinctions: we retain only what matters to act well.

### 3.4 Strategic InfoNCE as Soft SEC Alignment

Strategic InfoNCE implements a learnable approximation to belief refinement over soft Strategic Equivalence Classes (SECs). Rather than reconstructing full partner models, agents learn to compress co-policies into abstract representations that preserve distinctions in *influence* i.e., how a co-agent deforms the ego agent's soft best response distribution.

**Definition 16 (Soft Strategic Divergence)** *Let $BR_i^\tau(\pi_{-i})$ denote the soft best response of agent $i$ to co-policy $\pi_{-i}$:*

$$BR_i^\tau(a_i \mid \pi_{-i}) := \frac{\exp(Q_i^{\pi_{-i}}(a_i)/\tau)}{\sum_{a'} \exp(Q_i^{\pi_{-i}}(a')/\tau)}$$

*Then the soft strategic divergence between two co-policies is:*

$$SD_i^\tau(\pi_{-i}, \pi'_{-i}) := D_{\text{KL}}(BR_i^\tau(\pi_{-i}) \,\|\, BR_i^\tau(\pi'_{-i}))$$

*This measures how distinguishable two co-policies are based on the incentive gradients they induce in the ego agent.*

**Definition 17 (Strategic InfoNCE Loss)** *To compress co-policies into an influence-aligned embedding space, we train a representation $f(\pi_{-i})$ to reflect behavioral similarity via:*

$$\mathcal{L}_{InfoNCE} = -\mathbb{E}\left[\log \frac{\exp(sim(f(\pi_{-i}), f(a^+)))}{\sum_{a^-} \exp(sim(f(\pi_{-i}), f(a^-)))}\right]$$

*where $a^+ \sim BR_i^\tau(\pi_{-i})$ and negatives $a^- \sim BR_i^\tau(\pi_{-j})$ for unrelated $\pi_{-j}$. The function $sim(\cdot,\cdot)$ denotes similarity in embedding space.*

**Lemma 2 (InfoNCE Minimization Yields Soft SEC Alignment)** *If $SD_i^\tau(\pi_{-i}, \pi'_{-i}) \leq \varepsilon$, then under sufficient optimization of $\mathcal{L}_{InfoNCE}$:*

$$\|f(\pi_{-i}) - f(\pi'_{-i})\| \text{ is small.}$$

*Thus, proximity in the embedding space implies approximate strategic substitutability.*

**Theorem 3 (Strategic InfoNCE Bounds Mutual Information)** *Minimizing $\mathcal{L}_{InfoNCE}$ maximizes a lower bound on:*

$$I_{BR} := I(a_i; \pi_{-i}) = \mathbb{E}_{\pi_{-i}}\left[D_{\text{KL}}(BR_i^\tau(\pi_{-i}) \| p(a_i))\right]$$

*Strategic InfoNCE thus preserves distinctions relevant for epistemic legibility.*

**Definition 18 (Abstraction Embedding)** *We define the agent-relative abstraction as:*

$$\phi_t^i := f(\pi_{-i})$$

*where $f$ is trained to reflect ego agent $i$'s incentive response to $\pi_{-i}$. This serves as a compressed surrogate for posterior belief over SECs.*

***Remark.*** *In practice, $\phi_t^i = f(h_t)$, where $h_t$ is the interaction history consistent with $\pi_{-i}$. Strategic InfoNCE thus learns to map histories to influence-equivalent regions. Since preferences are ego-relative, the embedding encodes* agent-conditioned *abstractions.*

**Policy Update under Abstraction.** At execution time, each agent conditions behavior on its abstraction $\phi_t^i$:

- **Low Influence Filter (high entropy)**: Treat co-agent as strategically indeterminate; act greedily over ego Q-values.
- **High SC (small divergence)**: Treat co-agent as locally aligned; follow expected incentive gradients.
- **High LC (long-run divergence)**: Trigger hedging or exploratory strategies to disambiguate latent influence.

This supports soft belief refinement over influence space without identity modeling or full simulation.

**Epistemic Interpretation.** Strategic InfoNCE is not just a contrastive loss. It implements a tractable, learnable approximation to BAMDP-style belief compression:

- $\phi_t^i = f(\pi_{-i})$ acts as a sufficient statistic for influence-relevant distinctions.
- $\mathcal{L}_{\text{InfoNCE}}$ induces a soft clustering of co-policies by their effect on best response.
- $I(a_i; \pi_{-i})$ measures partner legibility, an epistemic gain signal over relevance.

Strategic InfoNCE is not just a contrastive loss. It implements a tractable, learnable approximation to BAMDP-style belief compression.[1]

**Mapping to BAMDP and SER:**

- $\phi_t^i = f(h_t) \leftrightarrow$ belief state over SECs in BAMDP
- $\mathcal{L}_{\text{InfoNCE}} \leftrightarrow$ compressive approximation of VOO-preserving abstraction
- $I(a_i; \pi_{-i}) \leftrightarrow$ VOI over influence (epistemic gain from behavior legibility)
- $\varepsilon$-Soft SEC $\leftrightarrow$ soft posterior support over strategically indistinct policies

**Generalization.** Because $f$ is trained over behaviorally grounded preference traces, the resulting abstraction $\phi_t^i$ generalizes across unseen co-policies. Rather than encoding identity, it encodes relevance while making the agent robust to novel partners by aligning incentives, not features.

**Conclusion.** Strategic InfoNCE operationalizes the first layer of epistemic filtering: learning to compress co-agent behavior into actionable latent regions over SECs. It provides a tractable foundation for all subsequent abstraction filters by approximating belief updates over what matters to act without directly modeling who the partner is.

### 3.5 $\Delta\mathrm{SEC}(t)$ as a Signal of Strategic Compression

To act well under uncertainty, agents must not only refine their beliefs about the environment but also compress their uncertainty over *influence-relevant distinctions* in co-agent behavior. We introduce

$$\Delta\mathrm{SEC}(t)$$

as an epistemic signal: it measures how much ambiguity the agent has resolved about the *strategic equivalence class* (SEC) to which its partner is currently believed to belong.

This section formalizes $\Delta\mathrm{SEC}(t)$ as entropy reduction over a belief distribution on soft SEC membership, tracks its expected monotonicity under informative interaction, and interprets it as a surrogate regret signal for abstraction convergence.

[1] An open limitation is that InfoNCE may fail when co-policies induce adversarial incentive gradients i.e., when multiple $\pi_{-i}$ lead to indistinguishable short-term preferences but diverge sharply later. This motivates future work on adversarial SEC boundaries or curriculum-learning abstraction alignment.

**Soft SEC Regions.** Let $\mathrm{SD}_\tau(\pi_{-i}, \pi'_{-i}) := D_{\mathrm{KL}}(\mathrm{BR}_i^\tau(\pi_{-i}) \| \mathrm{BR}_i^\tau(\pi'_{-i}))$ denote the soft best response divergence between two co-policies.

**Definition 19 ($\delta$-Soft Strategic Equivalence Neighborhood)** *Given divergence threshold $\delta > 0$, define the $\delta$-neighborhood of $\pi_{-i}$ as:*

$$\mathcal{C}_\delta(\pi_{-i}) := \left\{ \pi'_{-i} \in \Pi_{-i} \;\middle|\; SD_\tau(\pi_{-i}, \pi'_{-i}) \leq \delta \right\}$$

*This class contains all co-policies that are approximately equivalent in their influence on the ego agent's decision geometry.*

**Belief over Influence Equivalence.** At time $t$, the agent maintains a belief distribution $B_t$ over these soft SEC neighborhoods.

**Definition 20 (Belief over SEC Neighborhoods)** *Let $\{\pi_{-i}^{(k)}\}_{k=1}^N \subset \Pi_{-i}$ be a representative set of co-policies. Then $B_t$ is defined via kernel density over learned embeddings:*

$$B_t(\mathcal{C}_\delta(\pi_{-i}^{(k)})) := \frac{1}{Z} \sum_{j=1}^{N} \mathbb{I}\left[ SD_\tau(\pi_{-i}^{(j)}, \pi_{-i}^{(k)}) \leq \delta \right] \cdot sim(f(h_t), f(\pi_{-i}^{(j)}))$$

*where $Z$ is a normalizing constant and sim is a similarity function (e.g., cosine or dot product).*

**Definition 21 (Strategic Compression Signal)** *Define:*

$$\Delta\mathrm{SEC}(t) := \mathcal{H}[B_{t-1}] - \mathcal{H}[B_t]$$

*where $\mathcal{H}[B_t]$ is the Shannon entropy of the belief distribution over soft SEC classes. A positive $\Delta\mathrm{SEC}(t)$ reflects epistemic progress defined as resolution of previously plausible but now implausible strategic hypotheses.*

**Proposition 1 (Entropy Drop Implies Structural Compression)** *Let $B_t$ and $B_{t-1}$ be consecutive belief distributions over soft SECs. Then:*

$$Support(B_t) \subsetneq Support(B_{t-1}) \quad \Rightarrow \quad \Delta\mathrm{SEC}(t) > 0$$

**Proof 2** *If the support of $B_t$ is a strict subset of $B_{t-1}$, at least one previously viable equivalence class has been ruled out. Since entropy is maximized under uniform distributions and decreases as mass concentrates, this implies $\mathcal{H}[B_t] < \mathcal{H}[B_{t-1}]$.*

**Proposition 2 (Information Bound on SEC Compression)**

$$\Delta\mathrm{SEC}(t) \leq I(\mathcal{C}_\delta; h_t)$$

*where $I(\mathcal{C}_\delta; h_t)$ is the mutual information between the observed interaction history and SEC class identity. Thus, $\Delta\mathrm{SEC}(t)$ is bounded by how much can be learned about influence equivalence from interaction.*

**Bayesian Framing.** This signal tracks convergence not in identity space, but in epistemic relevance space. $\Delta\mathrm{SEC}(t)$ reflects compression over beliefs conditioned on influence, not full policy recovery. In BAMDP terms, this is posterior compression over latent SEC identity given history $h_t$.

**Mapping to BAMDP Decomposition:**

- **Posterior belief:** $b_t^i(\pi_{-i}) \approx$ belief over soft SEC class
- **Epistemic uncertainty:** $\mathcal{H}[B_t]$ measures ambiguity in influence-relevant distinctions
- **Refinement progress:** $\Delta\mathrm{SEC}(t)$ tracks VOI-style belief gain
- **Convergence signal:** Serves as an epistemic alternative to regret under bounded rationality

**Mapping to SER Structure:**

- **Strategic Equivalence:** $\mathcal{C}_\delta(\pi_{-i})$ softens the binary relation $\pi_{-i} \sim_i \pi'_{-i}$
- **SER DAG traversal:** Entropy drop mirrors node pruning in recursive abstraction graphs
- **Abstraction embedding:** InfoNCE approximates preimage partitioning via similarity in BR response

**Interpretation.** $\Delta\mathrm{SEC}(t)$ is not an identity-tracking metric. It is an epistemic signal of abstraction sufficiency: it indicates whether the learned filters are converging toward a minimal yet sufficient representation of influence-relevant behavior.

**Relation to Recursive SER Graphs.** Prior work on Strategic Equivalence Relations (Lauffer et al., 2023) frames abstraction as a recursive DAG over histories, where each node defines a co-policy partition inducing the same best response. Our approach softens this structure into a continuous embedding space, enabling smooth belief compression via learned metrics.

While we do not construct the SER graph explicitly, $\Delta\mathrm{SEC}(t)$ can be interpreted as an implicit traversal: tracking how many decision-relevant clusters are ruled out over time. This opens the door for future work integrating metric embeddings with symbolic abstraction hierarchies.

**Modular Interpretation.** Beyond serving as a convergence metric, $\Delta\mathrm{SEC}(t)$ can be operationalized in multiple downstream roles:

- **Adaptive learning rate:** Increase gradient update confidence when $\Delta\mathrm{SEC}(t)$ is large.
- **Trust calibration:** Treat partners as more predictable when SEC entropy is low.
- **Coordination readiness:** Trigger joint planning only when ambiguity over strategic relevance falls below a threshold.

This modular view emphasizes that abstraction convergence is not just a training diagnostic, but a general-purpose epistemic signal for coordination scaffolding.

$\Delta\mathrm{SEC}(t)$ implements epistemic feedback for abstraction learning. It measures what the agent has ruled out, not about who the partner is, but about what distinctions matter. It is the soft counterpart to Bayes regret, interpreted here as **epistemic sufficiency under bounded abstraction capacity**. This signal grounds our abstraction frames in a learnable convergence metric aligned with BAMDP and SER principles.

### 3.6 Coordination as Resolving Relevance Ambiguity

Coordination under uncertainty is traditionally framed as inference over partner identity or equilibrium strategy. In contrast, we propose that coordination is fundamentally an epistemic process: it is not about modeling who the partner is, but about resolving ambiguity over which distinctions in their behavior are strategically relevant to the ego agent's decision geometry.

This shift reframes coordination from a prediction problem to an abstraction alignment problem. Our framework operationalizes this alignment through a layered architecture of abstraction filters: Influence Filter, Short-horizon Compatibility (SC), and Long-horizon Compatibility (LC). Each of these frames compresses the ego agent's interaction history $h_t$ into a latent abstraction $\phi_t^i$, preserving only influence-relevant distinctions.

**Epistemic Objective.** The central epistemic question becomes:

> What distinctions in partner behavior must be retained to act well under uncertainty?

Rather than tracking identity, the agent filters history through abstraction frames that identify which features of partner behavior induce nontrivial deformation in its own soft best response. Strategic

InfoNCE implements this filtering by learning an embedding space over co-policy influence, and $\Delta\mathrm{SEC}(t)$ quantifies convergence over that abstraction space.

**Operational Reframing.** Formally, coordination emerges when multiple agents reduce ambiguity over the relevance structure of their partners:

$$\text{Coordination} := \text{Reduction in strategic ambiguity over influence-relevant features}$$

This redefinition displaces equilibrium computation and identity modeling. What matters is epistemic alignment: agreement on which behavioral distinctions affect one's own incentives.

**Bounded Rationality and Tractable Abstraction.** Full belief maintenance is intractable in general-sum Dec-POMDPs. Our method approximates Bayes-Adaptive refinement through influence-conditioned filters:

- **IF:** Filters for legibility: co-policies that induce determinate ego responses.
- **SC:** Groups co-policies with similar short-term impact on ego preferences.
- **LC:** Disentangles co-policies with diverging long-term influence.
- **InfoNCE:** Learns abstraction embedding via contrastive alignment over strategic effect.
- $\Delta\mathrm{SEC}(t)$**:** Tracks abstraction convergence through entropy reduction in influence space.

Each of these approximates belief refinement under bounded abstraction capacity. Together, they form a layered scaffold for resolving epistemic ambiguity without full posterior inference.

**Mapping to BAMDP and SER.** Our approach is a natural descendant of BAMDPs and Strategic Equivalence Relations:

- BAMDPs decompose value into VOI and VOO; our filters align with this decomposition under abstraction.
- SER introduces equivalence classes over partner behavior; our soft divergence metrics generalize this to continuous embeddings.
- $\Delta\mathrm{SEC}(t)$ tracks epistemic refinement akin to Bayes regret, but in abstraction space rather than latent state space.

The goal of coordination is not full modeling of others, but alignment over what distinctions are relevant to preserve. Our method achieves this through layered abstraction filters, soft belief compression via Strategic InfoNCE, and epistemic convergence measured by $\Delta\mathrm{SEC}(t)$. Coordination is thus reinterpreted as an emergent property of epistemic alignment when agents agree not on identity, but on influence.

$$\textbf{Coordination} \equiv \textbf{Strategic Relevance Agreement}$$

This simple redefinition shapes the architecture of Strategic Abstractions itself, a framework whose own structure reflects the very process it seeks to model, boundedly rational agents learning what distinctions matter in order to coordinate under uncertainty.

## 4 Broader Impact

This work introduces a framework for scalable coordination under uncertainty by reframing the problem as one of strategic relevance, not identity or mimicry. Rather than modeling who a partner is or replicating their behavior, agents learn to abstract over what influences their own incentives, attending only to distinctions that reshape their decision boundaries. Coordination emerges not through policy matching but through alignment on what matters, offering a tractable and epistemically grounded alternative to traditional multi-agent learning in general-sum, partially observable settings.

By formalizing this process, Strategic Abstractions enables agents to navigate decentralized, asymmetric environments without shared goals, language, or centralized control. It supports flexible adaptation to novel partners, ad hoc teaming, and pluralistic cooperation through abstraction alignment rather than full-policy inference. Applications range from human–AI interaction and collaborative autonomy to distributed planning systems where agents must act under strategic ambiguity.

However, abstraction mechanisms that discard "irrelevant" distinctions also carry risks. Compressing strategic influence may unintentionally suppress minority behaviors, obscure alternative strategies, or be exploited to hide incentives in adversarial contexts. Because our method enables inference over latent influence patterns (without relying on identity tracking) it may pose novel threats in surveillance, manipulation, or asymmetric negotiation where contestability is limited.

More subtly, by defining coordination through shared relevance, the framework introduces a normative filter on which distinctions are considered actionable. These filters, if left unexamined, may encode implicit bias, favor dominant strategies, or marginalize perspectives that deviate from learned conventions. Ensuring that these abstractions remain transparent and contestable is critical.

To address these concerns, we advocate for continued development of abstraction-aware interpretability tools, influence-sensitive safety constraints, and protocols that allow agents to negotiate or contest relevance assumptions. Coordination must not presume shared understanding; it must earn it, by making the terms of alignment legible, revisable, and accountable across agents with divergent roles, information, or values.

## References


Christopher Amato. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning, 2024. URL https://arxiv.org/abs/2409.03052.

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Louise Barrett. *Beyond the brain: How body and environment shape animal and human minds*. Princeton University Press, 2011.

Richard Bellman and Robert Kalaba. A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290, 1959.

George W. Brown. Iterative Solution of Games by Fictitious Play, 1951. URL https://bibbase.org/network/publication/brown-iterativesolutionofgamesbyfictitiousplay-1951.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination, 2020. URL https://arxiv.org/abs/1910.05789.

Juin-Kuan Chong, Colin F. Camerer, and Teck-Hua Ho. Cognitive hierarchy: A limited thinking theory in games. In Rami Zwick and Amnon Rapoport (eds.), *Experimental Business Research*, pp. 203–228. Springer US, 2005. ISBN 978-0-387-24244-6. DOI: 10.1007/0-387-24244-9_9.

David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL https://arxiv.org/abs/2405.06624.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with Opponent-Learning Awareness, September 2018. URL http://arxiv.org/abs/1709.04326. arXiv:1709.04326 [cs].

György Gergely and Gergely Csibra. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, 7(7):287–292, 2003.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, February 2025. URL http://arxiv.org/abs/2502.14143. arXiv:2502.14143 [cs].

Nigel Howard. *Paradoxes of rationality: Theory of metagames and political behavior*. The MIT Press, 2003.

Hengyuan Hu, Adam Lerer, Brandon Cui, David Wu, Luis Pineda, Noam Brown, and Jakob Foerster. Off-Belief Learning, August 2021a. URL http://arxiv.org/abs/2103.04000. arXiv:2103.04000 [cs].

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-Play" for Zero-Shot Coordination, May 2021b. URL http://arxiv.org/abs/2003.02979. arXiv:2003.02979 [cs].

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Edward Hughes, Joel Z. Leibo, Matthew G. Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas, September 2018. URL http://arxiv.org/abs/1803.08884. arXiv:1803.08884 [cs].

Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019. ISSN 23521546. DOI: 10.1016/j.cobeha.2019.04.010. URL https://linkinghub.elsevier.com/retrieve/pii/S2352154618302055.

Niklas Lauffer, Ameesh Shah, Micah Carroll, Michael Dennis, and Stuart Russell. Who Needs to Know? Minimal Knowledge for Optimal Coordination, July 2023. URL http://arxiv.org/abs/2306.09309. arXiv:2306.09309 [cs].

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. *arXiv:1702.03037 [cs]*, February 2017. URL http://arxiv.org/abs/1702.03037. arXiv: 1702.03037.

Aly Lidayan, Michael Dennis, and Stuart Russell. BAMDP Shaping: a Unified Framework for Intrinsic Motivation and Reward Shaping, March 2025. URL http://arxiv.org/abs/2409.05358. arXiv:2409.05358 [cs].

James John Martin. *Some Bayesian decision problems in a Markov chain.* PhD thesis, Massachusetts Institute of Technology, 1965.

Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2006.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind, March 2018. URL http://arxiv.org/abs/1802.07740. arXiv:1802.07740 [cs].

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling Others using Oneself in Multi-Agent Reinforcement Learning, March 2018. URL http://arxiv.org/abs/1802.09640. arXiv:1802.09640 [cs].

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive pomdps. *Advances in neural information processing systems*, 20, 2007.

Michael Scaife and Jerome S Bruner. The capacity for joint visual attention in the infant. *Nature*, 253(5489):265–266, 1975.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5): 675–691, 2005.

Michael P. Wellman, Karl Tuyls, and Amy Greenwald. Empirical game theoretic analysis: A survey. *Journal of Artificial Intelligence Research*, 82:1017–1076, February 2025. ISSN 1076-9757. DOI: 10.1613/jair.1.16146. URL http://dx.doi.org/10.1613/jair.1.16146.

# Supplementary Materials

*The following content was not necessarily subject to peer review.*

## 5 Functional Definitions of Strategic Abstraction Objects

We define the core epistemic objects used to formalize strategic abstraction as a process of belief refinement over relevance classes. Each object operates over histories and encodes a structural transformation on the agent's inference space, rather than prescribing concrete actions.

**Definition 22 (Influence Filter)** *The **Influence Filter** is a belief-indexed abstraction function that maps interaction histories to equivalence classes over soft best response behavior. Histories are grouped if they induce indistinguishable posterior soft best responses:*

$$\mathcal{F}_{\textit{influence}} : \mathcal{H}_t \to \mathcal{A}_{\textit{filter}} \subseteq \Sigma^{\textit{soft-BR}} \tag{1}$$

*where:*

- $\mathcal{H}_t$ *is the space of interaction histories up to time* $t$*,*
- $\mathcal{A}_{\textit{filter}}$ *is an abstraction space partitioning histories into relevance-equivalent soft BR classes,*
- $\Sigma^{\textit{soft-BR}}$ *is the space of stochastic best response strategies.*

*The induced equivalence relation is:*

$$h, h' \in \mathcal{H}_t \quad \textit{s.t.} \quad \pi^*(\cdot \mid h) \approx \pi^*(\cdot \mid h') \quad \Rightarrow \quad \mathcal{F}_{\textit{influence}}(h) = \mathcal{F}_{\textit{influence}}(h') \tag{2}$$

*That is, two histories are assigned to the same abstraction class if they yield indistinguishable beliefs about how co-agent structure deforms the agent's soft best response.*

**Definition 23 (Short-Term Strategic Compatibility (STSC))** *The **STSC** functional evaluates local compatibility between agent and co-agent strategies, conditioned on a fixed abstraction filter. It measures whether the co-agent's behavior lies within a strategically tolerable region of influence:*

$$\textit{STSC} : (\pi_i, \pi_j, \phi_t) \mapsto \mathbb{R} \tag{3}$$

*where:*

- $\pi_i$ *is the agent's soft best response policy under current abstraction* $\phi_t$*,*
- $\pi_j$ *is the co-agent's policy inferred from history,*
- $\phi_t$ *is a fixed abstraction filter determining relevance classes (e.g., SF or SC layer),*
- *Output is a scalar compatibility score indicating the extent to which* $\pi_j$ *lies within the abstraction-conditioned best response basin of* $\pi_i$*.*

*STSC encodes a localized incentive alignment test: it is high when* $\pi_j$ *does not induce deviation from the soft best response trajectory defined by* $\phi_t$*, and low otherwise.*

Operationally:

$$\text{STSC}(\pi_j \mid \pi_i, \phi_t) \propto \text{softBR}_i(\pi_j \mid \phi_t) \tag{4}$$

**Abstraction Layer Decision Boundaries**

Each abstraction layer (Intent Certainty, Short-Term Strategic Compatibility, Long-Term Strategic Compatibility) defines a decision boundary over relevance-filtered beliefs.

Let $\delta_{\text{Layer}}$ be the indicator function for whether a decision boundary is active:

$$\delta_{\text{Layer}} : \mathcal{H}_t \mapsto \{0, 1\} \tag{5}$$

For example, the Influence Filter (IF) gate is:

$$\delta_{\text{IF}}(h_t) = \begin{cases} 1 & \text{if } \mathcal{H}_t \text{ supports a confident influence filter class} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

These gates control whether abstraction-filtered soft best response updates are permitted at time $t$.

# 6 Operationalization: Formalization of Abstraction Filters as Epistemic Decision Boundaries

We define the two primary abstraction filters in our framework: Influence Filter (IF) and Short-horizon Compatibility (SC), as epistemic decision boundaries over the Bayes-Adaptive belief state. These filters gate whether co-agent influence is modeled explicitly in the agent's update, and are grounded in principles of epistemic sufficiency: an abstraction is activated when incorporating co-agent structure is expected to reduce uncertainty or improve the agent's value estimate under bounded rationality.

## 6.1 Influence Filter (IF)

Influence Filter defines when the agent has sufficiently collapsed its belief over the co-agent's behavior such that further modeling is justified. Formally:

$$\text{IF}_t^{i \leftarrow j} := \mathbb{I}\left[H\left(\hat{\pi}_j(\cdot \mid h_t^i)\right) < \tau_{\text{IF}}\right] \tag{7}$$

where:

- $\hat{\pi}_j(\cdot \mid h_t^i)$ is agent $i$'s posterior over agent $j$'s policy, induced by observation–action history $h_t^i$;
- $H(\cdot)$ denotes Shannon entropy;
- $\tau_{\text{IF}}$ is an entropy threshold that encodes an epistemic sufficiency condition: the minimum collapse required to consider partner intent legible.

IF is conceptually tied to belief convergence in BAMDPs: it activates when the posterior over co-agent behavior is concentrated enough to support strategic modeling.

## 6.2 Short-horizon Compatibility (SC)

SC determines whether incorporating co-agent information substantively deforms the ego agent's best response. It is defined by the divergence between action distributions with and without conditioning on the co-agent model:

$$\text{SC}_t^{i \leftarrow j} := \mathbb{I}\left[\text{KL}\left(\pi_i^{\text{BR}|\emptyset} \,\|\, \pi_i^{\text{BR}|\hat{\pi}_j}\right) > \tau_{\text{SC}}\right] \tag{8}$$

where:

- $\pi_i^{\text{BR}|\emptyset}$ is the soft best response policy of agent $i$ assuming a uniform prior over co-agent behavior;
- $\pi_i^{\text{BR}|\hat{\pi}_j}$ is the soft best response policy conditioned on the belief $\hat{\pi}_j$;
- $\text{KL}(\cdot \,\|\, \cdot)$ is the Kullback-Leibler divergence;

- $\tau_{\mathrm{STSC}}$ is a threshold representing minimal influence relevance.

SC thus evaluates whether the partner model $\hat{\pi}_j$ materially alters the ego agent's decision boundary i.e., whether co-agent influence should be treated as strategically relevant.

### 6.3 Strategic InfoNCE and Gated Alignment

These gates are used to selectively activate contrastive alignment updates. The loss is defined as:

$$\mathcal{L}^i_{\text{InfoNCE}} = \text{IF}_t^{i \leftarrow j} \cdot \text{SC}_t^{i \leftarrow j} \cdot \text{InfoNCE}\left((\phi_t^j, a_i^+), \{(\phi_t^k, a_i^-)\}_{k \neq j}\right) \tag{9}$$

where:

- $\phi_t^j$ is the co-agent embedding at time $t$;
- $a_i^+ \sim \pi_i^{\text{BR}|\hat{\pi}_j}$ is a positive action under the best response to $\hat{\pi}_j$;
- $a_i^-$ are negative actions under best responses to other agents or alternative abstractions.

Only when both gates are active does the agent align embeddings based on strategic relevanc defined by epistemic collapse (IF) and incentive deformation (SC).

### 6.4 Theoretical Note: Sufficient Approximation to BAMDP Value Alignment

Let $Q^{\hat{\pi}_j}$ and $Q^{\emptyset}$ denote value functions under BAMDP belief states with and without modeling $\hat{\pi}_j$. Then:

$$\mathbb{E}_{s,a}[Q^{\hat{\pi}_j}(s,a) - Q^{\emptyset}(s,a)] > \delta \quad \Rightarrow \quad \text{SC}_t^{i \leftarrow j} = 1$$

for some $\delta$ defined by $\tau_{\text{SC}}$. Thus, activation of SC can be interpreted as a minimal regret-reduction justification for modeling effort. Similarly, IF activation corresponds to posterior belief entropy falling below a decision-theoretic sufficiency threshold.

**Theorem 4** *Let agent $i$ maintain a belief $\hat{\pi}_j$ over co-agent $j$'s policy at time $t$, and let $h_t^i$ be agent $i$'s observation–action history. Define:*

- $\pi_i^{BR|\hat{\pi}_j}$*: agent $i$'s soft best response conditioned on $\hat{\pi}_j$;*
- $\pi_i^{BR|\emptyset}$*: soft best response under an uninformative co-agent model (e.g., uniform or marginalized prior);*
- $Q_i^{\psi}(s,a)$*: agent $i$'s estimated Q-value when conditioning on co-agent model $\psi$;*
- $H(\hat{\pi}_j)$*: Shannon entropy over $\hat{\pi}_j$;*
- $d_t$*: agent $i$'s marginal state distribution at time $t$.*

***Assumption 2*** *1.* ***Bounded Rationality****: $\pi_i^{BR|\psi}(a) \propto \exp\left(\frac{1}{\beta} Q_i^{\psi}(s,a)\right)$ for some $\beta > 0$.*

2. ***Monotonic Sensitivity****: For fixed $\beta$, the probability mass of $\pi_i^{BR|\psi}$ increases on actions with higher $Q_i^{\psi}$.*
3. ***Influence Filter Activation (IF)****: The co-agent's behavior is legible:*

$$IF_t^{i \leftarrow j} := \mathbb{I}\left[H\left(\hat{\pi}_j(\cdot \mid h_t^i\right) < \tau_{\text{IF}}\right] = 1$$

4. ***Short-horizon Compatibility Activation (SC)****: The co-agent's predicted behavior deforms the decision boundary:*

$$SC_t^{i \leftarrow j} := \mathbb{I}\left[\text{KL}\left(\pi_i^{BR|\emptyset} \,\|\, \pi_i^{BR|\hat{\pi}_j}\right) > \tau_{\text{SC}}\right] = 1$$

:

*Then under the joint activation of IF and SC, the expected value of acting under $\pi_i^{BR|\hat{\pi}_j}$ strictly improves over the uninformed baseline:*

$$\mathbb{E}_{s\sim d_t, a\sim\pi_i^{BR|\hat{\pi}_j}}\left[Q_i^{\hat{\pi}_j}(s,a)\right] > \mathbb{E}_{s\sim d_t, a\sim\pi_i^{BR|\emptyset}}\left[Q_i^{\hat{\pi}_j}(s,a)\right] - \epsilon$$

*for some $\epsilon > 0$ that decreases monotonically as $H(\hat{\pi}_j) \to 0$.*

### 6.5 Interpretation

When IF is active, the partner's behavior is sufficiently predictable to support modeling. When SC is active, modeling meaningfully alters the ego agent's strategy. Under bounded rationality, this guarantees that conditioning on $\hat{\pi}_j$ leads to improved expected value in the ego agent's decision. The abstraction gates thus define a decision boundary in belief space beyond which coordination modeling becomes epistemically justified.

**Lemma 3 (Gated Abstraction Reduces SEC Ambiguity)** *Let agent $i$ maintain a belief $b_t^{i\leftarrow j}$ over a finite set of Strategic Equivalence Classes $\mathcal{C}_j = \{C_1, \ldots, C_k\}$ describing possible strategic roles or incentive-relevant behavior patterns for co-agent $j$.*

*Suppose:*

- *At time $t$, the Influence Filter gate $\mathrm{IF}_t^{i\leftarrow j} = 1$;*
- *The Short-horizon Compatibility gate $\mathrm{SC}_t^{i\leftarrow j} = 1$;*
- *A co-agent embedding $\phi_t^j$ is updated via Strategic InfoNCE with positive sample $(\phi_t^j, a_i^+)$ and negatives $(\phi_t^k, a_i^-)$ for $k \neq j$.*

*Then under the assumption that co-agent embeddings preserve influence-relevant distinctions, the expected entropy of the belief over SEC class assignment decreases:*

$$\mathbb{E}\left[H(b_{t+1}^{i\leftarrow j})\right] < H(b_t^{i\leftarrow j})$$

*That is, updating $\phi_t^j$ gated by IF and SC leads to expected refinement of the ego agent's belief over which strategic role class the partner belongs to.*

## 7 $\Delta\mathrm{SEC}(t)$ as Regret Reduction

We reinterpret $\Delta\mathrm{SEC}(t)$ not merely as an agreement score over abstraction layers, but as an operational signal for resolving ambiguity over Strategic Equivalence Classes (SECs). This metric quantifies the epistemic utility of a partner abstraction $\phi_t^j$ by comparing induced regret under different co-agent models.

Formally, define the regret under a soft best response to a co-agent model $\psi$ as:

$$\mathrm{Regret}_{\mathrm{BR}_i}(a \mid \psi) := Q_i^{\psi}(s,a) - \mathbb{E}_{a'\sim\pi_i^{\mathrm{BR}_\psi}}[Q_i^{\psi}(s,a')]$$

We then define the differential regret of the abstraction-filtered model $\phi_t^j$ relative to a reference co-policy model $\hat{\pi}_j$ as:

$$\Delta\mathrm{SEC}(t) := \mathbb{E}_{s,a}\left[\mathrm{Regret}_{\mathrm{BR}_i}(a \mid \hat{\pi}_j) - \mathrm{Regret}_{\mathrm{BR}_i}(a \mid \phi_t^j)\right]$$

This formulation interprets $\Delta\mathrm{SEC}(t)$ as an expected regret reduction achieved by compressing the co-agent policy model into an abstraction $\phi_t^j$ that preserves distinctions relevant to the ego agent's

best response. When $\Delta\mathrm{SEC}(t) > 0$, the abstraction induces a policy that performs at least as well as the full model in expectation, suggesting that $\phi_t^j$ constitutes a sufficient filter over influence-relevant behavior.

We hypothesize that monotonic increases in $\Delta\mathrm{SEC}(t)$ reflect convergence in belief over the partner's SEC class. Empirically, this should correlate with improvements in coordination performance across tasks with strategic ambiguity (e.g., Overcooked, Harvest). This connects belief compression to alignment, and makes $\Delta\mathrm{SEC}(t)$ a candidate objective for learning abstraction layers that resolve relevance ambiguity.

## 8 Defining $\Delta\mathrm{SEC}(t)$ as a Regret-Reducing Abstraction Alignment Signal

The Strategic Equivalence Relation (SER) framework formalizes when modeling a co-agent is unnecessary: namely, when the partner's behavior does not deform the ego agent's incentive landscape. In Strategic Abstractions, the goal is not just to detect equivalence but to iteratively refine beliefs over abstraction layers that compress strategically irrelevant details while preserving influence-relevant ones.

To connect abstraction refinement with performance, we define a scalar metric, $\Delta\mathrm{SEC}(t)$, that quantifies the alignment between an agent's learned abstraction $\phi_t^j$ of its co-agent and the full behavioral model $\hat{\pi}_j$. Crucially, this metric is not defined over agreement in behavior, but over induced value: how abstraction affects the ego agent's own policy performance.

### 8.1 Definition: Regret-Reducing Abstraction Alignment

Let:

- $\phi_t^j$: abstraction-filtered model of co-agent $j$ at time $t$;
- $\hat{\pi}_j$: full model or empirical estimate of co-agent $j$'s behavior;
- $\pi_i^{\mathrm{BR}|\psi}$: soft best response of agent $i$ conditioned on co-agent model $\psi$;
- $Q_i^{\psi}(s, a)$: agent $i$'s value estimate when conditioning on $\psi$.

We define the differential regret of using $\phi_t^j$ in place of $\hat{\pi}_j$ as:

$$\Delta\mathrm{SEC}(t) := \mathbb{E}_{s,a}\left[Q_i^{\pi_i^{\mathrm{BR}|\phi_t^j}}(s, a) - Q_i^{\pi_i^{\mathrm{BR}|\hat{\pi}_j}}(s, a)\right]$$

This expression measures how much expected value agent $i$ loses or gains by acting under the abstraction $\phi_t^j$ rather than the full model $\hat{\pi}_j$. A positive $\Delta\mathrm{SEC}(t)$ implies that $\phi_t^j$ is a sufficient and possibly more generalizable representation of co-agent influence.

### 8.2 Interpretation and Empirical Use

- When $\Delta\mathrm{SEC}(t) > 0$, the abstraction induces equal or better policy performance relative to the full model. This implies that $\phi_t^j$ retains the influence-relevant distinctions and discards strategically inert variation.
- As coordination proceeds and agents refine their representations, we expect $\Delta\mathrm{SEC}(t)$ to increase signaling abstraction alignment.
- When paired with new partners, a decline in $\Delta\mathrm{SEC}(t)$ should correspond to performance degradation, while recovery in $\Delta\mathrm{SEC}(t)$ should precede or predict regained coordination.

This makes $\Delta\mathrm{SEC}(t)$ not only an internal validation signal for abstraction quality but also a **predictive metric of coordination readiness**. Unlike surface-level agreement metrics, it anchors abstrac-

tion learning in value difference: it asks whether the abstraction changes what the ego agent wants to do and whether that change improves performance.

### 8.3 Connection to SER

Where SER provides the minimal conditions under which modeling is unnecessary, $\Delta\mathrm{SEC}(t)$ serves as a continuous signal for when abstraction is sufficient. In this sense, it is the dual of strategic equivalence: not a binary filter over whether modeling is needed, but a gradient of abstraction relevance, grounded in regret.

This connection enables a soft extension of SER to learning settings, where agents do not reason over exact SEC classes, but iteratively compress and refine over observed behaviors. By quantifying abstraction alignment in terms of expected regret, $\Delta\mathrm{SEC}(t)$ provides an operational bridge between value-theoretic sufficiency and practical abstraction learning.