A Unified Theory of Strategic Representations

Sandy Tanwisuth, Meridian Goodman *

April 1, 2025

1 Abstract

In multi-agent learning, agents must form representations of their co-players. But modeling full policies is often unnecessary, unstable, or intractable. Instead, we ask: What is ***the least and (closest to) all*** I need to know about you in order to coordinate well? This paper presents a theoretical framework for answering that question through a hierarchy of strategic abstractions— compressed, behaviorally grounded representations of how one agent influences another's incentives, outcomes, and available actions. Each layer in the hierarchy (intent clarity, local incentive similarity, long-term outcome equivalence, structural world dynamics) formalizes **a distinct inductive bias for coordination under uncertainty**. Together, they define a principled, learnable, and architecture-agnostic substrate for robust coordination across all game. Whether agents are biological or artificial, alignment emerges not from shared goals or architectures, but from compatible compressions of what strategically matters. This work provides a foundation for aligning diverse agents through **minimal mutual understanding, enabling scalable coordination across environments, embodiments, and agent types**.

2 Introduction

Motivation. Have you ever wondered why reinforcement learning agents have made such remarkable progress in zero-sum games like Go, Chess, or Poker? Why strategic reasoning in competitive settings often appears more tractable than in cooperative ones? At first glance, this may seem paradoxical—shouldn't helping others be easier than outsmarting?

But the key insight is this: selfish agents are often **easier to predict**. In zero-sum environments, each agent is transparently optimizing for its own payoff. Their goals are self-contained, their actions tightly coupled to their interests. In contrast, agents that ***intend*** to help—such as assistive robots or cooperative LLMs—frequently underperform, miscoordinate, or even hinder the very partner they aim to support.

Why? Because **"helpfulness"** without **strategic clarity** is dangerous. When an agent's actions do not reliably reveal its intentions, or when its influence on the environment is inconsistent, it becomes harder to coordinate—even if its motives are aligned. This gap between **strategic influence** and **surface-level intent** lies at the heart of why collaboration is often more fragile than competition. Our framework embraces this asymmetry by modeling not just what agents do, but how ***legibly*** they affect others' decisions and outcomes. Our framework attempt to answer this question: "**How do your actions shape what I know and what I can do in response (what is best for me to do in response)?**".

Ultimately, this framework lays the **foundation for general-purpose coordination—across *agents* and across all types of games** that may be biological, artificial, or hybrid! By grounding trust in behaviorally meaningful abstractions rather than surface-level signals, it enables agents to **infer, adapt to, and align with others through structured interaction**. We envision this as a critical step toward

*We thank Niklas Lauffer for on-going advices and support during the development of this project. We thank CHAI for the community and compute resources with special thanks to Cassidy Laidlaw, Michelle Li, Karim Abdelsadek, Eli Bronstein, and Cam Allen for their advices. We also thanks MATS for the community and special thanks to Alex Cloud and Kola Ayonrinde. Draft of a work-in-progress shared for transparency and attribution. Not affiliated with any organization. Authored independently. Please contact the first author at sandy314@duck.com for licensing, collaboration, or reuse.

collective alignment: the ability for diverse agents to build mutual understanding, anticipate one another's strategic influence, and coordinate effectively—even in the absence of shared architecture, language, or reward functions.

Real Intro. Strategic coordination in multi-agent environments requires agents to form simplified yet decision-relevant representations of their co-players' behaviors. Modeling full co-player policies is often **intractable and unnecessary**: agents rarely need to know exactly what their partner will do, only how their behavior affects strategic outcomes. In this work, we propose a unified framework for strategic abstraction learning: a structured approach to **compressing co-player policies into progressively refined, behaviorally meaningful representations.**

Rather than asking what an agent *must know* to coordinate optimally, we ask: ***Given what an agent currently knows, is this representation sufficient to coordinate effectively—or is further refinement needed to disambiguate the co-player's strategic influence?*** This question motivates a hierarchy of abstractions that capture increasingly fine-grained notions of strategic relevance. Each level introduces an inductive bias that **filters**, **compresses**, or **refines co-player behavior** according to the demands of coordination.

We formalize this process using tools from reinforcement learning, information theory, and representation learning. Specifically, we introduce:

- **Intent Certainty via Mutual Information (IC)** — a mutual information-based filter that identifies co-policies with clear, high-confidence influence signal on the ego agent's behavior.
- **Strategic InfoNCE** — a contrastive learning objective that learns compressed representations of co-policies based on their short-term influence on ego agent actions (via soft best response distributions).
- **SFR-SS (Successor Feature Representation Strategic Similarity)** — a trajectory-level abstraction that groups co-policies according to the long-term features they induce, independent of reward.
- **MASR-SS (Model-Aware Strategic Strategic Similarity)** — a refinement that considers how co-player policies affect the ego agent's world dynamics, supporting planning and model-based generalization.
- **Strategic Refinement as Learning Dynamics** — a principled method for measuring the reduction in strategic ambiguity over time, as learned representations evolve from coarse filters to structured, goal-relevant abstractions.

Together, these components define **a general-purpose theory of strategic abstraction through learning** The resulting framework enables agents to **coordinate** robustly across behaviorally diverse partners, adaptively **refine** its own beliefs, and **act** based on the **minimum necessary knowledge** for strategic success. Our results provide both theoretical guarantees and practical tools for scalable, interpretable coordination across fully cooperative (common-payoff), mixed-motive (general-sum), and competitive (zero-sum) multi-agent settings. Crucially, by learning compressed representations that retain only strategically relevant information, this framework also enables **zero-shot transfer of learned policy abstractions** across new environments and partners—allowing agents to generalize coordination strategies without retraining from scratch.

2.1 Connection to Social Sciences

The Role of Agreement Across Abstractions. Each level of abstraction in our framework captures a distinct facet of strategic influence: intent certainty (IC), short-term behavioral incentives (strategic CPC), long-term outcome trajectories (SFR-SER), and induced world structure (MASR-SER). When all levels align—i.e., when the agent's signal is clear, its behavior elicits **predictable best responses**, leads to desirable long-run outcomes, and induces stable transition dynamics—then the ego agent can act with confidence. This convergence constitutes a form of **alignment across abstraction layers**, indicating that the co-player's behavior is interpretable, coherent, and trustworthy.

Conversely, if these abstractions diverge—e.g., if a co-player's actions suggest cooperation but ultimately undermine long-term goals or distort the transition model—then the agent should recognize the mismatch and either abstain, seek clarification, or learn more. In this way, **our framework offers an operational and computational interpretation of trust**: strategic consistency across abstraction layers provides a data-driven, learnable basis for determining whether to coordinate with or adapt to a given co-policy.

Remark 1 (Stereotypes as Strategic Abstractions)**.** *Our framework also provides a computational lens on the formation, utility, and refinement of stereotypes. Here, stereotypes are not treated as socially loaded judgments, but as* ***functional abstractions: compressed representations of co-player behavior that retain only decision-relevant influence.*** *Each abstraction layer—IIE, CPC, SFR, MASR—defines a different form of stereotype, grounded in increasingly rich models of strategic interaction. Importantly, the agent refines these stereotypes over time, measuring their predictive adequacy through entropy reduction ($\Delta_{SEC}(t)$). This offers a formal answer to when stereotypes are useful, when they should be updated, and how they support trust under uncertainty.*

Broader Impact

This work contributes foundational tools for building trustworthy, coordination-ready agents in open multi-agent environments. By redefining alignment not in terms of fixed preferences or goals, but through a hierarchy of learnable, strategic abstractions, we provide a new operational language for intent modeling, influence detection, and trust calibration. Our framework supports scalable and interpretable coordination across biological, artificial, and hybrid systems—including LLMs, robots, humans, and agents of unknown or uncertain type.

One of the most pressing concerns in AI safety is not just individual misalignment, but *multi-agent miscoordination*: the possibility that independently trained, goal-oriented agents will fail to coordinate despite nominal alignment with their respective objectives. This includes risks identified in the "multi-polar failure" scenario (Critch, 2020), where even well-intentioned agents act at cross-purposes due to insufficient shared modeling or incompatible abstractions. Our theory directly addresses this by formalizing the minimum conditions under which strategic understanding, anticipation, and trust become possible—and when further refinement is required to prevent breakdowns.

In doing so, we offer a learnable substrate for what we might call *computational coordination*: the ability to coordinate effectively under uncertainty, across boundaries of architecture, embodiment, or training data. We envision this as a critical building block for long-term alignment: one that enables agents to recognize when they are truly aligned—not because they have the same utility function, but because they converge on the same actionable understanding of how to interact in a shared world.

3 Strategic Equivalence as Compression

Let Π_{-i} be the space of co-player policies. The central idea behind strategic abstraction is to learn a function $h : \Pi_{-i} \to \mathbb{R}^d$ that discards all information in π_{-j} not relevant to the best-response set $BR_i(\pi_{-i})$. This function induces a clustering of policies into what we call Strategic Equivalence Classes (SECs).

Definition 1 (Strategic Equivalence Class (this is definition 4.3 from SER paper Lauffer et al., 2023))**.** *Two policies π_{-i}, π'_{-i} are strategically equivalent, written $\pi_{-i} \sim_i \pi'_{-i}$, if they induce the same best response set:*

$$BR_i(\pi_{-i}) = BR_i\left(\pi'_{-i}\right).$$

This relation defines a partition of Π_{-i}, forming the coarsest possible abstraction under which the best response of agent i is invariant.

Remark 2. *This structure mirrors the Information Bottleneck (IB) principle, where $Z = BR_i(\pi_i)$ acts as a compressed representation retaining only behaviorally relevant content.*

4 Soft Best Response and Strategic Divergence

Prose: We introduce the **soft best response distribution** as a principled way to model how real-world agents, particularly those trained via reinforcement learning, respond to the strategies of their co-players in a more **realistic** and **tractable** manner. Unlike the hard best response, which assumes perfect rationality and yields a deterministic action choice, the soft best response accounts for ***bounded rationality***, ***noise***, and ***uncertainty*** in strategic decision-making.

This stochastic formulation enables us to represent ambiguity in an agent's incentives, which is crucial when reasoning about coordination under limited information or in the early stages of learning. It also supports smooth, differentiable computations, making it compatible with gradient-based learning algorithms and contrastive representation objectives like InfoNCE.

We specifically choose the **Boltzmann (or softmax)** distribution because it provides a well-understood, temperature-controlled interpolation between uniform random behavior (at high temperatures) and deterministic maximization (as temperature approaches zero meaning that the agent is fully rationale). This functional form has desirable mathematical properties—it is continuous, differentiable, and monotonic in the Q-values—allowing us to define soft notions of best response that are both **analytically tractable** and **behaviorally interpretable**. Moreover, it **aligns naturally with established models of bounded rationality** in behavioral game theory and biological reinforcement learning, offering a principled foundation for defining soft strategic equivalence classes and reasoning about graded, rather than binary, distinctions in strategic behavior.

Definition 2 (Soft Best Response Distribution)**.** *Let $Q^{\pi_{-i}}(a_i)$ denote agent i's expected value for action a_i under co-policy π_{-i}, and let $\tau > 0$ be a temperature parameter. The soft best response distribution is defined as:*

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

This defines a probability distribution over agent i's action space A_i.

Explanation. We now describe the components of the soft best response distribution in detail:

- $Q^{\pi_{-i}}(a_i) \in \mathbb{R}$ denotes **the expected return** to agent i when taking action $a_i \in A_i$ against co-player policy $\pi_{-i} \in \Pi_{-i}$, under a fixed environment and reward structure. Formally, this quantity represents the value of agent i's action under the assumption that the other agents follow the joint policy π_{-i}, and that the agent behaves according to a one-step deviation from a default policy by playing a_i.

  This value captures how favorable a given action a_i is in expectation, taking into account both the stochastic dynamics of the environment and the interactive behavior of co-players. We assume that $Q^{\pi_{-i}}(a_i)$ is finite and bounded, typically within the reward range defined by the environment (e.g., if per-timestep rewards lie in $[r_{\min}, r_{\max}]$, then so does $Q^{\pi_{-i}}(a_i)$). These values are fundamental to strategic reasoning and learning: they allow the agent to rank actions and compute probabilistic preferences over them through mechanisms like softmax, which in turn enables smooth adaptation in uncertain or changing environments.

- $\mathrm{BR}_i^\tau(a_i \mid \pi_{-i}) \in [0, 1]$ defines **a probability distribution** over agent i's action space A_i, where actions with higher Q-values under co-policy π_{-i} are assigned higher probabilities. The softmax function used here is strictly positive and normalized, which guarantees that every action $a_i \in A_i$ receives non-zero probability for any finite $Q^{\pi_{-i}}(a_i)$ and temperature $\tau > 0$. This ensures the distribution has *full support* over A_i, meaning that the agent never entirely rules out any action.

  This full support property is especially important in early stages of learning or when facing high uncertainty: it allows the agent to maintain exploratory flexibility, sampling all available actions while still biasing toward higher-value ones. Over time, as the agent becomes more confident (e.g., due to better estimates of $Q^{\pi_{-i}}$), the distribution can naturally become more peaked, smoothly transitioning toward deterministic behavior as $\tau \to 0$. Thus, the soft best response not only represents graded preferences but also provides a built-in mechanism for balancing exploitation and exploration.

- $\tau > 0$ is a scalar temperature parameter that controls the entropy of the distribution. As $\tau \to 0$, the distribution becomes sharply peaked on the highest-valued action, recovering the deterministic best response. As $\tau \to \infty$, the distribution approaches the uniform distribution over A_i, reflecting maximum entropy and indifference between actions.

  Maximum entropy here corresponds to maximum uncertainty. From the perspective of agent i, this uncertainty reflects a lack of useful information about which action is preferable. In such a case, the agent behaves as though all actions are equally likely to be optimal, leading to uniform random exploration. This interpretation aligns with early learning phases, where the agent may play arbitrary actions and learn from the outcomes.

- The soft best response operator defines a mapping from co-player policies $\pi_{-i} \in \Pi_{-i}$ to probability distributions over actions:

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

  where $\Delta(A_i)$ denotes the probability simplex over A_i.

  Intuitively, this means that for any given co-player strategy π_{-i}, agent i responds not with a single deterministic action, but with a distribution over actions that reflects both the expected values of each action and the agent's uncertainty (as modulated by τ). This formulation captures the idea that even when an agent has preferences over actions, it may still randomize among them when those preferences are weak or noisy, or when strategic exploration is beneficial.

This soft best response distribution provides a continuous and differentiable relaxation of the classic (hard) best response operator. It captures notions of bounded rationality and stochasticity in strategic behavior, and forms the basis for defining soft strategic equivalence relations and learning-based coordination strategies.

Prose: Intuitively, a low strategic divergence implies that the two co-player policies lead to nearly identical preferences over actions for agent i, while a high divergence indicates substantial differences in how agent i would act in response. Crucially, due to the softmax's invariance to additive shifts in Q-values (Lemma 1), the strategic divergence depends only on the *relative* preferences among actions—not their absolute magnitudes. This interpretation motivates the use of soft BR distributions as a **behaviorally grounded, differentiable proxy** for measuring strategic similarity. Theorem 1 formalizes this connection: if two co-player policies induce identical soft best response distributions (i.e., zero KL divergence), they must induce the same hard best response set, and therefore belong to the same strategic equivalence class (under a specific condition).

Definition 3 (Strategic Divergence)**.** *Let $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ be two co-player policies. Their strategic divergence from the perspective of agent i is defined as:*

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

Lemma 1 (Softmax Invariance to Additive Shift)**.** *Let $Q \in \mathbb{R}^n$ be a vector of real-valued scores. The softmax over Q with temperature τ is invariant to additive shifts:*

$$Softmax_\tau(Q + c \cdot \mathbf{1}) = Softmax_\tau(Q) \quad \text{for any constant } c \in \mathbb{R}$$

Proof. We compute:

$$\mathrm{Softmax}_\tau(Q_i + c) = \frac{\exp((Q_i + c)/\tau)}{\sum_j \exp((Q_j + c)/\tau)} = \frac{\exp(Q_i/\tau) \cdot \exp(c/\tau)}{\sum_j \exp(Q_j/\tau) \cdot \exp(c/\tau)} = \mathrm{Softmax}_\tau(Q_i)$$

□

Remark 3. *Two Q-vectors that differ only by an additive constant are behaviorally equivalent under the soft best response distribution. That is, if $Q = Q' + c \cdot \mathbf{1}$, then $Softmax_\tau(Q) = Softmax_\tau(Q')$. This justifies that soft best responses encode only relative preferences between actions, not absolute values.*

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

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

Then agent i's best response sets are equal:

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

Proof. Let $p(a_i) := \mathrm{BR}_i^\tau(a_i \mid \pi_{-i})$ and $q(a_i) := \mathrm{BR}_i^\tau(a_i \mid \pi'_{-i})$. By assumption, $D_{\mathrm{KL}}(p\|q) = 0$, which implies $p(a_i) = q(a_i)$ for all a_i.

Since soft best responses are softmax distributions over Q-values, this means that:

$$\frac{\exp(Q^{\pi_{-i}}(a_i)/\tau)}{\sum_j \exp(Q^{\pi_{-i}}(a_j)/\tau)} = \frac{\exp(Q^{\pi'_{-i}}(a_i)/\tau)}{\sum_j \exp(Q^{\pi'_{-i}}(a_j)/\tau)}$$

This equality holds only when:

$$Q^{\pi_{-i}}(a_i) = Q^{\pi'_{-i}}(a_i) + c \quad \forall a_i, \text{ for some } c \in \mathbb{R}$$

By Lemma 1, this shift-invariance implies identical best response sets:

$$\arg\max_{a_i} Q^{\pi_{-i}}(a_i) = \arg\max_{a_i} Q^{\pi'_{-i}}(a_i)$$

Hence, $\mathrm{BR}_i(\pi_{-i}) = \mathrm{BR}_i(\pi'_{-i})$, and $\pi_{-i} \sim_i \pi'_{-i}$. □

Explanation. Theorem 1 establishes that if two co-player policies induce exactly the same soft best response distribution for agent i, then they must also induce the same hard best response. This condition holds if and only if the Q-values induced by the two policies differ by a constant shift (i.e., they preserve the same *relative* preferences over actions).

Example 1 (Overcooked). Imagine two co-player policies: one always fetches tomatoes, the other fetches tomatoes 95% of the time and onions 5% of the time. From agent i's perspective, both lead to very similar Q-values for deciding when and whether to fetch an onion. If the Q-values differ only by a constant offset (e.g., one just consistently completes tasks slightly faster), then the soft best responses will be identical, and Theorem 1 applies. But if the slight behavioral variation causes different task timing incentives (e.g., when to deliver), then even with similar soft BRs, the hard BRs could diverge—and Theorem 1 would no longer hold.

In short, Theorem 1 highlights an important connection between soft and hard strategic reasoning. But it only applies when Q-values differ by a constant. In practice, this distinction matters in near-tie or noisy environments, where soft preferences may mask strategic divergence.

Definition 4 (Soft Strategic Equivalence)**.** *Two co-policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ are said to be ε-soft strategically equivalent if:*

$$D_{KL}(BR_i^\tau(\pi_{-i}) \| BR_i^\tau(\pi'_{-i})) \leq \varepsilon$$

Explanation. This definition introduces a relaxed notion of strategic equivalence based on the similarity between soft best response distributions. It formalizes the idea that two co-player policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ can be considered behaviorally similar from agent i's perspective if they induce nearly identical soft best response distributions.

- $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$: These are two co-player policies—i.e., strategies used by all agents other than agent i. Each co-policy defines how agent i's partners behave in the environment.
- $\mathrm{BR}_i^\tau(\pi_{-i}) \in \Delta(A_i)$: This is agent i's soft best response distribution to π_{-i}, defined over the action space A_i. It represents a stochastic policy computed via a softmax over expected Q-values (see Definition 2), and lies in the probability simplex $\Delta(A_i)$.

- $D_{\mathrm{KL}}(\cdot\|\cdot) \in [0, \infty)$: This is the Kullback–Leibler divergence between two probability distributions. It quantifies how different the induced soft best responses are. A divergence of 0 means the distributions are identical; larger values imply greater behavioral dissimilarity.

- $\varepsilon \in \mathbb{R}_{\geq 0}$: This is a user-defined tolerance threshold that determines how close the soft best response distributions must be for the two co-policies to be considered equivalent. Smaller values correspond to stricter equivalence.

  Remark 4 (Choosing the Soft Equivalence Tolerance ε)**.** *The choice of ε in the definition of ε-soft strategic equivalence determines the level of behavioral tolerance allowed when grouping co-player policies into equivalence classes. It plays a crucial role in balancing robustness to noise against fidelity to strategic distinctions. The appropriate value of ε may depend on several factors:*

  - ***Softmax Temperature** τ:* *Higher values of τ produce smoother soft best response distributions that are less sensitive to Q-value variation. In this case, ε should be set smaller to avoid over-grouping dissimilar policies. Conversely, lower τ makes the distributions sharper and more sensitive, allowing for a larger ε.*
  - ***Agent Rationality and Policy Noise:*** *In environments where agents are boundedly rational or where behavioral noise is common (e.g., due to exploration or human variability), a larger ε may be appropriate to avoid overreacting to minor deviations that do not reflect strategic intent.*
  - ***Empirical Distribution of KL Divergences:*** *If soft best response distributions are computed across a dataset of co-policies, the empirical distribution of pairwise KL divergences can be used to calibrate ε. For example, one may select ε as a fixed quantile or based on a clustering threshold.*
  - ***Application-Specific Precision Requirements:*** *In safety-critical or high-stakes settings where precise distinctions in agent behavior matter, a smaller ε is preferred. In contrast, for tasks that emphasize generalization or coordination robustness, a more permissive ε may be more effective.*

  In practice, ε can be treated as a tunable hyperparameter and selected via cross-validation or sensitivity analysis. Its value controls the granularity of the induced clustering over co-player policies and therefore the robustness and expressiveness of the learned strategic representations.

Intuitively, this definition says that two co-player policies are ε-soft strategically equivalent if they induce nearly indistinguishable soft best responses from the ego agent's perspective. That is, from agent i's perspective, the two co-policies are effectively interchangeable for purposes of decision-making, up to a small tolerance. This soft equivalence is particularly useful in noisy or ambiguous environments where small deviations in behavior (e.g., due to bounded rationality or exploration) should not lead to strict reclassification of strategic intent.

Implications for Learning under Uncertainty and Connections to Representation Learning. Definitions 2 and 3 introduced the soft best response distribution and strategic divergence as smooth, probabilistic tools for quantifying how agent i's behavior changes in response to variation in co-player policy π_{-i}. Theorem 1 then established that if two co-player policies induce identical soft best response distributions, they also yield the same hard best response—and hence belong to the same strategic equivalence class. Crucially, this holds when the Q-values differ only by an additive constant, which softmax effectively ignores.

This softmax invariance has several implications for learning in practice:

- **Numerical stability:** shifting Q-values by a constant is harmless and commonly done in deep reinforcement learning pipelines for normalization or stability.

- **Bounded rationality:** the softmax formulation captures decision stochasticity due to uncertainty or cognitive limitations, aligning with models of bounded rationality in both behavioral game theory and practical MARL.

- **Compression for representation learning:** since soft best response distributions depend only on the *relative* ordering of Q-values, they implicitly define an abstraction over the co-policy space. Policies that induce behaviorally indistinguishable soft BRs can be grouped into equivalence classes that preserve only the decision-relevant information.

This perspective motivates a representation learning objective grounded in strategic reasoning: if we wish to learn a compact embedding of co-player behavior that preserves distinctions relevant for ego coordination, then we should aim to separate co-policies that induce different soft best responses. This leads us naturally to define a *contrastive learning objective*—specifically, a strategic InfoNCE loss—that aligns co-player representations with their induced soft BR distributions. This connection is formalized in the next definition.

5 Contrastive Strategic Repesentation Learning

Connection. The preceding definitions and theorem establish a soft, information-theoretic foundation for reasoning about strategic equivalence. Definition 2 introduces the soft best response distribution BR_i^τ, which captures how agent i's action preferences vary with respect to the co-player policy π_{-i}. Definition 3 introduces the strategic divergence SD_i^τ as a fine-grained, continuous measure of behavioral dissimilarity, and Theorem 1 formalizes when identical soft best responses imply hard strategic equivalence.

Intro. Our next goal is to *learn* a compressed representation $h(\pi_{-i})$ of the co-player policy space that preserves these behavioral distinctions. In particular, we want the embedding space to reflect strategic distinctions relevant to agent i's decision-making. To achieve this, we adopt a contrastive learning framework analogous to Contrastive Predictive Coding [$CPCEq.1$] Oord et al., 2019.

At a high level, CPC seeks to learn embeddings by predicting future representations z_{t+k} from a context c_t using a density ratio:

$$f(x, c) \propto \frac{p(x|c)}{p(x)} \qquad \text{[CPC Eq. 2]}$$

Rather than predicting the future, our setting involves inferring best responses. The underlying principle is the same: we aim to maximize mutual information between an observed strategic response (e.g., a sampled action from $\mathrm{BR}_i^\tau(\pi_{-i})$) and its source policy π_{-i}, while minimizing similarity to irrelevant (negative) samples.

Intuition. In contrastive predictive coding, the goal is to learn embeddings that preserve information useful for predicting future outcomes. In contrast to the original InfoNCE that asks "what future does this context lead to?", we're asking "**what kind of partner policy does this action suggest?**" In our setting, rather than predicting future observations, we aim to infer the underlying co-player policy that best explains the observed actions. Specifically, we wish to learn a representation $h(\pi_{-i})$ that maximizes the mutual information between a sampled strategic response (e.g., an action $a^+ \sim \mathrm{BR}_i^\tau(\pi_{-i})$) and the co-policy that induced it, while minimizing similarity to actions drawn from irrelevant or strategically dissimilar co-policies.

Example (Overcooked). This intuition becomes particularly salient in multi-agent environments like Overcooked or Cleanup. In Overcooked, for example, consider a teammate who always navigates clockwise in a narrow circuit kitchen. Observing this pattern, a best-responding agent should infer that the co-player belongs to a class of policies that commit to clockwise motion all the time, and thus adapt by also moving clockwise to avoid interference. A contrastive loss would help cluster such co-player policies together, because they elicit similar best responses.

Example (Cleanup). On the other hand, environments like Cleanup expose more subtle mismatches. Suppose an agent initially signals cooperation by standing near the river, implicitly suggesting a willingness to help with the scarce cleaning task. However, when a valuable apple spawns, the same agent quickly takes it instead of cleaning. This action contradicts what their earlier behavior implied. A contrastive objective helps disambiguate such cases: it encourages the embedding $h(\pi_{-i})$ to be updated based on the actual

observed response (taking the apple), pulling the representation toward a more self-serving class of policies if this pattern persists.

However, it is also important to recognize that boundedly rational agents can make occasional mistakes—or simply act opportunistically without violating their underlying cooperative intent. For example, it may have been locally convenient for the agent to collect the apple because it spawned very close by. Strategic CPC naturally accommodates this type of ambiguity: by relying on soft best response distributions rather than hard classifications, it learns to update co-player representations with a degree of tolerance. This makes it more robust to noisy observations and one-off deviations, enabling a smoother and more reliable clustering of strategic behavior over time.

Connection. These examples demonstrate why contrastive learning is well-suited to the problem of strategic abstraction. It allows agents to cluster co-player policies not based on surface-level signals or prior expectations, but based on the actual strategic incentives they induce—as revealed through best response behavior.

Definition 5 (Strategic InfoNCE loss)**.** *To do this, we define a Strategic InfoNCE loss that mirrors the structure of CPC's loss (Equation 4 in the CPC paper), replacing future samples with soft BR samples:*

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

where f is a similarity function (e.g., dot product or cosine similarity), $a^+ \sim BR_i^\tau(\pi_{-i})$ is a positive sample from the true soft best response, and a_j^- are negatives sampled from the responses to other co-player policies. The loss encourages embeddings $h(\pi_{-i})$ to cluster policies that induce similar soft best responses, while pushing apart those with divergent incentives.

Explanation. The function $f(a, h(\pi_{-i}))$ scores how well the action a aligns with the embedding of the policy π_{-i}. If the positive action a^+ is highly compatible with the embedding $h(\pi_{-i})$, and the negatives are less so, the loss is minimized. As a result, the learned representation encourages co-player policies that induce similar best response distributions to lie closer in embedding space, while separating those that induce distinct responses. This process creates a representation space that is strategically meaningful: policies that lead to the same coordination strategies for agent i are grouped together, enabling efficient generalization, adaptation, and recognition of partner behavior during interaction.

In effect, this contrastive objective learns a strategic representation space where co-policies that induce similar distributions over agent i's actions are embedded nearby. Thus, the learned embedding compresses the co-policy space into clusters that preserve best-response equivalence—a key prerequisite for coordination, imitation, and adaptation in multi-agent learning settings.

Lemma 2 (InfoNCE Minimization Implies Embedding Consistency)**.** *Let $h : \Pi_{-i} \to \mathbb{R}^d$ be a learned embedding function trained to minimize the Strategic InfoNCE loss. Then, under sufficient optimization and representative sampling, minimizing $\mathcal{L}_{InfoNCE}$ implies that co-policies inducing similar soft best responses are embedded close together:*

$$\pi_{-i} \approx \pi'_{-i} \quad \Rightarrow \quad \|h(\pi_{-i}) - h(\pi'_{-i})\| \text{ is small,}$$

where $\approx$ denotes behavioral similarity under soft best response distributions, e.g.,

$$D_{KL}(BR_i^\tau(\pi_{-i}) \| BR_i^\tau(\pi'_{-i})) \leq \varepsilon,$$

and the norm may be taken as Euclidean or cosine distance depending on the similarity function f used in the contrastive loss.

Remark 5. *This lemma formalizes the intuition that minimizing the Strategic InfoNCE loss encourages embeddings to reflect the behavioral structure of the co-policy space. That is, policies which induce similar strategic incentives for agent i (as captured by their soft best response distributions) will be mapped to similar points in the embedding space $\mathbb{R}^d$. This underpins the view of h as learning a soft clustering over strategic equivalence classes.*

Connection. The Strategic InfoNCE loss introduced above is more than just a heuristic objective for learning embeddings, it is a principled estimator of strategic structure. Theorem 2 formalizes this connection by showing that minimizing $\mathcal{L}_{\text{InfoNCE}}$ maximizes a lower bound on the *strategic mutual information* between the co-player policy π_{-i} and the induced action $a_i \sim \text{BR}_i^\tau(\pi_{-i})$. This mutual information quantifies how much information about the co-player's policy is preserved in the distribution over ego-agent responses. A high value indicates that the ego-agent's action distribution is highly dependent on the co-player's strategic type.

By minimizing the contrastive loss, we encourage the learned embedding $h(\pi_{-i})$ to retain the distinctions that are most informative for predicting best responses. In this way, the representation function approximates a soft clustering over the co-policy space, where policies that induce indistinguishable soft BRs (and thus low KL divergence) are embedded close together. This naturally aligns with the notion of soft strategic equivalence classes, and provides a scalable, differentiable mechanism for inferring them from interaction data.

Theorem 2 (Mutual Information Lower Bound)**.** *Minimizing $\mathcal{L}_{InfoNCE}$ maximizes a lower bound on the strategic mutual information:*

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

Thus, contrastive representation learning approximates a soft clustering over SECs.

Explanation. This theorem provides a formal justification for using the Strategic InfoNCE loss as a learning objective. It states that minimizing $\mathcal{L}_{\text{InfoNCE}}$ is equivalent to maximizing a lower bound on the mutual information $I(a_i; \pi_{-i})$ between the co-player policy and the ego agent's soft best response action. This quantity, which we refer to as *strategic mutual information*, measures how much agent i's response reveals about the identity or behavioral type of the co-player policy.

Let us break down each component: the expectation $\mathbb{E}_{\pi_{-i}}$ is taken over a distribution of co-player policies, and $\text{BR}_i^\tau(\pi_{-i})$ denotes the soft best response distribution induced by each policy (Definition 2). The KL divergence compares this distribution to the marginal action distribution $P(a_i)$, which captures the average response across all policies. When $\text{BR}_i^\tau(\pi_{-i})$ differs significantly from $P(a_i)$, the co-player policy induces a unique strategic response, contributing to a higher mutual information score.

Therefore, minimizing $\mathcal{L}_{\text{InfoNCE}}$ encourages embeddings $h(\pi_{-i})$ that preserve distinctions in how different co-player policies affect agent i's action preferences. In effect, this process induces a soft clustering over the co-policy space, where policies that induce similar strategic incentives are embedded nearby. As such, this theorem provides both a theoretical foundation and an operational criterion for learning behaviorally meaningful representations that approximate soft strategic equivalence classes.

6 Intent Certainty (IC) as an Strategic Ambiguity Filter

Before learning long-term strategic abstractions, we first identify which co-player policies emit reliable and unambiguous behavioral signals. We formalize this using *Intentional Information Equivalence (IC)*, which measures the certainty or intentionality of a policy based on its influence on the ego agent's action distribution.

Definition 6 (Intent Certainty via Mutual Information.)**.** *Let $\pi_{-i} \in \Pi_{-i}$ denote a co-player policy and $a_i \in A_i$ the ego agent's action. We define the intentional information score as the mutual information between the co-policy and the ego action:*

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

This measures how much observing the co-player policy reduces uncertainty over the ego agent's best response. Policies that induce highly certain or sharply peaked best responses yield higher IC scores.

Definition 7 (Contrastive Estimation of IC.)**.** *We estimate $\text{IC}(\pi_{-i})$ via contrastive learning. Let $h_{IC}(\pi_{-i})$ be an embedding of the co-policy, and define a loss function $f(a_i, h_{IC}(\pi_{-i}))$. The contrastive loss for intentional*

influence is:

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

where $a_i^+ \sim BR_i^\tau(\pi_{-i})$ is a soft best response action and $a_i^- \sim BR_i^\tau(\pi_{-j})$ for $j \neq i$ are sampled negatives.

Definition 8 (Intent Filtering.)**.** *We retain only co-player policies whose mutual information score exceeds a predefined threshold:*

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

This forms the set of high-certainty co-policies that are passed to the next stage of representation learning.

By filtering out policies with ambiguous or noisy influence, this stage ensures that downstream abstraction (e.g., via SFR-SER) is grounded in interpretable and robust behavioral signals.

Remark 6 (Interpreting Intent Certainty)**.** *It is important to note that IC does not measure whether a co-player is cooperative or adversarial—it measures whether they are* ***strategically*** **clear***. For example, a policy that always cooperates and one that always defects may lead to opposite behaviors, but both produce* ***a highly predictable influence*** *on the ego agent's decision-making. Under IC, they are considered intentionally equivalent because each yields a sharply peaked distribution over the ego agent's best response.*

*At this level of abstraction, we are not asking "****what is this agent trying to achieve?****" but rather "****how certain am I about what I should do in response?****" IC serves as an intent filter that distinguishes between agents with high strategic influence (i.e., they reliably shape ego behavior) and those that are ambiguous, inconsistent, or noisy. It provides a principled, learnable signal of who the agent is in terms of their capacity to influence:* ***not what their intentions are, but how clearly those intentions are expressed.***

7 Long-term Strategic Abstraction via Successor Representation

Motivation. While soft strategic equivalence provides a powerful framework for clustering co-player policies based on short-term incentives—i.e., how they influence the ego agent's immediate best response—it does not fully capture alignment in long-term intent. Two policies may appear equivalent under soft best response (e.g., inducing the same action distribution), yet lead to very different trajectories and long-term consequences. This distinction becomes especially important in environments with delayed rewards, sparse feedback, or temporally extended goals. For example, in the Cleanup environment, a co-player who hovers near the river may initially appear to be cooperative—inducing the ego agent to fetch apples while assuming the partner will clean. However, if that partner consistently ignores the cleaning task over time and only takes the apple when it spawns nearby, their behavior, though superficially aligned, ultimately undermines the shared goal of maintaining cleanliness. Despite similar short-term incentives, the long-term outcomes induced by such a policy diverge from those of a truly cooperative partner who sacrifices short-term rewards to clean proactively. This highlights the need for a representation that captures the full trajectory of strategic consequences, not just momentary coordination patterns. To reason robustly about such cases, we require a refinement of strategic equivalence that incorporates an agent's ***impact on future outcomes***. We introduce **SFR-SER** (Successor Feature Representation Strategic Equivalence Relation) to address this: by comparing the long-term successor features induced by co-player policies, we gain a representation of their strategic trajectory. This allows us to reason not only about *how* a partner influences our current behavior, but also about ***what long-term outcomes*** their behavior is driving us toward. Intuitively, the successor feature-based strategic equivalence relation (SFR-SER) asks the question: ***"Do these co-player policies result in the same long-term consequences for me, under arbitrary goals?"*** While soft strategic equivalence focuses on short-term incentives by comparing induced best response distributions, SFR-SER incorporates trajectory-level reasoning by evaluating whether different co-policies lead the ego agent to similar long-term state occupancies. This perspective allows us to group together co-policies that may differ in immediate behavior but are aligned in their long-term strategic consequences. In doing so, SFR-SER provides a principled lens for detecting long-term alignment or misalignment especially in multi-agent coordination where commitment to the same goal may take different forms as stated in the Cleanup example.

Long-Horizon Strategic Abstraction via Successor Feature Representations. To reason about the long-term influence of a policy on the environment, we consider the successor feature representation (SFR) framework, which captures the distribution over features encountered under a policy. Specifically, the ξ-*function* encodes the cumulative discounted visitation density over features $\phi \in \Phi$ under policy π:

$$\xi^{\pi}(s, a, \phi) = \sum_{k=0}^{\infty} \gamma^k \, p(\phi_{t+k} = \phi \mid s_t = s, a_t = a; \, \pi).$$

This representation serves as a reward-independent signature of a policy's long-term behavior in the feature space. It characterizes the full trajectory-level consequences of selecting action a in state s and continuing under policy π, regardless of the task-specific reward function. Importantly, the ξ-function captures not just which states are visited, but which *semantic features* those states instantiate, enabling transfer and reasoning about abstract goals [SFR Eq. 11] Reinke and Alameda-Pineda, 2022.

Connection. We now define a strategic equivalence relation based on SFR, which we term *SFR-SER*. This abstraction groups co-player policies according to whether they induce indistinguishable long-term outcomes for the ego agent, measured through their ξ-function

Definition 9 (SFR-based Strategic Similarity (SFR-SS))**.** *Let π_i be the ego agent's policy and $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ be two co-player policies. We say $\pi_{-i} \sim_i^{SFR} \pi'_{-i}$ if, for all states $s \in S$, actions $a \in A_i$, and features $\phi \in \Phi$, the successor feature representations experienced by the ego agent are equivalent:*

$$\xi^{\pi_i, \pi_{-i}}(s, a, \phi) \approx \xi^{\pi_i, \pi'_{-i}}(s, a, \phi).$$

That is, the co-player policies are SFR-strategically equivalent if they induce the same long-run feature distributions from the perspective of agent i executing policy π_i.

Explanation. This definition introduces a trajectory-aware notion of strategic equivalence, grounded in the successor feature representation (SFR) framework. It formalizes when two co-player policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ are indistinguishable to the ego agent i, based on their long-term influence on the environment.

- $\pi_i \in \Pi_i$: The fixed policy of the ego agent.
- $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$: Two co-player policies under comparison. These may differ in their surface behavior but may nonetheless have the same long-term impact on agent i's experience.
- $\xi^{\pi_i, \pi_{-i}} : S \times A_i \times \Phi \to \mathbb{R}_{\geq 0}$: The *successor feature representation* for agent i under joint policy (π_i, π_{-i}). It defines the discounted cumulative density over features $\phi \in \Phi$ encountered when taking action $a \in A_i$ in state $s \in S$, and continuing under (π_i, π_{-i}).
- $\xi^{\pi_i, \pi_{-i}}(s, a, \phi) = \sum_{k=0}^{\infty} \gamma^k \, p(\phi_{t+k} = \phi \mid s_t = s, a_t = a; \pi_i, \pi_{-i})$: This expansion makes clear that ξ encodes the full discounted trajectory over feature space—not merely next-step transitions or one-step predictions.
- Φ: The feature space, typically describing semantically relevant state aspects (e.g., object properties, roles fulfilled, goals achieved). The SFR aggregates future visitation information over this space.
- $\pi_{-i} \sim_i^{\text{SFR}} \pi'_{-i}$: Strategic equivalence under SFR, from the perspective of agent i. This relation holds if the co-player policies induce *indistinguishable ξ-trajectories* for all ego actions and states.

Intuition. This definition abstracts away from the literal actions of co-players and focuses instead on their long-run influence on the ego agent's future feature occupancy. If $\pi_{-i} \sim_i^{\text{SFR}} \pi'_{-i}$, then for all practical purposes, agent i experiences the same future distribution over semantically meaningful features regardless of whether it is paired with π_{-i} or π'_{-i}. Thus, the two policies can be treated as strategically interchangeable.

Example (Overcooked). For example, in the Overcooked environment, consider two co-player policies: one that always moves clockwise to deliver onions to the soup pot, and another that moves counterclockwise but ultimately delivers onions with similar timing and consistency. Although their action-level behavior differs, both policies reliably enable the ego agent to cook and serve dishes at the same rate. The ξ-function over features such as "onion delivered," "pot filled," and "dish served" is nearly identical across both co-policies. From the ego agent's perspective, they are strategically interchangeable.

Notice that while the examples use hand-coded, semantically meaningful features for intuitive explanation, the actual feature representations used in practice may be much more granular or abstract: e.g., "the agent is standing at grid position (2,5) and facing left." These low-level or latent features can still support meaningful strategic abstraction when aggregated over time through the successor feature representation.

Example (Cleanup). Similarly, in the Cleanup environment, consider a co-player that proactively cleans the river before collecting apples, and another that collects one apple before beginning to clean. While their initial actions differ, both policies ensure the river is cleaned early enough to sustain apple spawning. As a result, the ego agent experiences the same trajectory of meaningful features—"river cleaned," "apple collected," and "goal reached"—over time. These policies induce similar ξ-functions and belong to the same SFR-based strategic equivalence class.

These examples highlight how SFR-SER captures the essence of long-term cooperation: it abstracts from transient behavioral variability and focuses instead on the co-player's impact on the agent's strategic future.

Why this matters. This formulation enables abstraction and compression: multiple co-player policies may behave differently at the action level, yet induce the same ξ-function from the ego's perspective. This allows for learned representations $h(\pi_{-i})$ that preserve only the aspects of the co-policy relevant to the ego agent's strategic outcomes. It also supports generalization across behaviorally diverse but strategically aligned co-players.

Definition 10 (SFR-Consistent Representation)**.** *A representation function $h : \Pi_{-i} \to \mathbb{R}^d$ is said to be SFR-consistent with respect to ego policy π_i if for all co-policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$,*

$$\|h(\pi_{-i}) - h(\pi'_{-i})\| \text{ is small} \quad \text{whenever} \quad D_{KL}\left(\xi^{\pi_i,\pi_{-i}}(s,a,\cdot) \,\|\, \xi^{\pi_i,\pi'_{-i}}(s,a,\cdot)\right) \le \varepsilon \quad \forall s,a.$$

Theorem 3 (Representation Consistency Implies Approximate SFR Equivalence)**.** *Let $h : \Pi_{-i} \to \mathbb{R}^d$ be an SFR-consistent representation function with respect to a fixed ego policy π_i. Suppose there exists $\varepsilon > 0$ such that for all co-player policies $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$,*

$$\|h(\pi_{-i}) - h(\pi'_{-i})\| \le \delta \quad \Rightarrow \quad D_{\mathrm{KL}}\left(\xi^{\pi_i,\pi_{-i}}(s,a,\cdot) \,\|\, \xi^{\pi_i,\pi'_{-i}}(s,a,\cdot)\right) \le \varepsilon \quad \forall s \in S, a \in A_i.$$

Then, all policies within a radius-δ neighborhood in the representation space are ε-soft strategically equivalent under SFR:

$$\|h(\pi_{-i}) - h(\pi'_{-i})\| \le \delta \quad \Rightarrow \quad \pi_{-i} \sim_i^{SFR(\varepsilon)} \pi'_{-i}.$$

8 World Dynamics Compression via Multi-agent Successor Representation

Model-Aware Strategic Similarity (MASR-SS). While SFR-SER captures long-term outcome equivalence through feature occupancy, it assumes the environment dynamics remain fixed. However, in many multi-agent settings, the co-player's policy not only shapes incentives but also modifies the transition dynamics experienced by the ego agent. For example, a partner who blocks key pathways in Overcooked, or who delays river cleaning in Cleanup, changes the effective affordances available to the ego agent—altering which actions are possible, efficient, or safe.

To capture these effects, we introduce *Model-Aware Strategic Equivalence (MASR-SER)*: a model-aware strategic equivalence relation based on multi-agent successor representations, a refinement of SFR-SER that clusters co-player policies based on their induced transition models. Two policies are MASR-equivalent if they produce the same transition dynamics for the ego agent, regardless of the long-run feature outcomes.

This allows the agent to reason not only about what co-players are trying to achieve, but how their behavior structurally modifies the environment. MASR-SER supports planning, simulation-based reasoning, and fine-grained generalization across behaviorally similar but dynamically distinct partners.

Definition 11 (Model-Aware Strategic Equivalence (MASR-SER))**.** *Let $\pi_i \in \Pi_i$ be the ego agent's fixed policy, and let $\pi_{-i}, \pi'_{-i} \in \Pi_{-i}$ be two co-player policies. We say that $\pi_{-i} \sim_i^{\text{MASR}} \pi'_{-i}$ (i.e., they are MASR-strategically equivalent from the perspective of agent i) if the induced transition dynamics experienced by the ego agent under both co-policies are equivalent:*

$$p(s' \mid s, a_i; \pi_{-i}) = p(s' \mid s, a_i; \pi'_{-i}) \quad \forall s \in S, a_i \in A_i, s' \in S.$$

That is, the two policies are MASR-strategically equivalent if the world evolves in the same way for agent i, regardless of which policy it is paired with.

Intuition. MASR-SER answers the question: ***"Do these two co-player policies give me the same experiential world?"*** In other words, **"will my environment respond in the same way when I act—regardless of which co-player I'm paired with?"** This goes beyond shared goals or outcome trajectories (as in SFR-SER) to encompass how a co-player modifies the transition dynamics themselves.

Example (Overcooked). Imagine two co-player policies in the Cramped Room layout. One agent moves quickly and predictably through the hallway without blocking. Another moves with slight delays but never obstructs the ego agent's path. Both result in the ego agent having unobstructed access to key stations, inducing nearly identical transition dynamics. Despite timing differences, they are MASR-equivalent because they create the same experiential structure.

By contrast, a third policy that unpredictably pauses in the hallway forces the ego agent to reroute or wait, thus altering the transition probabilities conditioned on ego actions. This breaks MASR equivalence, even if the long-term outcomes (e.g., dish delivered) remain similar.

Example (Cleanup). Consider two co-player policies: one that cleans the river immediately and one that cleans it after picking up a nearby apple. Both result in the river being clean soon enough for apples to spawn regularly. From the ego agent's perspective, the environment responds similarly over time to its apple-fetching actions—they can expect apples to be present and accessible. These policies are MASR-equivalent.

However, a third policy that neglects the river entirely changes the environment's effective model: apple spawn rates decrease, and the ego agent's actions now lead to different downstream transitions. This policy belongs to a different MASR-equivalence class.

9 Strategic Refinement as Learning Dynamics

Motivation. Strategic abstraction is not static — it evolves as the agent interacts, observes, and updates its representations of others. In this section, we formalize strategic abstraction as a learning process: a dynamic refinement of co-player equivalence classes over time. As the agent collects more data and improves its representations through objectives like IC filtering, contrastive learning (CPC), successor feature reasoning (SFR), and model-aware abstraction (MASR), it resolves uncertainty about which distinctions are strategically meaningful.

We quantify this progression using a metric of ***strategic ambiguity reduction***, which tracks how entropy over co-player policy clusters decreases as learning unfolds. The trajectory of this refinement illustrates how coordination capabilities improve—not by modeling co-policies in full detail, but by compressing them into increasingly precise, decision-relevant abstractions. ***This lens frames strategic learning as a principled process of uncertainty elimination, aligned with the agent's capacity to coordinate robustly in complex, uncertain environments.***

Definition 12 (Strategic Ambiguity Reduction with Filtering and Refinement)**.** *Let Π_{-i} denote the space of all co-player policies and let $\mathcal{P}_{intent} \subseteq \Pi_{-i}$ be the subset of co-policies that exceed an intentional influence threshold:*

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

Let E_0 be the initial coarse clustering over $\mathcal{P}_{intent}$, and let E_t be the learned SEC partition at time t based on a representation h_t trained with CPC, SFR, and MASR objectives.

We define the total strategic ambiguity reduction over time as:

$$\Delta_{\text{SEC}}(t) := H[\mathcal{P}_{intent} \mid E_0] - H[\mathcal{P}_{intent} \mid E_t],$$

where $H[\cdot \mid E_t]$ denotes the conditional entropy of the co-policy distribution under the current clustering.

The refinement trajectory then becomes a composition of stages:

$$\Pi_{-i} \xrightarrow{IC} \mathcal{P}_{intent} \xrightarrow{CPC+SFR} E_t \xrightarrow{MASR} E_{t+\Delta},$$

reflecting progressive resolution of ambiguity from intent clarity to behavioral impact to structural dynamics.

Explanation and Intuition. This definition formalizes strategic learning as a process of progressive uncertainty reduction. The agent begins with a large, undifferentiated space of co-player policies Π_{-i}, which may include both ambiguous and uninformative behaviors. The first filtering stage uses Intentional Information Equivalence (IC) to retain only policies that exert clear, unambiguous influence on the ego agent—those with high mutual information between co-policy and ego response. This filtered subset, $\mathcal{P}_{\text{intent}}$, defines the starting point for abstraction.

The learning dynamics then refine the agent's beliefs over time by applying contrastive and representation learning objectives. At each stage, the ego agent improves its ability to distinguish co-policies that matter strategically: first based on short-term influence (via CPC), then based on long-run feature trajectories (via SFR), and finally on induced transition structure (via MASR). Each refinement step narrows the agent's clustering of co-policies—reducing entropy over the space of strategic types and sharpening the agent's ability to respond appropriately.

Example. Consider an Overcooked environment in the Coordination Ring layout. Initially, the agent may face several co-player policies—some chaotic, some systematic. After IC filtering, it keeps only those whose movement patterns reliably constrain the ego agent's choices (e.g., always clockwise vs. always counter-clockwise). Early CPC learning might cluster these based on how they influence the ego's next action—e.g., "wait" vs. "grab pot." Later, SFR reveals that some of these partners consistently lead to burned dishes due to delayed delivery, even if they seem aligned in the short term. MASR disambiguates partners who induce the same outcomes but do so via different paths—e.g., by blocking choke points or delaying the ego agent's movement options. Through each stage, the ego agent's understanding becomes more fine-grained and actionable.

Connection to Trust. This refinement process operationalizes a notion of trust grounded in consistency across abstraction layers. When a co-player consistently reduces strategic ambiguity—i.e., when its influence is clear, behaviorally predictable, aligned in outcomes, and structurally stable—the agent has reason to trust it. That trust is not binary, but emerges as the cumulative result of abstraction layers aligning. Conversely, if a policy passes IC but diverges later in SFR or MASR, the agent should remain cautious, recognize uncertainty, or request further evidence. In this way, $\Delta_{\text{SEC}}(t)$ provides a learnable, measurable proxy for the ***refinement of trust*** over time, driven by data and grounded in strategic consequence.

Remark 7 (Strategic Clarity Redefines Cooperation and Coordination)**.** *Traditional notions of cooperation often rely on outcome-based judgments—e.g., did the agent help achieve a common goal? In contrast, our framework redefines coordination in terms of* strategic clarity*: whether a co-player's behavior reliably informs the ego agent about what it should do. From this perspective, cooperation is not about being helpful—it is about giving legible signal and strategically stable.*

This leads to a surprising but useful insight: agents that always defect *can be highly valuable for coordination. Their behavior, though uncooperative in outcome terms, is maximally predictable. The ego agent can adapt quickly and confidently when paired with such agents. In contrast, co-players who sometimes cooperate and sometimes defect—without a stable signal or contextual rationale—introduce ambiguity, miscoordination, and learning instability.*

Example (Overcooked). *Consider two co-players: one always blocks the pot and never helps serve dishes, and another appears cooperative at first but unpredictably switches between assisting and sabotaging. The always-defecting agent, though adversarial, is easier to adapt to: the ego agent can route around it or lower its expectations. The unpredictable agent introduces behavioral noise, making trust and coordination harder. Under our framework, the first agent may score high on IC and MASR (consistent influence and model structure), while the second scores low—despite being occasionally helpful.*

This redefinition allows us to separate alignment in outcomes *from* alignment in influence*. Agents that behave with intent—even if their goals diverge from ours—can still support stable, interpretable coordination. This is especially important in multi-agent safety, where strategic clarity can be more valuable than goodwill.*

Self-Alignment through Strategic Consistency. While this framework defines abstraction layers for modeling others, it also supports introspective reasoning. Each level—intent clarity (IC), short-term incentives (CPC), long-term outcomes (SFR), and world modeling (MASR)—can be applied not just to co-player policies, but to the agent's own behavior. This allows an agent to reflect: *Are my actions consistent with my inferred intentions? Do I reliably influence the world in ways that match my stated goals or internal beliefs?* When these layers align internally, the agent can be said to be *strategically self-aligned*. When they diverge, this inconsistency becomes a signal—either of misgeneralization, emergent conflict, or miscommunication. Our framework thus provides a computational lens on self-awareness, enabling agents to verify whether they themselves are trustworthy according to the very standards they apply to others.

10 Intrinsic Motivation through Strategic Refinement

Traditional approaches to intrinsic motivation focus on driving exploration through novelty, prediction error, or information gain—guiding agents toward unfamiliar states in the hope of discovering value. While effective in single-agent environments, these approaches can fall short in multi-agent settings where what matters is not just state novelty, but the strategic influence of others. In such cases, the key question is not "What is new?" but "What is informative about how others affect my behavior?"

Our framework reframes intrinsic motivation as a process of *strategic refinement*: the structured reduction of uncertainty about co-player influence over time. Rather than seeking novelty for its own sake, agents are motivated to minimize ambiguity across a hierarchy of abstraction layers—Intentional Information Equivalence (IIE), soft best responses (CPC), long-term successor features (SFR), and induced world dynamics (MASR). This process is quantified by the strategic ambiguity reduction metric $\Delta_{\text{SEC}}(t)$, which tracks entropy decay over equivalence classes of co-policies. The larger the reduction, the more the agent has learned about the strategic roles of its partners.

This refinement process serves as a powerful form of intrinsic motivation:

- **At the IIE layer**, agents seek partners who provide maximally legible behavioral influence—those who consistently steer the ego agent's action distribution. This aligns with mutual information-based curiosity, but grounded in influence rather than state novelty.
- **At the CPC and SFR layers**, agents are intrinsically motivated to cluster policies by how they alter short-term incentives and long-run outcomes. Strategic divergence between clusters creates contrastive learning signals that reward epistemic progress.
- **At the MASR layer**, refinement corresponds to understanding how others alter the ego agent's transition dynamics. This introduces a model-based analogue of empowerment: the more predictable and stable your environment is under a partner's influence, the more strategic control you have over your own planning horizon.

Importantly, this form of intrinsic motivation is *interaction-centered* rather than self-centered. It does not seek control in isolation, but understanding through relation—driving agents to explore and refine their models of others not just to improve prediction, but to enable robust, zero-shot coordination.

Example. Consider an agent in Cleanup faced with several co-player policies. A partner who always takes apples is uncooperative, but clear. Another partner who hesitates near the river sends ambiguous signals—does this mean they will clean, or defect later? Strategic refinement drives the ego agent to engage further with the ambiguous policy, seeking more interaction data to resolve its SEC assignment. This curiosity is not about novelty, but about the need to converge on a minimal, decision-relevant abstraction of the partner's role.

In this way, strategic abstraction naturally induces a form of intrinsic motivation aligned with coordination: the drive to reduce ambiguity over how others influence one's own policy. This motivates exploration of behavior space in a way that supports safety, trust, and long-term generalization.

11 Open-Endedness through Strategic Abstraction

Open-ended systems are defined by their ability to produce artifacts that are both novel and learnable to an observer Hughes et al., 2024. In our case, these artifacts are strategic representations—compressed abstractions of co-player behavior that evolve over time. From the observer's perspective (the ego agent), a co-player policy is open-ended if each new observation contributes to the refinement of the strategic abstraction space: either by revealing new best-response distinctions, clarifying long-run outcomes, or shifting the model of environmental dynamics.

Our framework satisfies both dimensions of open-endedness:

- **Novelty.** As the ego agent interacts with new co-policies, it encounters behaviors that are initially unpredictable under its current clustering. These surprises drive partition refinement—e.g., splitting previously indistinguishable classes when CPC, SFR, or MASR layers diverge.
- **Learnability.** Over time, longer interaction histories enable more accurate representation learning. As $t \to \infty$, the learned abstraction h_t compresses the behavior space more effectively, reducing uncertainty and enabling faster alignment and generalization.

The abstraction refinement metric $\Delta_{\text{SEC}}(t)$ (defined in Section **??**) naturally corresponds to a measure of information gain. Each refinement step produces an artifact (e.g., a finer-grained SEC partition) that is both novel—because it surprises the current model—and learnable—because it results in a stable, predictive refinement.

State Marginal Matching As part of the next phase, we plan to incorporate the State Marginal Matching (SMM) Lee et al., 2019 objective into the strategic abstraction framework. SMM provides a mathematically grounded and behaviorally meaningful way to drive exploration, not merely through novelty, but via principled distribution matching over strategically relevant state marginals. This aligns naturally with the goal of abstraction learning: rather than training agents to explore arbitrarily, SMM can help them match the distribution of interactional roles or strategic niches, forming a tractable outer loop to scaffold strategic refinement. Our goal is to adapt SMM as a meta-exploration signal that encourages the discovery of co-player abstractions that are maximally informative yet compressible—potentially guiding the system toward convergent abstraction layers (IC, CPC, SFR, MASR) in a more self-supervised and open-ended fashion.

Strategic Open-Endedness. This view positions strategic abstraction not just as a tool for coordination, but as a *driver of open-ended interaction.* The ego agent learns to distinguish more nuanced behavioral types, respond more precisely, and align more reliably—not by scaling data or parameters, but by progressively internalizing the structure of others' influence.

This reframing also connects to safety: if the agent observes that multiple abstraction layers (e.g., IIE and MASR) disagree, it halts or defers. If they align and converge, the agent gains epistemic confidence. In this way, abstraction agreement becomes an actionable indicator of open-ended but safe generalization—moving toward ASI not by expanding capability space, but by deepening relational understanding.

References

Hughes, E., Dennis, M., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., & Rocktaschel, T. (2024, June). Open-Endedness is Essential for Artificial Superhuman Intelligence [arXiv:2406.04268 [cs]]. https://doi.org/10.48550/arXiv.2406.04268

Lauffer, N., Shah, A., Carroll, M., Dennis, M., & Russell, S. (2023, July). Who Needs to Know? Minimal Knowledge for Optimal Coordination [arXiv:2306.09309 [cs]]. https://doi.org/10.48550/arXiv.2306.09309
Comment: To be published at ICML 2023.

Lee, L., Eysenbach, B., Parisotto, E., Salakhutdinov, R., & Levine, S. (2019). STATE MARGINAL MATCHING WITH MIXTURES OF POLICIES.

Oord, A. v. d., Li, Y., & Vinyals, O. (2019, January). Representation Learning with Contrastive Predictive Coding [arXiv:1807.03748 [cs]]. https://doi.org/10.48550/arXiv.1807.03748

Reinke, C., & Alameda-Pineda, X. (2022). Successor Feature Representations. *Transactions on Machine Learning Research*. Retrieved March 9, 2025, from https://openreview.net/forum?id=MTFf1rDDEI