Well-Founded Arbitration Under Internal Epistemic Pluralities

1 Problem Statement: A Computational Theory of Justified Action Under Internal Pluralities

Agents do not fail because they lack predictions, they fail when they act on predictions whose relevance they cannot justify. We propose ***a computational theory of justified action under internal pluralities***. When should an intelligent agent act on diverse, conflicting internal signals versus abstaining to preserve ***epistemic integrity***—the state where internal components have achieved sufficient coordination about what distinctions matter for principled action given their current internal diverging perspectives? This question arises whenever agents must coordinate without shared objectives [Arrow, 1951; Ngo, 2025a], from neural network components processing contradictory gradients [Elhage et al., 2021] to human-AI teams navigating value misalignment [Russell, 2019].

The Universal Challenge. Every intelligent system faces this fundamental dilemma. A large language model must reconcile conflicting training signals when generating responses [Bai et al., 2022]. A human weighs competing intuitions before making decisions. A multi-agent system coordinates despite divergent objectives [List and Pettit, 2011]. In each case, the core question remains: ***when is action epistemically justified given internal conflict?*** This challenge appears whether we view the agent as a computational system managing internal diverging perspectives or as a coalition of subagents that must achieve normative coordination before committing to action.

Beyond Classical Decision Theory. Traditional approaches assume agents either possess well-defined utility functions [Savage, 1954] or can learn them through environmental feedback [Sutton and Barto, 2018]. But intelligence systems often operates in the absence of clear optimization targets [Russell, 2019], relying instead on diverse, potentially misaligned internal signals. We formalize when action based on such signals is ***epistemically coherent***—what we term **well-foundedness** [Ngo, 2025b]: when internal epistemic structure warrants action, rather than merely optimal. This requires not just computational tractability, but also principled coordination among internal components that may hold conflicting views about what distinctions matter for decision-making. Current approaches exhibit systematic failures when internal components disagree: neural networks display inconsistent behavior across semantically equivalent inputs, human-AI teams reach unstable compromises that satisfy neither party's values, and multi-agent systems oscillate between conflicting strategies. These failures stem from acting before achieving sufficient internal epistemic coordination, highlighting the need for principled mechanisms that preserve future options for coherent action.

From External to Internal Strategic Equivalence. Our approach builds on Strategic Equivalence Relations (SER) [Lauffer et al., 2023], which provide a principled method for discarding distinctions among strategies in game-theoretic settings. In classical SER, agents ignore differences between strategies unless they lead to different best responses against all possible opponent strategies. We generalize this framework from external multi-agent coordination to internal epistemic arbitration: rather than determining which strategic distinctions matter against external opponents, we establish which expert distinctions matter for internal coordination. This shift from inter-agent to intra-agent equivalence relations enables principled abstraction over conflicting internal signals while preserving decision-relevant information.

The One-Step Pluralist Bandit (1SPB) Setting. We study this challenge in its most essential form: a stateless decision problem where an agent must arbitrate among K internal experts, each producing scalar predictions $Q_k(a) \in [0, 1]$ for every action $a \in \mathcal{A}$. These predictions represent fixed epistemic inputs: priors, heuristics, or biases that may diverge significantly. The agent applies an abstraction function φ to compress expert configurations into equivalence classes, paralleling accessibility relations in Kripke models of knowledge [Fagin et al., 1995; Chellas, 1980]. There is no feedback, no reward signal, no iterative learning—only the pure problem of epistemic arbitration among potentially disagreeing components.

Dual Theoretical Perspectives. We develop this framework through two complementary lenses. From a *computational* perspective, we establish conditions for well-foundedness Ngo [2025b]—when an agent's internal structure has achieved sufficient epistemic coherence to warrant action. From a *coordination-theoretic* perspective, we analyze this as a problem of norm emergence among internal coalition members who must achieve common knowledge about decision-relevance before proceeding. These perspectives yield identical mathematical procedures while offering different conceptual foundations for understanding epistemically justified action under plurality. **Operationalizing Epistemic Responsibility.** Our central innovation lies in making abstract philosophical concepts computationally tractable [Goldman, 1999]. We introduce two complementary epistemic conditions that must be satisfied before action is permitted:

- **Epistemic Sufficiency**: Verifying that internal distinctions are irrelevant to the choice boundary
- **Epistemic Stability**: Ensuring that uncertainty about expert weighting does not deform behavior

These conditions are both necessary and jointly sufficient for epistemic justification. Sufficiency alone is insufficient, an agent might preserve decision-relevant distinctions yet remain paralyzed by disagreement about how to weight them. Stability alone is also insufficient, consistent weighting of experts provides no guarantee that meaningful distinctions haven't been discarded. Only their conjunction ensures that both the abstraction preserves what matters and the internal coalition agrees on how to use it. This abstention-capable approach addresses a critical failure mode: premature action under internal disagreement often leads to brittle, unstable behavior that performs worse than withholding action until internal coordination is achieved. Consider an AI system with conflicting safety and capability considerations—acting before

these are reconciled risks systematic misalignment, while abstention preserves future options for principled coordination. Action proceeds only when both conditions are satisfied simultaneously. This joint requirement transforms the philosophical question of ***epistemic justification***, whether internal coordination has reached the threshold where action becomes warranted, into concrete computational checkpoints[Chow, 1970; Geifman and El-Yaniv, 2017].

Rather than optimizing for expected outcomes, the agent first verifies that its internal components have achieved sufficient normative agreement about what distinctions matter for action. When either epistemic condition fails, the agent abstains—not due to uncertainty about the world, but due to insufficient internal coordination about decision-relevance. This preserves both computational tractability and normative legitimacy: action proceeds only when internal coalition members can trust that their shared abstraction preserves all strategically relevant distinctions. **From Simulation to Arbitration.** This reframes agency itself [Ngo, 2025a, 2023], providing **a scale-invariant grammar for coordination**. Rather than viewing intelligence as world-modeling followed by optimization, we propose it as ***epistemic arbitration under constraint***. The question is not ***"what should I do given complete world knowledge?"*** but ***"when is my internal diverging perspectives sufficiently resolved to warrant any action that represents my plural beliefs in a stable, principled manner?"*** This applies whether internal coordination emerges through computational mechanisms (hierarchical agency Kulveit [2024]) or through norm coordination processes (common knowledge achievement). Where memory-centric theories classify what an agent can represent Kirtland et al. [2025], we define when representation is sufficient to act. The result is a planner-free, **abstention-capable substrate for principled action under plurality**—whether in neural modules learning to coordinate their outputs, human-AI teams establishing shared trust boundaries, or civilizational coalitions navigating value plurality [List and Pettit, 2011]. Justified coordination, under this lens, begins when ***epistemic agreement*** has converged [Aumann, 1976; Moses and Tennenholtz, 1995]—a form of norm emergence focused not on behavioral coordination (how to act) but on epistemic justification (when action itself becomes warranted).

2 Epistemic Sufficiency

The principle of ***epistemic sufficiency*** defines the minimal condition under which internal plurality may be collapsed into action. It governs when an abstraction over expert predictions preserves all distinctions that matter to the agent's behavior. Condition (1) operationalises the modal axiom K for the internal coalition: indistinguishable expert profiles must induce identical soft best-response policies, just as indistinguishable worlds satisfy the same propositions in Kripke semantics [Fagin et al., 1995, Ch. 2].

However, epistemic sufficiency governs *when* an abstraction preserves decision-relevant distinctions, but does not specify ***how*** to construct such abstractions in the first place. We first address this foundational question by defining principled abstraction construction, then establish the sufficiency criterion that evaluates these abstractions.

2.1 Principled Abstraction Construction

This section formalizes the canonical abstraction that preserves exactly the decision-relevant distinctions and derives tractable approximations via constrained optimization.

2.1.1 Canonical Decision Equivalence

The fundamental question in epistemic arbitration is: when should two different expert configurations be treated as equivalent? Intuitively, configurations that lead to identical decision outcomes contain no strategically relevant distinctions and may be safely merged. This insight leads us to define equivalence in terms of policy identity rather than prediction similarity. We begin with the observation that two expert configurations should be considered equivalent if and only if they induce identical decision policies at the current temperature setting.

Motivation. Consider two expert configurations that, despite numerical differences in their predictions, yield identical soft best-response policies after temperature-weighted aggregation. From the agent's decision-theoretic perspective, these configurations are functionally indistinguishable, meaning that any choice between them is arbitrary. Conversely, configurations that induce different policies represent genuine epistemic distinctions that matter for action selection. This motivates a canonical notion of equivalence based purely on decision boundaries. The soft best-response policy follows the Boltzmann distribution [Jaynes, 2003], where temperature τ controls the agent's sensitivity to internal disagreement. This connection formalizes the intuition that epistemic arbitration should preserve exactly the distinctions that matter for the agent's knowledge-to-action translation, grounded in modal logic foundations [Chellas, 1980].

Definition 1 (Canonical epistemic equivalence)**.** *For expert configurations $Q, Q' \in [0,1]^{K\times|A|}$ and fixed temperature τ, define*

$$Q \sim_{\pi^*,\tau} Q' \quad \Longleftrightarrow \quad \pi^*_{raw}(\cdot|Q) = \pi^*_{raw}(\cdot|Q') \tag{1}$$

where π^*_{raw} is the unabstracted policy from Equation (1).

This relation is an equivalence relation since policy equality is reflexive, symmetric, and transitive. Equivalence classes are defined point-wise for the temperature τ in force for that decision episode.

Temperature dependence. The equivalence relation $\sim_{\pi^*,\tau}$ explicitly depends on the temperature parameter τ, reflecting that the same prediction differences may be strategically relevant at one temperature but negligible at another. When τ is small, the softmax policy becomes more pointed, and minor prediction differences can yield distinct action preferences. As τ increases, the policy smooths, and more configurations

become equivalent. This temperature dependence ensures that epistemic equivalence respects the agent's current sensitivity to internal disagreement.

The canonical abstraction is then:

$$\varphi_{\text{can},\tau}(Q) := [\![Q]\!]_{\sim_{\pi^*,\tau}} \tag{2}$$

which maps each configuration to its equivalence class under decision-boundary preservation.

Canonical abstraction properties. The canonical abstraction $\varphi_{\text{can},\tau}$ represents the coarsest possible grouping of expert configurations that preserves all decision-relevant information. It collapses exactly those distinctions that do not affect the agent's choice distribution, while maintaining all strategic differences. This makes it the natural benchmark against which other abstractions should be measured.

Optimality of the canonical abstraction. Having defined the canonical equivalence relation, we now establish its fundamental optimality properties. The key insight is that $\varphi_{\text{can},\tau}$ represents the coarsest possible abstraction that maintains epistemic sufficiency, making it the natural benchmark for evaluating other abstraction schemes.

Proposition 1 (Maximally coarse safe abstraction)**.** *The canonical abstraction $\varphi_{can,\tau}$ satisfies $\Delta_{SEC}(\varphi_{can,\tau}) = 0$. Furthermore, any abstraction φ with $\Delta_{SEC}(\varphi) = 0$ must refine $\varphi_{can,\tau}$; that is, $\varphi^{-1}(\varphi(Q)) \subseteq \varphi^{-1}_{can,\tau}(\varphi_{can,\tau}(Q))$ for all Q.*

Understanding the refinement condition. The refinement condition $\varphi^{-1}(\varphi(Q)) \subseteq \varphi^{-1}_{\text{can},\tau}(\varphi_{\text{can},\tau}(Q))$ formalizes the intuition that any epistemically safe abstraction must be at least as fine-grained as the canonical one. In concrete terms, if two configurations Q and Q' are grouped together by some abstraction φ, they must also be grouped together by $\varphi_{\text{can},\tau}$. This means that φ cannot merge configurations that the canonical abstraction keeps separate, doing so would violate epistemic sufficiency by conflating strategically distinct expert profiles.

Proof. If $\varphi(Q) = \varphi(Q')$ and $\Delta_{\text{SEC}}(\varphi) = 0$, then $\text{KL}(\pi^*_{\text{raw}}(\cdot|Q) \| \pi^*_{\text{raw}}(\cdot|Q')) = 0$. By non-negativity of KL divergence, zero implies equality of distributions, hence $Q \sim_{\pi^*,\tau} Q'$. Therefore $\varphi^{-1}(\varphi(Q)) \subseteq \varphi^{-1}_{\text{can},\tau}(\varphi_{\text{can},\tau}(Q))$. □

Proof interpretation. The proof establishes the refinement property through a straightforward logical chain. When an abstraction φ achieves perfect epistemic sufficiency ($\Delta_{\text{SEC}}(\varphi) = 0$), any two configurations it groups together must induce identical policies. But this policy identity is precisely the defining criterion for canonical equivalence. Therefore, every grouping made by φ must respect the canonical equivalence classes, meaning φ can only create finer partitions, never coarser ones. The KL divergence [Cover and Thomas, 2006] provides the natural measure for policy differences under this epistemic framework.

Uniqueness and maximality. This proposition establishes that $\varphi_{\text{can},\tau}$ occupies a unique position in the space of all possible abstractions: it is the unique maximally coarse abstraction that preserves epistemic sufficiency. Any attempt to create a coarser abstraction, one that merges configurations separated by $\varphi_{\text{can},\tau}$, must necessarily violate the sufficiency condition by conflating configurations with different induced policies.

Practical implications for abstraction design. The optimality result provides crucial guidance for practical abstraction construction. First, it establishes an upper bound on compression: no abstraction can be both epistemically safe and more aggressive than $\varphi_{\text{can},\tau}$ in merging configurations. Second, it suggests a principled evaluation criterion: practical abstractions should be measured by how closely they approximate the canonical partition while remaining computationally tractable.

Connection to information theory. From an information-theoretic perspective, $\varphi_{\text{can},\tau}$ achieves the minimal sufficient statistic [Lehmann and Casella, 2005] for decision-making under the current temperature regime. It retains exactly the information needed to preserve the soft best-response policy while discarding all strategically irrelevant details about expert predictions. This connects our epistemic framework to classical statistical decision theory, where sufficient statistics play a similar role in preserving decision-relevant information.

Limitations and computational reality. While theoretically elegant, the canonical abstraction faces significant computational challenges. In continuous or high-dimensional prediction spaces, enumerating all equivalence classes under policy identity may be intractable. Moreover, checking whether two configurations induce identical policies requires exact computation, which may be numerically unstable. These practical limitations motivate the need for tractable approximations that sacrifice some optimality for computational feasibility.

Interpretation. This proposition establishes $\varphi_{\text{can},\tau}$ as the unique maximally coarse abstraction that preserves epistemic sufficiency. Any abstraction that achieves perfect sufficiency ($\Delta_{\text{SEC}} = 0$) must be at least as fine-grained as the canonical abstraction. This provides a theoretical upper bound on compression: no abstraction can be both epistemically safe and coarser than $\varphi_{\text{can},\tau}$.

Connection to modal epistemology. The canonical equivalence relation mirrors the accessibility relations in Kripke semantics for epistemic logic. Just as possible worlds are equivalent when they satisfy the same epistemic formulas, expert configurations are equivalent when they induce the same decision policies. This connection formalizes the intuition that epistemic arbitration should preserve exactly the distinctions that matter for the agent's knowledge-to-action translation.

Practical implications. While $\varphi_{\text{can},\tau}$ provides the theoretical standard, it may be computationally intractable for continuous or high-dimensional prediction spaces. The canonical abstraction serves as a theoretical foundation rather than a practical algorithm. Its primary value lies in establishing the epistemic boundary: any tractable approximation must respect the constraint that strategically equivalent configurations remain grouped together.

2.2 Tractable Approximation via Constrained Optimization

Since $\varphi_{\text{can},\tau}$ may be computationally intractable, we derive practical abstractions by solving:

$$\varphi^* = \arg\min_{\varphi \in \mathcal{F}} I(Q; \varphi(Q)) \quad \text{subject to} \quad \Delta_{\text{SEC}}(\varphi) \leq \varepsilon \tag{C.1}$$

where $I(\cdot;\cdot)$ denotes mutual information and $\mathcal{F}$ is a tractable family of abstraction functions corresponding to the decision-boundary preserving clustering, sparse attention masks, and information bottleneck methods detailed in Section 6.3.

This approach shifts the burden of justification from *post-hoc testing* φ to *principled construction*: φ is admissible *only* if it preserves the decision boundary up to ε. To ensure feasibility, choose ε no smaller than $0.5 \cdot \min_{\text{class}} \text{KL}$ observed on validation samples.

Any optimizer φ^* solving (C.1) automatically satisfies the epistemic sufficiency requirement $\Delta_{\text{SEC}}(\varphi^*) \leq \varepsilon$, though stability (Section 3) must be verified separately via Algorithm 2 or by extending (C.1) with an additional constraint $\Delta_{\text{SEC}_{\text{plural}}}(\varphi) \leq \varepsilon$.

The constraint can be enforced via Lagrangian methods or penalty approaches. In practice, we employ the empirical proxy:

$$\mathcal{L}_{\text{EIB}}(\varphi) = \widehat{H}[\varphi(Q)] - \beta \widehat{I}[\pi^*_\varphi; Q] + \lambda \widehat{\Delta}_{\text{SEC}}(\varphi) \tag{C.2}$$

with λ chosen large enough to enforce the epistemic boundary, where $\widehat{\Delta}_{\text{SEC}}$ uses the sample-bounded estimator from Section 6.1.

3 Epistemic Stability

Beyond individual knowledge, coherent action demands robustness akin to *distributed knowledge*—what the coalition would know if their beliefs were pooled Fagin et al. [1995]. Epistemic stability therefore requires that every admissible pooling W yield an indistinguishable policy. While epistemic sufficiency ensures that abstraction preserves decision-relevant distinctions, it does not guarantee that internal disagreement has been meaningfully reconciled. Even within a sufficient abstraction class, different aggregations of expert predictions may yield divergent policies. *Epistemic stability* addresses this residual uncertainty: it ensures that no admissible arbitration weighting over experts induces a materially different action distribution.

Motivation. Rather than assuming uniform expert weighting, the agent must consider multiple plausible trust distributions, realizing distributed knowledge aggregation [Fagin et al., 1995, §2]. This expert aggregation problem connects to classical social choice theory [Arrow, 1951], though here applied to internal epistemic arbitration rather than social preferences. This internal consensus problem connects to classical convergence results [DeGroot, 1974], though here applied to epistemic arbitration rather than belief updating. We provide principled methods for constructing the admissible weighting set $\mathcal{W} \subset \Delta_K$.

Construction Method 1: Competence-based weights Define expert competence scores $c_k \in [0, 1]$ based on:

- Historical accuracy on validation tasks
- Confidence calibration metrics
- Domain-specific expertise indicators

Then construct $\mathcal{W}_{\text{comp}} = \{W \in \Delta_K : W_k \propto \exp(\alpha c_k), \alpha \in [0, \alpha_{\max}]\}$ for temperature parameter $\alpha_{\max}$.

Construction Method 2: Uncertainty-aware weights For each expert k, estimate prediction uncertainty $u_k = \text{Var}(Q_k)$ and define:

$$\mathcal{W}_{\text{unc}} = \left\{ W \in \Delta_K : W_k = \frac{(1/u_k)^\beta}{\sum_j (1/u_j)^\beta}, \beta \in [0, \beta_{\max}] \right\}$$

This emphasizes confident experts while maintaining robustness to uncertainty estimation errors.

Construction Method 3: Adversarial robustness Consider worst-case expert reliability by including:

$$\mathcal{W}_{\text{robust}} = \{W \in \Delta_K : \|W - W_{\text{uniform}}\|_1 \leq \rho\}$$

for robustness radius $\rho > 0$. This ensures stability even if trust assumptions are violated.

Practical recommendation: Use the union $\mathcal{W} = \mathcal{W}_{\text{comp}} \cup \mathcal{W}_{\text{unc}} \cup \mathcal{W}_{\text{robust}}$ to capture multiple sources of weighting uncertainty. These construction methods instantiate principled aggregation functions [Grabisch et al., 2009] adapted for epistemic arbitration under internal disagreement.

Weighted soft best response. For any weighting $W \in \Delta_K$ over experts, define the induced soft best response:

$$\pi^*_W(a) \propto \exp\left(\frac{1}{\tau} \sum_{k=1}^{K} W_k Q_k(a)\right)$$

This generalizes the uniform averaging used in the abstraction policy. The agent may consider a finite set $\mathcal{W} \subset \Delta_K$ of admissible arbitration weights constructed using the methods above.

Stability criterion. We define the maximum policy divergence across these admissible weightings as:

$$\Delta\text{SEC}_{\text{plural}} := \max_{W_1, W_2 \in \mathcal{W}} \text{KL}\left(\pi^*_{W_1} \,\|\, \pi^*_{W_2}\right)$$

The abstraction is said to be *epistemically stable* if:

$$\Delta\text{SEC}_{\text{plural}} < \epsilon$$

for the same divergence threshold $\epsilon > 0$. This ensures that the agent's uncertainty over expert aggregation does not result in unstable or contradictory actions.

Interpretation. Epistemic stability imposes a behavioral coherence condition: all admissible aggregations must lead to policies that are effectively indistinguishable. Unlike sufficiency, which tests stability across different prediction configurations within an abstraction class, stability tests robustness to internal weighting under fixed predictions. Both are needed to justify action.

Relation to abstention. If $\Delta\text{SEC}_{\text{plural}} \geq \epsilon$, then internal arbitration remains unresolved: different interpretations of the same abstraction lead to different behaviors. In this case, the agent abstains. Stability thus provides a second gate, complementing sufficiency, on whether internal plurality has truly been reconciled.

Tractability considerations. Computing $\Delta\text{SEC}_{\text{plural}}$ over all of Δ_K is intractable. However, our construction methods provide finite, tractable weighting sets. In practice, the agent may:

- Use the principled construction methods above to generate $\mathcal{W}$;
- Apply sparse priors or trust-based kernels within each construction method;
- Bound divergence using closed-form approximations for distributions over bounded Q-values.

Summary. Epistemic stability ensures that abstraction-induced aggregation yields coherent action, even under uncertainty about internal weighting. It complements epistemic sufficiency by verifying that compression is not just justified in form, but stable in outcome. Together, these two criteria define a minimal epistemic standard for justified arbitration.

3.1 Adaptive Weighting Set Construction

Algorithm 2: Dynamic weighting set construction

1. Initialize with uniform weighting: $\mathcal{W} = \{W_{\text{uniform}}\}$
2. For each construction method $m \in \{\text{comp}, \text{unc}, \text{robust}\}$:
   (a) Generate candidate weights $\mathcal{W}_m$
   (b) Compute $\Delta_{\text{SEC}}^{\text{plural}}$ over $\mathcal{W} \cup \mathcal{W}_m$
   (c) If $\Delta_{\text{SEC}}^{\text{plural}} < \epsilon$, add $\mathcal{W}_m$ to $\mathcal{W}$
3. Return largest admissible weighting set

Proposition 2 (Weighting set optimality)**.** *Algorithm 2 returns the largest weighting set satisfying epistemic stability, ensuring maximal robustness to trust specification while maintaining epistemic guarantees.*

Computational complexity: Each iteration requires $O(|\mathcal{W}_m| \cdot |A|)$ evaluations, making the total complexity $O(K^2 \cdot |A|)$ for dense weighting exploration or $O(K \cdot |A|)$ for sparse parameterizations.

4 Well-Foundedness and Epistemic Norm Coordination

This section presents our core theoretical contribution through two complementary lenses. We first develop the technical framework of well-foundedness, grounded in hierarchical agency and computational constraints. We then present an equivalent coordination-theoretic interpretation based on epistemic norm emergence. Finally, we establish formal equivalence between these frameworks, demonstrating that they yield identical computational procedures while providing different conceptual foundations for understanding justified action under internal disagreement.

4.1 Well-Foundedness

We now define *well-foundedness* as the structural condition under which an agent's arbitration procedure is epistemically justified in proceeding to action. In the one-step pluralist bandit (1SPB) setting, well-foundedness is not a property of the environment or expert predictions per se—it is a property of the agent's internal epistemic structure. It holds only when abstraction is both strategically sound and behaviorally stable. The dual test below recreates common knowledge of readiness internally: only when both sufficiency and stability hold [Fagin et al., 1995, Prop. 6.1.2] may the agent act.

Definition. An agent's decision is said to be *well-founded* if and only if both of the following conditions are satisfied:

1. **Epistemic sufficiency:** For any two expert configurations Q, Q' that map to the same abstraction class under ϕ, their induced policies are nearly identical:

$$\Delta\text{SEC} := \max_{Q,Q' \in \varphi^{-1}(\varphi(Q))} \text{KL}(\pi^*(\cdot \mid Q) \,\|\, \pi^*(\cdot \mid Q')) < \epsilon$$

2. **Epistemic stability:** For any two admissible expert weightings W_1, W_2, their induced policies are nearly identical:

$$\Delta\text{SEC}_{\text{plural}} := \max_{W_1,W_2 \in \mathcal{W}} \text{KL}(\pi^*_{W_1} \,\|\, \pi^*_{W_2}) < \epsilon$$

These two criteria jointly ensure that: (a) abstraction does not erase distinctions that matter for decision-making, and (b) remaining internal disagreement does not deform the action boundary beyond an acceptable threshold.

Interpretation. Well-foundedness expresses minimal epistemic coherence. It ensures that the agent neither collapses meaningful distinctions nor acts amid unresolved ambiguity. If either condition fails, the agent abstains—not due to uncertainty about outcomes, but due to internal misalignment about what distinctions matter. This reflects a form of epistemic discipline: action is permitted only when internal representations have converged to relevance-preserving and influence-stable forms.

Practical implications. The well-foundedness criterion can be implemented as a static check prior to action. It does not require modeling external states, estimating expected returns, or simulating future rollouts. Instead, it verifies that the agent's internal compression and arbitration mechanisms satisfy a minimal agreement condition. This makes it applicable in planner-free, feedback-free, and decentralized settings where overcommitment may be costly or unjustified.

Theoretical connection. Well-foundedness generalizes the logic of Strategic Equivalence Relations (SER) Lauffer et al. [2023] to internally plural agents. In SER, agents discard distinctions among co-policies unless they induce different best responses. Here, we apply the same epistemic logic internally: distinctions among expert predictions (and their aggregations) are retained only when they alter the soft best response. The result is a principled mechanism for coordination under epistemic plurality.

Summary. Well-foundedness defines when an agent's abstraction and aggregation cohere into a stable, justified action. It ensures that internal disagreement has been reconciled to the degree required for principled commitment. When this threshold is not met, the agent abstains preserving epistemic integrity over premature convergence.

4.2 Coordination-Theoretic Interpretation: Epistemic Norm Coordination

We now present an equivalent formulation based on coordination theory and convention emergence. From this perspective, *epistemic norm coordination* is the structural condition under which a coalition's arbitration procedure is epistemically justified in proceeding to action. In the 1SPB setting, norm coordination is not a property of expert predictions per se—it is a property of the coalition's capacity to achieve common knowledge about decision-relevance [Fagin et al., 1995].

Epistemic Norm Coordination Principle. Action is epistemically justified if and only if the coalition satisfies three qualitative coordination conditions:

1. **Norm Existence**: A canonical abstraction $\varphi_{can,\tau}$ exists and is well-defined
2. **Norm Coherence**: $\varphi_{can,\tau}$ preserves all and only decision-relevant distinctions
3. **Norm Consensus**: All admissible coalition interpretations respect $\varphi_{can,\tau}$

When any condition fails, the coalition lacks common knowledge about decision-relevance, making abstention necessary to preserve epistemic integrity.

Formal Coordination Conditions. Condition 1: Norm Existence (Canonical Abstraction Well-Definedness)

$$\exists! \varphi_{can,\tau} : [0,1]^{K \times |\mathcal{A}|} \to \Phi \text{ such that } \varphi_{can,\tau} \text{ is unique and stable}$$

This condition ensures the coalition can formulate a coherent norm about what distinctions matter. The canonical abstraction represents the coalition's emergent convention about decision-relevance.

Condition 2: Norm Coherence (Internal Consistency)

$$\varphi_{can,\tau} \text{ preserves all strategically relevant distinctions} \Leftrightarrow \tag{3}$$

$$\forall Q, Q' \in \varphi^{-1}_{can,\tau}(\varphi_{can,\tau}(Q)) : \pi^*(|Q) = \pi^*(|Q') \tag{4}$$

This condition ensures the proposed norm is internally consistent—it preserves exactly the distinctions it claims to preserve. Configurations grouped together under the canonical abstraction must induce identical policies.

Condition 3: Norm Consensus (Shared Understanding)

$$\forall W_1, W_2 \in \mathcal{W} : \pi^*_{W_1} \approx \pi^*_{W_2} \text{ under } \varphi_{can,\tau}$$

This condition ensures different members of the coalition interpret the norm the same way. All admissible interpretations of the canonical abstraction must converge to consistent action policies.

Coordination-Theoretic Interpretation. **Epistemic integrity** emerges from the coalition's capacity to maintain coherent shared norms about decision-relevance, not from keeping disagreement below arbitrary thresholds. **Convention emergence** occurs when the coalition spontaneously coordinates on $\varphi_{can,\tau}$ without external enforcement—a computational instantiation of common knowledge achievement. This connects to recent work on learning normativity [Demski, 2020, 2021], which addresses how agents can learn appropriate norms without perfect feedback or gold-standard training signals. **Abstention as coordination failure**: When norm coordination fails, abstention becomes necessary not because "disagreement is too high," but because there is no coherent "coalition decision" to be made. The coalition lacks the epistemic common ground required for justified action.

Connection to Norm Learning. This framework connects to recent advances in computational norm learning [Oldenburg and Zhi-Xuan, 2024; Tan and Ong, 2019; Demski, 2020], where agents achieve coordination by "assuming there exists a shared set of norms that most others comply with while pursuing their individual desires" [Oldenburg and Zhi-Xuan, 2024]. This builds on foundational work in norm learning [Tan et al., 2019; Demski, 2020] and Bayesian approaches to social norm inference [Tan and Ong, 2019].

4.3 Equivalence and Translation Between Frameworks

We now establish formal equivalence between well-foundedness and epistemic norm coordination, demonstrating that they represent different conceptual lenses on identical mathematical procedures.

Fundamental Equivalence Theorem.

Theorem 1 (Framework Equivalence)**.** *An agent's decision is well-founded if and only if the agent's internal coalition has achieved epistemic norm coordination. Formally:*

$$\textit{Well-Founded}(\textit{Agent}) \Leftrightarrow \textit{Norm-Coordinated}(\textit{Coalition}) \tag{5}$$

$$\Delta\text{SEC} < \epsilon \wedge \Delta\text{SEC}_{plural} < \epsilon \Leftrightarrow \textit{Norm Existence} \wedge \textit{Norm Coherence} \wedge \textit{Norm Consensus} \tag{6}$$

Translation Mappings. The frameworks are related by the following conceptual translations:

Well-Foundedness	**Norm Coordination**	**Mathematical Reality**
Canonical abstraction $\varphi_{can,\tau}$	Emergent epistemic convention	Same optimality result (Prop. 1)
Epistemic sufficiency	Norm coherence achievement	Same: $\Delta\text{SEC} < \epsilon$
Epistemic stability	Norm consensus achievement	Same: $\Delta\text{SEC}_{\text{plural}} < \epsilon$
Internal misalignment	Coordination failure	Same computational gate
Abstention threshold ϵ	Context-dependent consensus req.	Same divergence bound

Computational Procedures. Both frameworks yield identical algorithms:

1. **Canonical abstraction construction**: Same constrained optimization (Section 2.2)
2. **Sufficiency testing**: Same ΔSEC computation via sampling bounds
3. **Stability testing**: Same $\Delta\text{SEC}_{\text{plural}}$ computation over admissible weightings
4. **Action/abstention decision**: Same dual gate condition

Conceptual Benefits of Dual Formulation. The equivalence reveals deep connections between apparently distinct research traditions:

From well-foundedness: Connects to hierarchical agency theory [Ngo, 2025a], Strategic Equivalence Relations [Lauffer et al., 2023], and computational tractability constraints.

From norm coordination: Connects to common knowledge theory [Aumann, 1976], convention emergence [Lewis, 1969], computational social choice [Arrow, 1951], and recent advances in computational norm learning [Tan and Ong, 2019; Oldenburg and Zhi-Xuan, 2024; Demski, 2020].

Threshold interpretation: Well-foundedness treats ϵ as a computational necessity; norm coordination treats ϵ as context-dependent coordination requirements.

Summary. This dual formulation demonstrates that our framework captures fundamental principles that emerge independently from both computational constraints (well-foundedness) and coordination requirements (norm emergence). The mathematical inevitability of the same procedures arising from different theoretical starting points suggests that we have identified essential structural features of justified action under internal disagreement, rather than arbitrary design choices.

5 Illustrative Example: Abstraction Failure and Diagnostic Divergence

We now illustrate the epistemic diagnostics using a minimal example in the one-step pluralist bandit (1SPB) setting. The goal is to demonstrate how abstraction can fail either sufficiency or stability, and how these failures are detected by the ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ signals.

Setup. Let the action set be $\mathcal{A} = \{a_1, a_2, a_3\}$, and let the agent consist of $K = 3$ internal experts. Each expert provides a prediction vector $Q_k : \mathcal{A} \to [0, 1]$. Consider two configurations of expert predictions:

$$Q^{(1)} = [Q_1 = [1.0, 0.0, 0.0], \quad Q_2 = [1.0, 0.0, 0.0], \quad Q_3 = [1.0, 0.0, 0.0]]$$
$$Q^{(2)} = [Q_1 = [1.0, 0.0, 0.0], \quad Q_2 = [1.0, 0.1, 0.0], \quad Q_3 = [1.0, 0.0, 0.0]]$$

Suppose the agent uses a simple averaging abstraction:

$$\varphi(Q) := \frac{1}{K} \sum_{k=1}^{K} Q_k$$

and defines its soft best response using temperature $\tau = 0.2$:

$$\pi^*(a) \propto \exp\left(\frac{1}{\tau} \cdot \varphi(Q)(a)\right)$$

Computing soft best responses. The averaged prediction vectors are:

$$\varphi(Q^{(1)}) = [1.0, 0.0, 0.0]$$
$$\varphi(Q^{(2)}) = \left[\frac{3}{3}, \frac{0.1}{3}, 0\right] = [1.0, 0.0333, 0.0]$$

Plugging into the softmax yields:

$$\pi^*(\cdot \mid Q^{(1)}) \approx \text{softmax}\left(\frac{1}{0.2} \cdot [1.0, 0.0, 0.0]\right) = \text{softmax}([5.0, 0.0, 0.0]) \approx [0.88, 0.06, 0.06]$$
$$\pi^*(\cdot \mid Q^{(2)}) \approx \text{softmax}([5.0, 0.167, 0.0]) \approx [0.82, 0.12, 0.06]$$

KL divergence diagnostic. The induced divergence is:

$$\Delta\text{SEC} := \text{KL}\left(\pi^*(\cdot \mid Q^{(1)}) \,\|\, \pi^*(\cdot \mid Q^{(2)})\right) \approx 0.019$$

This nonzero divergence reflects that φ masks a subtle but strategically relevant distinction: a small deviation in expert 2's belief increases the probability of a_2 by a factor of 2.

If the agent sets $\epsilon = 0.01$, this abstraction would fail the sufficiency criterion.

Testing epistemic stability. Now consider arbitration via weighted aggregation. Let:

$$W^{(1)} = [1.0, 0.0, 0.0] \quad \text{(trust only expert 1)}$$
$$W^{(2)} = [0.0, 1.0, 0.0] \quad \text{(trust only expert 2)}$$

Then:

$$\pi^*_{W^{(1)}} = \text{softmax}\left(\frac{1}{0.2} \cdot Q_1\right) = \text{softmax}([5.0, 0.0, 0.0]) \approx [0.88, 0.06, 0.06]$$
$$\pi^*_{W^{(2)}} = \text{softmax}\left(\frac{1}{0.2} \cdot Q_2\right) = \text{softmax}([5.0, 0.5, 0.0]) \approx [0.78, 0.16, 0.06]$$

The induced divergence is:

$$\Delta\text{SEC}_{\text{plural}} := \text{KL}(\pi^*_{W^{(1)}} \,\|\, \pi^*_{W^{(2)}}) \approx 0.041$$

This exceeds $\epsilon = 0.01$, indicating instability: different arbitration weightings within the same abstraction class produce different policies. The agent must abstain, even if the abstraction is sufficient. Here, abstention is the exact analogue of the generals' failure to coordinate when common knowledge is absent in the classic Coordinated-Attack problem [Fagin et al., 1995, Prop. 6.1.2].

Summary. This example illustrates how minor deviations in expert predictions or weighting can cause soft best response policies to diverge. The diagnostics ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ serve as principled tests for epistemic readiness. Abstention is triggered when abstraction either discards decision-relevant distinctions (sufficiency failure) or leaves arbitration unresolved (stability failure).

6 Computational Tractability

The exact computation of Δ_{SEC} and $\Delta_{\text{SEC}_{\text{plural}}}$ requires optimization over continuous, high-dimensional sets, which is generally intractable. We address this by developing sample-based approximations with provable guarantees.

Both diagnostics admit sample-bounded estimators with runtime $\mathcal{O}((K+S)|A|)$, making previously intractable checks linear-time subroutines. Sampling approximates ε-common knowledge [1] gates through concentration bounds. The key insight is replacing exact optimization over continuous sets with sampling plus Lipschitz-based error bounds.

Specifically, the softmax policy $\pi^*_{\varphi,\tau}$ is Lipschitz with constant $L(\tau) = \frac{|A|-1}{2\tau}$, enabling concentration bounds: for δ-dense proposal distributions over $\varphi^{-1}(\varphi(Q))$, sampling S pairs and computing the empirical maximum KL divergence yields

$$\Pr[\Delta_{\text{SEC}} - \widehat{\Delta}_{\text{SEC}} \leq L\delta] \geq 1 - 2|A|e^{-S\delta^2/2}$$

where $\widehat{\Delta}_{\text{SEC}}$ is the sample-based estimator. The same bound applies to $\Delta_{\text{SEC}_{\text{plural}}}$ using finite covers of the admissible weighting set $\mathcal{W}$.

This computational tractability is crucial for deploying epistemic arbitration in real-time applications where both principled decision-making and efficient performance are required.

7 Epistemic Parameter Sensitivity

The epistemic arbitration framework relies on two key parameters that govern its abstention behavior: the divergence threshold $\epsilon > 0$, and the temperature parameter $\tau > 0$. These control, respectively, how much behavioral variation is tolerated within an abstraction class, and how strongly the agent responds to differences in aggregated predictions. Together, they define the resolution and sensitivity of the epistemic gate. Parameters τ and ϵ control epistemic sensitivity; lower values increase abstention frequency but provide stronger guarantees.

References

Kenneth J Arrow. *Social Choice and Individual Values*. Wiley, New York, 1951.

Richard Ngo. Towards a scale-free theory of agency. *AI Alignment Forum*, 2025a. URL `https://www.alignmentforum.org/posts/gHefoxiznGfsbiAu9/towards-a-scale-free-theory-of-agency`.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.

Leonard J Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Richard Ngo. Well-foundedness as an organizing principle of healthy minds and societies. *Mind the Future*, 2025b. URL `https://www.mindthefuture.info/p/well-foundedness-as-an-organizing`.

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].

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

Brian F Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

Alvin I Goldman. *Knowledge in a Social World*. Oxford University Press, 1999.

CK Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 4878–4887, 2017.

[1] See [Fagin et al., 1995, Thm. 11.6.2] on ε-coordination in unreliable networks.

Richard Ngo. Trust develops gradually via making bids and setting boundaries. *LessWrong*, 2023. URL https://www.lesswrong.com/s/qXZLFGqpD7aeEgXGL/p/7CKF6r8MegtcCWDbT.

Jan Kulveit. Hierarchical agency: A missing piece in ai alignment. AI Alignment Forum, November 2024. URL https://www.alignmentforum.org/posts/xud7Mti9jS4tbWqQE/hierarchical-agency-a-missing-piece-in-ai-alignment. Accessed: July 26, 2025.

Aaron Kirtland, Alexander Ivanov, Cameron Allen, Michael L. Littman, and George Konidaris. Memory as state abstraction over trajectories. https://memory-as-abstraction.github.io/, 2025. Accessed: July 23, 2025.

Robert J Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

Yoram Moses and Moshe Tennenholtz. Distributed epistemic algorithms: Knowledge, coordination and common knowledge. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 289–306, 1995.

Edwin T Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2006.

Erich L Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2005.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap. *Aggregation Functions*. Cambridge University Press, 2009.

Abram Demski. Learning normativity: A research agenda. AI Alignment Forum, 2020. URL https://www.alignmentforum.org/posts/2JGu9yxiJkoGdQR4s/learning-normativity-a-research-agenda.

Abram Demski. Four motivations for learning normativity. AI Alignment Forum, 2021. URL https://www.alignmentforum.org/posts/oqghwKKifztYWLsea/four-motivations-for-learning-normativity.

Ninell Oldenburg and Tan Zhi-Xuan. Learning and sustaining shared normative systems via bayesian rule induction in markov games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024.

Zhi-Xuan Tan and Desmond C. Ong. Bayesian inference of social norms as shared constraints on behavior. 2019. URL https://arxiv.org/abs/1905.11110.

Zhi-Xuan Tan, Jake Brawer, and Brian Scassellati. That's mine! learning ownership relations and norms for robots, 2019. URL https://arxiv.org/abs/1812.02576.

David K. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.

A Appendix

A.1 Extensions and Future Directions

A.1.1 Multi-Step Extension with Feedback

While our core framework addresses one-step decisions, extension to sequential settings is natural. Suppose the agent receives a binary signal $s \in \{0, 1\}$ indicating (out-of-distribution) regret after acting or abstaining. Each expert updates a trust weight w_k via:

$$w_k \leftarrow \frac{\exp\big(\eta \, [\![s = 1]\!] \, Q_k(a_t)\big) w_k}{\sum_j \exp\big(\eta \, [\![s = 1]\!] \, Q_j(a_t)\big) w_j}$$

A.1.2 Strategic Experts and Mechanism Design

Future work could address experts who strategically misreport predictions, requiring truthful elicitation mechanisms while preserving epistemic guarantees.

We provide comprehensive complexity bounds and practical performance analysis for all computational routines in our framework.

A.2 Detailed Complexity Analysis

A.2.1 Algorithmic Complexity Bounds

Table 1 summarizes the asymptotic costs for the main algorithmic components.

Routine	Time	Space
Approx ΔSEC	$\mathcal{O}(S \cdot \lvert A \rvert)$	$\mathcal{O}(\lvert A \rvert)$
Approx $\Delta\text{SEC}_{plural}$	$\mathcal{O}(S \cdot \lvert A \rvert)$	$\mathcal{O}(\lvert A \rvert)$
Soft BR eval ($\pi^*_{\varphi,\tau}$)	$\mathcal{O}(K \cdot \lvert A \rvert)$	$\mathcal{O}(\lvert A \rvert)$
Well-foundedness check	$\mathcal{O}(S \cdot \lvert A \rvert)$	$\mathcal{O}(\lvert A \rvert)$

Table 1: Computational complexity bounds. S is the number of samples, K is the number of experts, $|A|$ is the number of actions. All bounds assume IEEE 754 double-precision arithmetic and pre-computed exponentials where applicable.

A.3 Detailed Breakdown

Sample-bounded diagnostics. Both Approx ΔSEC and Approx $\Delta\text{SEC}_{plural}$ require:

- S policy evaluations, each taking $\mathcal{O}(|A|)$ time for softmax computation
- S KL divergence calculations, each taking $\mathcal{O}(|A|)$ time
- Finding the maximum over S values, taking $\mathcal{O}(S)$ time

This yields $\mathcal{O}(S \cdot |A|)$ total time complexity.

Soft best response evaluation. Computing $\pi^*_{\varphi,\tau}(\cdot|Q)$ requires:

- Aggregating K expert predictions: $\mathcal{O}(K \cdot |A|)$
- Computing softmax over $|A|$ actions: $\mathcal{O}(|A|)$
- Total: $\mathcal{O}(K \cdot |A|)$

Well-foundedness check. The complete epistemic gate requires both sufficiency and stability checks, dominated by the $\mathcal{O}(S \cdot |A|)$ diagnostic computations.

A.4 Optimizations and Special Cases

Early termination conditions. Several scenarios allow for early termination with $\mathcal{O}(|A|)$ complexity:

- **Expert consensus**: When $\max_{i,j} \|Q_i - Q_j\|_\infty < \delta$ for small δ, skip diagnostic computation
- **Empty preimage**: When $|\varphi^{-1}(\varphi(Q))| = 1$, automatically pass sufficiency test
- **Extreme disagreement**: When initial sampling reveals $\widehat{\Delta\text{SEC}} > 2\varepsilon$, early rejection without full sampling. Our ε-thresholds instantiate ε-common-knowledge gates, while eventual convergence reflects the 'eventual common knowledge' achievable in lossy communication settings [Fagin et al., 1995, Thm. 11.6.3].

Practical optimizations.

- **Incremental sampling**: Start with small S, increase until confidence threshold met
- **Cached evaluations**: Store policy evaluations for repeated abstraction classes
- **Parallel computation**: Both diagnostics are embarrassingly parallel over samples
- **Adaptive precision**: Use $S = 100$–1000 samples with $\delta = 0.1$–0.01 for $\geq 99\%$ confidence