Well-Founded Arbitration Under Internal Epistemic Disagreement

Technical manuscript - please cite if referencing specific results

Project Description

Epistemically Well-Founded Arbitration and Formal Criteria for Principled Action under Internal Pluralism

Abstract

We study the problem of principled expert aggregation in AI systems where internal disagreement may signal genuine epistemic uncertainty rather than noise to be averaged away. Standard ensemble methods like weighted voting [Arr51], bagging [Bre96], or Bayesian model averaging [Hoe+99] often force consensus even when expert predictions fundamentally conflict, potentially erasing decision-relevant information and producing overconfident outputs in safety-critical domains. We propose a framework for **epistemically well-founded arbitration**, in which agents abstain from action when internal expert disagreements cannot be resolved without losing strategically relevant distinctions. Using the one-step pluralist bandit (1SPB) setting, a minimal decision problem with no feedback or iteration, we formalize two key diagnostics: $\Delta\mathrm{SEC}$ measures whether abstraction over expert predictions preserves decision-relevant information, while $\Delta\mathrm{SEC}_{\mathrm{plural}}$ ensures stability across different expert weighting schemes. An agent's decision is well-founded only when both diagnostics remain below a divergence threshold ϵ. We provide computationally tractable approximation algorithms with theoretical guarantees, demonstrate the framework through illustrative examples, and establish connections to Strategic Equivalence Relations [Lau+23]. The approach enables **principled abstention** in ensemble systems: maintaining epistemic discipline rather than defaulting to potentially premature consensus. While this work is scoped to internal arbitration within single agents, it points toward a broader foundation for epistemic alignment grounded in shared relevance rather than forced coherence.

Risk Factors

Current AI ensemble deployment creates a systematic pathway to catastrophic failure through **epistemic collapse under centralized decision-making**. As AI systems scale and consolidate into fewer, more powerful ensemble architectures [Amo+16b], a dangerous dynamic emerges: centralized aggregation methods systematically mask genuine uncertainty precisely when distributed expert disagreement should signal caution about high-stakes decisions. The failure pathway operates as follows: (1) Expert models within large-scale ensembles disagree about novel or edge-case scenarios that pose existential or systemic risks, (2) Centralized aggregation algorithms (weighted voting [Arr51], Bayesian averaging [Hoe+99]) smooth over this disagreement to maintain operational efficiency, (3) The resulting overconfident predictions get amplified through concentrated deployment across critical infrastructure, and (4) Synchronized failures occur across multiple domains simultaneously when the underlying uncertain scenarios materialize.

Key amplification factors driving catastrophic risk include: *Institutional concentration*—as AI capability concentrates in fewer organizations [Rus19], uniform ensemble methods create correlated failure modes across systems; *Scale acceleration*—larger ensembles paradoxically increase overconfidence as aggregation methods treat genuine expert disagreement as noise to be eliminated [Guo+17]; *Feedback isolation*—centralized systems lack mechanisms to preserve minority expert opinions that might detect novel failure modes [Con85]; and *Deployment synchronization*—concentrated AI deployment means epistemic failures propagate across critical systems simultaneously [Ham+25].

This risk factor is **urgent and neglected** because it sits at the intersection of two accelerating trends: the concentration of AI capability in centralized ensemble architectures, and the deployment of these systems in domains where overconfident decisions can trigger cascading failures (markets, infrastructure control, autonomous weapons). Unlike traditional AI safety concerns focused on superintelligence, epistemic collapse is occurring *now* in deployed systems, with severity scaling directly with centralization and deployment scope.

The risk is **newly emerging** because previous ensemble methods were designed for predictive accuracy [Bre96; Zho12] rather than preserving epistemic diversity under uncertainty, a distinction that becomes critical as AI systems gain real-world agency and as deployment concentrates in fewer, more powerful architectures. Current uncertainty quantification [Ova+19; NC05] assumes centralized optimization is superior to distributed deliberation, creating systematic bias toward false consensus in exactly the scenarios where expert disagreement signals genuine uncertainty about catastrophic outcomes.

Our framework addresses this by enabling **principled abstention** when expert disagreement cannot be resolved without losing strategically relevant distinctions, preventing the epistemic overreach that drives catastrophic failures in concentrated AI systems.

Theory of Change

Our research operates on the premise that **epistemic infrastructure shapes coordination outcomes** [Ngo25b]: by providing formal criteria for when expert aggregation erases strategically relevant distinctions, we enable a transition from forced consensus to principled abstention in AI systems. This technical capability creates a pathway toward distributed coordination architectures that preserve rather than collapse epistemic diversity [Ngo25c].

The causal mechanism operates through **adoption incentives in safety-critical domains**. Organizations deploying AI systems in high-stakes contexts face increasing liability for overconfident decisions that ignore genuine uncertainty [Amo+16b; Rus19]. Our framework provides technical cover: formal diagnostics that demonstrate when abstention was epistemically justified, and computational tools that enable real-time implementation. Early adopters in medical diagnosis, autonomous systems, and risk assessment gain competitive advantage through reduced catastrophic failure rates [Guo+17; Ova+19].

Scaling occurs through regulatory and institutional pressure. As high-profile failures reveal the inadequacy of naive ensemble methods [Bre96; Zho12], regulatory bodies will demand epistemic justification for AI decisions in safety-critical domains [Ebe+21]. Our framework provides the technical infrastructure to meet these requirements, creating market pressure for adoption across ensemble AI systems. This regulatory capture accelerates transition from optimization-driven to epistemically-disciplined coordination [Doc23].

The **long-term transformation** operates through network effects: as more systems adopt principled abstention criteria, the coordination landscape shifts from centralized aggregation to distributed epistemic negotiation [Arr51; Con85]. Systems that can abstain and coordinate through shared relevance standards gain advantages in multi-agent environments [Lau+23], driving evolutionary pressure toward epistemic humility rather than overconfident optimization [Bos14].

Success metrics include: (1) Integration into production ensemble systems within 3-5 years, (2) Adoption by regulatory frameworks for AI safety evaluation [Ben+23], (3) Reduced catastrophic failure rates in early-adopting organizations, and (4) Citation by subsequent research developing distributed coordination protocols [Ngo25a].

Failure modes and contingencies: If adoption proves limited due to conservatism bias, the framework still succeeds as a diagnostic tool, revealing domains where current AI systems are epistemically underprepared. If computational overhead proves prohibitive, simplified versions targeting specific applications (medical diagnosis, autonomous systems) provide stepping stones toward broader adoption [LPB17].

Importance

This work establishes foundational infrastructure for a paradigmatic shift from centralized AI optimization to distributed epistemic coordination [Ngo25b]. Rather than treating expert disagreement as noise to be eliminated, our framework provides the **theoretical** and **computational basis** for AI systems that preserve and navigate genuine uncertainty through **principled abstention** and **wellfounded pluralistic arbitration** [Ngo25c]. The significance extends far beyond ensemble safety: we demonstrate how **well-founded pluralistic arbitration can serve as a minimal substrate for coordination without centralized control**, offering an alternative to the concentration of AI capability in monolithic architectures that drive enshittification dynamics [Doc23]. By enabling agents to maintain epistemic discipline rather than forcing premature consensus, this approach points toward AI development trajectories that enhance rather than diminish human autonomy through preserved epistemic diversity—countering the extractive optimization paradigms that characterize current platform capitalism [Doc21; Doc23]. If successful, this framework could fundamentally alter how AI systems scale and coordinate. Instead of capability concentration driving toward superintelligent centralized agents [Bos14], we envision distributed networks of epistemically disciplined systems that earn coherence through relevance convergence rather than optimization pressure. This represents a technical instantiation of pluralist coordination theory: systems that can abstain, maintain ambiguity, and coordinate through shared epistemic standards rather than imposed uniformity [Gra15].

The broader implications span multiple domains where current AI development risks epistemic collapse through premature consensus: from scientific research collectives to distributed autonomous organizations to human-AI collaborative systems. By providing formal criteria for when aggregation preserves versus erases strategically relevant distinctions [Lau+23], this work offers minimal foundational substrate for an infrastructure for coordination mechanisms that resist centralizing tendencies while maintaining operational effectiveness. This aligns with emerging visions of techno-humanist AI development [Ngo25a] that **prioritize epistemic humility** over false confidence, **distributed coordination** over centralized control, and **preservation of human agency** over optimization efficiency [And23], representing a foundational step toward AI systems that enhance rather than replace human epistemic capabilities. For extended philosophical grounding, see *Optional Appendix for Philosophers*.

Neglectedness

The specific problem of **epistemic abstention criteria** for ensemble systems represents a surprising gap in both AI safety and uncertainty quantification literature. Despite extensive research on ensemble methods [Bre96; Zho12] and uncertainty estimation [Guo+17; Ova+19], virtually no work addresses when aggregation should be avoided entirely. The field operates under an implicit assumption that producing some output is always preferable to abstention, even when expert disagreement signals genuine uncertainty about catastrophic scenarios.

This neglect stems from three disciplinary blind spots. Machine learning research focuses on predictive performance metrics that inherently favor output over abstention, creating institutional incentives to solve aggregation problems rather than question when aggregation is appropriate. AI safety research has concentrated on alignment problems assuming centralized agents [Bos14; Amo+16a; Rus19], largely overlooking distributed epistemic coordination as an alternative to centralized control. **Uncertainty quantification** treats expert disagreement as a calibration problem rather than a potential signal for principled non-action [NC05; Ova+19].

The theoretical foundations remain underdeveloped. While Strategic Equivalence Relations exist in game theory [Lau+23], their application to internal expert arbitration is entirely novel. No existing framework formalizes conditions under which compression of expert predictions erases strategically relevant distinctions, nor provides computationally tractable diagnostics for epistemic sufficiency in ensemble systems.

Existing uncertainty quantification work assumes that better aggregation methods solve the disagreement problem, rather than recognizing scenarios where disagreement itself carries crucial information. This gap is particularly concerning given recent critiques suggesting that safety frameworks systematically overestimate the reliability of centralized coordination mechanisms [Joe25]. The field lacks both theoretical foundations and practical tools for systems that can maintain epistemic humility when internal arbitration remains genuinely unresolved. Our framework addresses this specific neglected area: providing formal criteria and computational methods for principled abstention in multi-expert systems, representing the first systematic approach to epistemic justification for non-action in ensemble architectures.

Tractability

We have completed the theoretical foundations including:

- Formalization of the one-step pluralist bandit (1SPB) framework
- Definition and analysis of ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ diagnostics
- Sample-bounded approximation algorithms with Lipschitz guarantees
- Computational complexity analysis showing linear scaling $O(S \cdot |A|)$
- Theoretical connections to Strategic Equivalence Relations

Technical Details (Full Version)

This paper introduces a framework for **epistemically well-founded arbitration** in AI systems where internal expert disagreement may signal genuine epistemic uncertainty rather than noise to be averaged away. Standard ensemble methods [Rok10; Zho12] like weighted voting or Bayesian model averaging often force consensus even when expert predictions fundamentally conflict, potentially erasing decision-relevant information and producing overconfident outputs in safety-critical domains.

Our approach addresses the one-step pluralist bandit (1SPB) setting, where an agent must arbitrate among K internal experts, each producing scalar predictions $Q_k(a) \in [0, 1]$ for every action a. These predictions encode fixed priors, heuristics, or inductive biases rather than learned estimates. The agent acts once—if at all—based on a compressed representation of its internal disagreement, with no feedback or iterative structure [LS20].

The framework centers on two key epistemic diagnostics. First, Δ_{SEC} measures whether abstraction over expert predictions preserves decision-relevant information by quantifying the worst-case KL divergence [CT06] between soft best response policies across expert configurations grouped together by the abstraction ϕ. Second, $\Delta_{\text{SEC}}^{\text{plural}}$ ensures stability across different expert weighting schemes by measuring the maximum policy divergence across admissible arbitration weights $\mathcal{W} \subset \Delta_K$. An agent's decision is well-founded only when both diagnostics remain below a divergence threshold ε.

We provide computationally tractable approximation algorithms with theoretical guarantees. The exact computation of both diagnostics requires optimization over continuous, high-dimensional sets, which is generally intractable [Vap98]. We address this through sample-based approximations with Lipschitz bounds, yielding runtime $\mathcal{O}((K + S)|A|)$ where S is the number of samples. For the softmax policy $\pi^*_{\phi,\tau}$, we prove L-Lipschitz continuity with constant $L(\tau) = \frac{|A|-1}{2\tau}$, enabling confidence bounds for our sample-based estimators [MRT18].

The framework includes principled methods for constructing the admissible weighting set $\mathcal{W}$ through three approaches: competence-based weights using historical accuracy and expertise indicators, uncertainty-aware weights emphasizing confident experts while maintaining robustness, and adversarial robustness ensuring stability under worst-case expert reliability assumptions. We provide Algorithm 2 for dynamic weighting set construction that returns the largest set satisfying epistemic stability.

For abstraction design, we analyze three methods with formal guarantees: epistemic k-means clustering with decision-boundary preservation, information bottleneck optimization [TPB00] that directly penalizes epistemic violations, and sparse attention masks suitable for homogeneous expert ensembles. Each method comes with constructive sufficiency guarantees and practical implementation guidance.

The approach establishes theoretical connections to Strategic Equivalence Relations (SER) [Lau+23], generalizing their logic from external game-theoretic settings to internal epistemic arbitration. Just as external SER preserves strategic distinctions between different strategies, our framework preserves epistemic distinctions between different expert configurations when they alter the soft best response.

We demonstrate the framework through illustrative examples showing how minor deviations in expert predictions can cause soft best response policies to diverge, with the diagnostics serving as principled tests for epistemic readiness. Abstention is triggered when abstraction either discards decision-relevant distinctions (sufficiency failure) or leaves arbitration unresolved (stability failure).

The computational complexity analysis shows linear scaling in both sample size S and action space $|A|$, making the framework practical for real-time applications. This represents a dramatic improvement over the exponential complexity of exact computation over continuous abstraction classes, crucial for deploying epistemic arbitration in safety-critical systems where both principled decision-making and real-time performance are required.

The framework enables principled abstention in ensemble systems: maintaining epistemic discipline rather than defaulting to potentially unsafe consensus. While scoped to internal arbitration within single agents, it points toward a broader foundation for epistemic alignment grounded in shared relevance rather than forced coherence, offering a principled alternative to ad-hoc ensemble methods particularly valuable for safety-critical applications where unjustified confidence can be dangerous.

Progress Report and Future Plan

Current Status (as of Week 4)

Our current research focuses on the theoretical foundations of epistemically well-founded arbitration in one-step multi-agent bandits for submission to ICLR2026. We have formalized the framework with ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ diagnostics, established the well-foundedness criterion combining epistemic sufficiency and stability, and provided sample-bounded approximation algorithms with Lipschitz guarantees. The theoretical connections to Strategic Equivalence Relations and computational tractability analysis are complete, with an illustrative example demonstrating how epistemic failures trigger principled abstention. We have formalized the framework with ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ diagnostics, established the well-foundedness criterion combining epistemic sufficiency and stability, and provided sample-bounded approximation algorithms with Lipschitz guarantees. The theoretical connections to Strategic Equivalence Relations and computational tractability analysis are complete, with an illustrative example demonstrating how epistemic failures trigger principled abstention. Complete technical framework with proofs and algorithms is provided in the appendix (full technical version), demonstrating that we will be ready for ICLR2026 submission.

MATS Main Program Timeline (Weeks 5-10)

Weeks 5-6: Implementation and Baseline Experiments

- Implement core framework with ΔSEC and $\Delta\text{SEC}_{\text{plural}}$ diagnostics
- Create synthetic multi-expert scenarios with known ground truth
- Validate theoretical predictions against empirical results in controlled settings
- Establish baseline performance metrics

Weeks 7-8: Realistic Domain Applications

- Medical diagnosis ensemble: Implement expert models for rare condition detection using public medical datasets
- fMRI decoding: Create ensemble of brain state classifiers using different preprocessing pipelines and feature extraction methods for cognitive task prediction
- Document abstention rate patterns and correlations with true uncertainty

Weeks 9-10: Comparison with Existing Methods and Parameter Selection and Analysis

- Implement deep ensemble baselines [LPB17] and Bayesian model averaging [Hoe+99]
- Compare abstention decisions with uncertainty estimates from existing methods
- Analyze cases where methods agree/disagree on abstention
- Measure calibration and reliability across different uncertainty quantification approaches
- Develop adaptive calibration methods for ϵ and τ based on domain-specific risk tolerances
- Analyze sensitivity of abstention rates to parameter choices
- Create guidelines for practitioners on parameter selection
- Prepare initial draft for ICLR2026 submission

MATS Extension Timeline (Months 1-4)

Month 1: Paper Completion and Submission

- Finalize empirical evaluation results and analysis
- Complete related work section and positioning relative to existing uncertainty quantification
- Refine theoretical presentation and proofs
- Submit to ICLR2026 (deadline typically in October)

Month 2: Sequential Extensions

- Develop temporal consistency conditions for abstraction sequences
- Investigate dynamic expert reliability through trust weight updates [CL06]
- Implement prototype sequential decision-making framework

Month 3: Strategic Behavior and Mechanism Design

- Explore strategic expert behavior with mechanism design approaches
- Develop truthfulness conditions for expert reporting
- Analyze robustness to adversarial expert manipulation

Month 4: Scalability and Foundation Model Applications

- Address computational scalability for large expert ensembles through hierarchical methods
- Investigate framework application to foundation model uncertainty (parameter subsets, attention heads)
- Prepare follow-up publications based on extension work

Deliverables and Milestones

End of Main Program:

- Complete ICLR2026 submission with empirical validation
- Open-source implementation of the framework
- Practitioner guidelines for parameter selection

End of Extension:

- Sequential decision-making extension with temporal consistency
- Strategic expert behavior analysis with mechanism design
- Foundation model uncertainty application prototype
- 2-3 additional paper submissions to top-tier venues

Risk Mitigation:

If empirical evaluation proves more challenging than expected, we will focus on synthetic domains and theoretical analysis for the ICLR submission, while pursuing realistic applications during the extension period.

References

[Amo+16a] Dario Amodei et al. *Concrete Problems in AI Safety.* July 25, 2016. DOI: `10.48550/arXiv.1606.06565`. arXiv: `1606.06565[cs]`. URL: `http://arxiv.org/abs/1606.06565` (visited on 02/17/2025).

[Amo+16b] Dario Amodei et al. "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[And23] Marc Andreessen. *The Techno-Optimist Manifesto.* `https://a16z.com/the-techno-optimist-manifesto/`. Accessed: 2025-01-17. 2023.

[Arr51] Kenneth J. Arrow. *Social Choice and Individual Values.* 1st. New Haven, CT: Yale University Press, 1951.

[Ben+23] Yoshua Bengio et al. "Managing AI Risks in an Era of Rapid Progress". In: *arXiv preprint arXiv:2310.17688* (2023).

[Bos14] Nick Bostrom. "Superintelligence: Paths, dangers, strategies." In: (2014).

[Bre96] Leo Breiman. "Bagging predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140.

[CL06] Nicolo Cesa-Bianchi and Gábor Lugosi. "Prediction, learning, and games". In: (2006).

[Con85] Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* English translation: Essay on the Application of Analysis to the Probability of Majority Decisions. Paris: Imprimerie Royale, 1785.

[CT06] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2006.

[Doc21] Cory Doctorow. *How to Destroy Surveillance Capitalism.* `https://onezero.medium.com/how-to-destroy-surveillance-capitalism-8135e6744d59`. Accessed: 2025-01-17. 2021.

[Doc23] Cory Doctorow. "Social Quitting". In: *Locus Magazine* (Jan. 2023). Accessed: 2025-01-17.

[Ebe+21] Martin Ebers et al. "The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails)". In: *J* 4.4 (2021), pp. 589–603.

[Gra15] David Graeber. *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy.* Melville House, 2015.

[Guo+17] Chuan Guo et al. "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[Ham+25] Lewis Hammond et al. *Multi-Agent Risks from Advanced AI*. Feb. 19, 2025. DOI: 10.48550/arXiv.2502.14143. arXiv: 2502.14143[cs]. URL: http://arxiv.org/abs/2502.14143 (visited on 02/21/2025).

[Hoe+99] Jennifer A Hoeting et al. "Bayesian model averaging: a tutorial". In: *Statistical Science* 14.4 (1999), pp. 382–417.

[Joe25] Joe Collman, Joe Rogero, William Brewer. *Existing Safety Frameworks Imply Unreasonable Confidence*. https://intelligence.org/2025/04/09/existing-safety-frameworks-imply-unreasonable-confidence/. Accessed: 2025-01-17. Apr. 2025.

[Lau+23] Niklas Lauffer et al. *Who Needs to Know? Minimal Knowledge for Optimal Coordination*. July 13, 2023. DOI: 10.48550/arXiv.2306.09309. arXiv: 2306.09309[cs]. URL: http://arxiv.org/abs/2306.09309 (visited on 02/18/2025).

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6402–6413.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. 2nd. Cambridge, MA: MIT Press, 2018. ISBN: 978-0-262-03940-6.

[NC05] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning* (2005), pp. 625–632.

[Ngo25a] Richard Ngo. *Techno-humanism is techno-optimism for the 21st century*. https://www.mindthefuture.info/p/techno-humanism-is-techno-optimism. Accessed: 2025-01-17. 2025.

[Ngo25b] Richard Ngo. *Towards a scale-free theory of intelligent agency*. https://www.alignmentforum.org/posts/5tYTKX4pNpiG4vzYg/towards-a-scale-free-theory-of-intelligent-agency. Accessed: 2025-01-17. 2025.

[Ngo25c] Richard Ngo. *Well-foundedness as an organizing principle*. https://www.mindthefuture.info/p/well-foundedness-as-an-organizing. Accessed: 2025-01-17. 2025.

[Ova+19] Yaniv Ovadia et al. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13991–14002.

[Rok10] Lior Rokach. "Ensemble-based classifiers". In: *Artificial Intelligence Review* 33.1-2 (2010), pp. 1–39.

[Rus19] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[TPB00] Naftali Tishby, Fernando C Pereira, and William Bialek. "The information bottleneck method". In: *arXiv preprint physics/0004057* (2000).

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory*. New York: John Wiley & Sons, 1998. ISBN: 0-471-03003-1.

[Zho12] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.