The Fundamental Impossibility of Safe Capability Distribution: From Rice's Theorem to Social Coordination Limits

[Your Name Here]
with Claude (Anthropic)
Draft for Discussion

August 3, 2025

Abstract

We establish a fundamental impossibility result connecting computational undecidability to social coordination problems. Building on Rice's theorem and the undecidability of program equivalence, we prove that no computable mechanism can safely distribute powerful capabilities among goal-seeking agents in the general case. This result generalizes Arrow's impossibility theorem and explains why both centralized control and democratization of powerful technologies inevitably produce forms of systemic suffering. The theorem applies to any collection of agents with arbitrary computable utility functions, including humans, artificial intelligences, and hybrid systems.

1 Introduction

The distribution of powerful capabilities—whether computational tools, weapons, information, or other resources—presents a fundamental challenge in the design of social systems. Intuition suggests that either concentrating such capabilities (risking oppression) or distributing them broadly (risking misuse) leads to suboptimal outcomes. We formalize this intuition and prove it represents a fundamental impossibility rooted in computational theory.

Our main contribution is establishing a connection between Rice's theorem on the undecidability of semantic properties of programs and the impossibility of achieving consensus on capability distribution among rational agents. This connection reveals that the difficulties in managing powerful technologies are not merely practical or political, but reflect deep mathematical limits on what any coordination mechanism can achieve.

2 Formal Framework

Definition 1 (Goal-Seeking Agent)**.** *A goal-seeking agent A is a computational process characterized by:*

- *A utility function $U_A : \Omega \to \mathbb{R}$ over possible world states*
- *A capability set $C_A \subseteq C$ representing available actions*
- *A decision procedure $D_A : \Omega \times 2^C \to C$ for selecting actions*

Definition 2 (Capability Distribution Problem)**.** *Given a set of agents* $\{A_1, A_2, \ldots, A_n\}$ *and a set of powerful capabilities* $P \subseteq C$*, find a distribution function* $\delta : P \to 2^{\{A_1,\ldots,A_n\}}$ *that assigns capabilities to agents such that system-wide welfare is maximized.*

Definition 3 (Safe Distribution)**.** *A distribution* δ *is* safe *if it satisfies:*

1. ***Beneficence****: Agents use assigned capabilities to improve overall welfare*
2. ***Non-maleficence****: No agent uses capabilities to deliberately harm others*
3. ***Sustainability****: The distribution remains stable over time*

3 Main Results

Theorem 1 (Fundamental Capability Distribution Impossibility)**.** *For any system of* $n \geq 2$ *goal-seeking agents with arbitrary computable utility functions, no computable mechanism can guarantee a safe distribution of powerful capabilities in the general case.*

Proof Sketch. The proof proceeds by reduction from Rice's theorem:

1. **Undecidability of Agent Behavior**: By Rice's theorem, determining whether two agents will exhibit equivalent behavior when given the same capabilities is undecidable in general.

2. **Impossibility of Trust Verification**: Without the ability to verify behavioral equivalence, no mechanism can distinguish between agents who will use capabilities beneficially versus maleficently.

3. **Temporal Inconsistency**: Even if current agent behavior could be verified, agents may modify their goals or strategies over time, making any safety guarantee temporary.

4. **Computational Intractability**: The space of possible capability-agent interactions grows exponentially, making exhaustive analysis impossible for any realistic system size.

Therefore, any distribution mechanism must operate under fundamental uncertainty about agent intentions and future behaviors, precluding guaranteed safety. □

Corollary 1 (Inevitability of Systemic Suffering)**.** *In any system of goal-seeking agents faced with capability distribution decisions:*

- *Withholding capabilities guarantees constraint-based suffering*
- *Distributing capabilities guarantees misuse-based suffering*
- *No distribution policy can eliminate both forms of suffering*

Corollary 2 (Generalization to Artificial Systems)**.** *The impossibility result applies to any collection of goal-seeking computational processes, including artificial intelligences, economic agents, and hybrid human-AI systems.*

4 Connections to Existing Results

Our theorem generalizes and connects several important impossibility results:

4.1 Arrow's Impossibility Theorem

Arrow showed that no voting system can perfectly aggregate individual preferences into collective decisions. Our result shows that even if such aggregation were possible, the undecidability of program behavior would prevent safe implementation of collective decisions regarding capability distribution.

4.2 Tragedy of the Commons

The classic tragedy emerges when individual rationality leads to collective irrationality. Our theorem explains why this tragedy is mathematically inevitable: any mechanism for preventing it requires solving undecidable problems about agent behavior.

4.3 AI Alignment Problem

The difficulty in ensuring AI systems pursue intended goals reflects our more general result. Even with perfect goal specification, Rice's theorem implies we cannot verify that an AI system actually implements those goals without running it—potentially causing the very harms we seek to prevent.

5 Implications

5.1 Technology Policy

Our result suggests that debates over technology regulation, from AI governance to platform moderation, face fundamental rather than merely practical limitations. No regulatory framework can guarantee safe outcomes while preserving beneficial uses.

5.2 AI Safety Research

The undecidability connection implies that AI safety cannot be solved through formal verification alone. Safety measures must account for the fundamental uncertainty about system behavior in novel situations.

5.3 Economic Theory

The theorem provides theoretical grounding for observed market failures in technology adoption and suggests why both laissez-faire and centrally planned approaches to capability distribution exhibit systematic problems.

5.4 Social Philosophy

The result formalizes intuitions about the tragic nature of social coordination problems and suggests that utopian solutions to capability distribution are not merely difficult but mathematically impossible.

6 Open Questions and Future Work

Several important questions emerge from this work:

1. **Approximation Algorithms**: While perfect safety is impossible, what approximation bounds can be achieved for specific classes of agents or capabilities?

2. **Probabilistic Guarantees**: Can weaker probabilistic safety guarantees be provided under reasonable assumptions about agent behavior distributions?

3. **Mechanism Design**: What properties should practical capability distribution mechanisms optimize for, given that perfect safety is unachievable?

4. **Empirical Validation**: How do real-world capability distribution systems perform relative to the theoretical limits established here?

5. **Recursive Applications**: How does the impossibility result apply to systems that must distribute capabilities for managing capability distribution (meta-governance)?

7 Conclusion

We have established that the safe distribution of powerful capabilities among goal-seeking agents faces fundamental computational limits rooted in Rice's theorem and the undecidability of program equivalence. This result explains why both centralized control and broad democratization of powerful technologies inevitably produce systemic suffering, and why this tragedy extends beyond human systems to any collection of goal-seeking computational processes.

The theorem does not counsel despair but rather realistic expectations about what social coordination mechanisms can achieve. Understanding these fundamental limits may guide the development of more robust, adaptive approaches to capability distribution that acknowledge uncertainty and work within rather than against mathematical constraints.

As artificial intelligence systems become more capable and autonomous, these theoretical limits become increasingly practical concerns. The impossibility of guaranteed safe capability distribution among computational agents may be one of the most important constraints on the development of advanced AI systems and the societies that deploy them.

Acknowledgments

This work emerged from discussions about software packaging and the democratization of computational tools. We thank the open source software community for providing concrete examples of both the benefits and risks of capability distribution at scale.