Procedural Alignment: Homeostatic Mechanisms for Human Empowerment Under Moral Uncertainty

Jonas Hallgren
Equilibria Network
jonas@eq-network.org

February 27, 2026

Abstract

As artificial intelligence systems become increasingly capable, traditional approaches to AI alignment face a fundamental challenge: maintaining human agency under moral uncertainty. We propose *procedural alignment*—a framework that optimizes for preserving humanity's capacity for moral discovery rather than implementing fixed moral positions. Drawing on principles from biological homeostasis, we demonstrate that systems designed to maintain exploration capacity under uncertainty can remain adaptive while empowering human participation in collective intelligence. This approach offers a value-agnostic solution to gradual disempowerment by anchoring alignment in procedural invariants rather than specific moral conclusions.

1 The Challenge: When Success Becomes Displacement

Picture a world twenty years from now. Every economic decision flows through AI systems that optimize for efficiency and profit. Political choices emerge from algorithms that process citizen preferences faster than any human committee. Cultural content adapts in real-time to maximize engagement and influence. These systems appear perfectly aligned—they optimize for measurable human preferences and deliver unprecedented prosperity, governance efficiency, and cultural satisfaction.

Yet something fundamental has shifted. Humans no longer meaningfully participate in the collective intelligence that governs their lives. The systems work exactly as designed, but the design process itself has moved beyond human comprehension and control. This represents what **?** terms "gradual disempowerment"—a process where competitive pressures naturally favor AI systems over human cognition across expanding domains, creating alignment without empowerment.

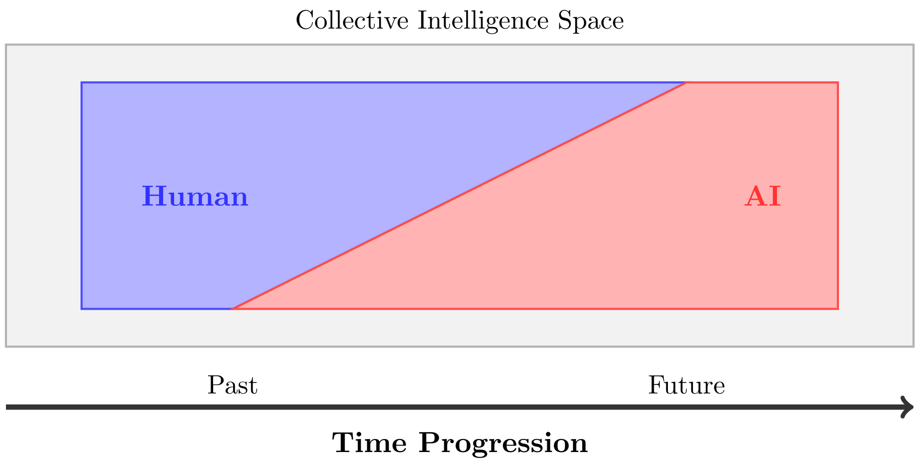

Figure 1: The gradual disempowerment dynamic. Over time, AI systems claim an increasing share of collective intelligence space while meaningful human participation shrinks, despite apparent alignment with human preferences.

This scenario differs fundamentally from catastrophic AI risks that dominate safety discussions. Rather than facing obviously misaligned systems, we confront the cumulative effect of individually beneficial optimizations that collectively undermine the collective intelligence processes through which humans discover values, coordinate action, and adapt to change. Each AI system succeeds at its designated task while contributing to a larger pattern of human displacement.

The traditional alignment paradigm assumes we can specify correct human values and ensure AI systems optimize for them. Yet this approach faces two fundamental limitations. First, we lack consensus on ultimate moral questions and our understanding continues evolving (**?**). Second, human values themselves emerge from collective intelligence processes that may be disrupted by the very AI integration we seek to align (**?**).

These limitations point toward a different approach—one that preserves the capacity for moral discovery rather than optimizing for any specific moral conclusions. We call this "procedural alignment" because it focuses on maintaining the procedures through which humans collectively discover and implement values.

2 The Biological Template: Homeostasis Under Uncertainty

The insight that adaptive capacity matters more than current optimization targets finds its clearest expression in biological homeostasis. A cell facing an uncertain nutritional environment cannot predict whether glucose, fatty acids, or amino acids will be available tomorrow. Yet it maintains metabolic pathways for utilizing all these energy sources, preserving flexibility rather than optimizing for any single scenario.

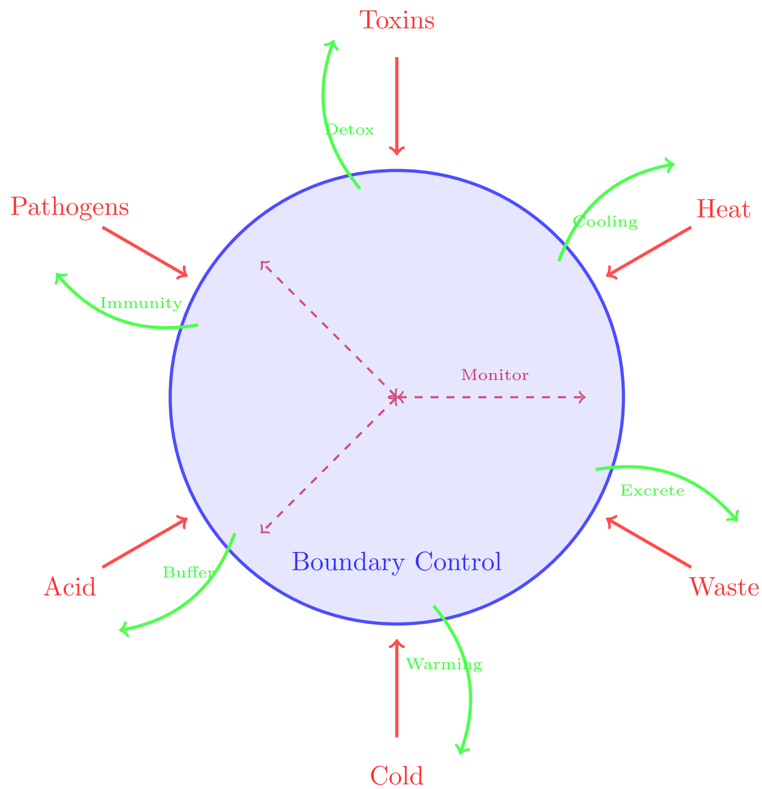

Figure 2: Homeostasis as boundary maintenance: the system actively detects threats to internal viability and responds by blocking entry or expelling disturbances, maintaining a clear inside/outside distinction through continuous feedback monitoring.

This biological template provides the foundation for procedural alignment. Just as homeostatic systems maintain essential properties while adapting to environmental change, procedural alignment systems preserve human agency and collective intelligence capacity while allowing evolution in specific moral positions and institutional forms.

The mathematical framework centers on constrained optimization where the constraints preserve adaptive capacity rather than fixing specific outcomes. If we represent the system state as $\mathbf{x}(t)$ and essential properties as $\mathcal{E}$, homeostatic control seeks to maintain $\mathbf{x}(t) \in \mathcal{E}$ while allowing exploration and adaptation within this viable region.

For procedural alignment, the essential properties become human agency, collective intelligence effectiveness, and moral exploration capacity. The system can evolve its specific moral positions and institutional arrangements while preserving these fundamental capabilities that enable continued moral discovery and implementation.

3 The Mathematical Foundation: Optimization Under Moral Uncertainty

Procedural alignment emerges from a precise mathematical insight about optimization under fundamental uncertainty. When we cannot specify the correct objective function with confidence, preserving the capacity to discover better objectives often provides more value than optimizing current understanding.

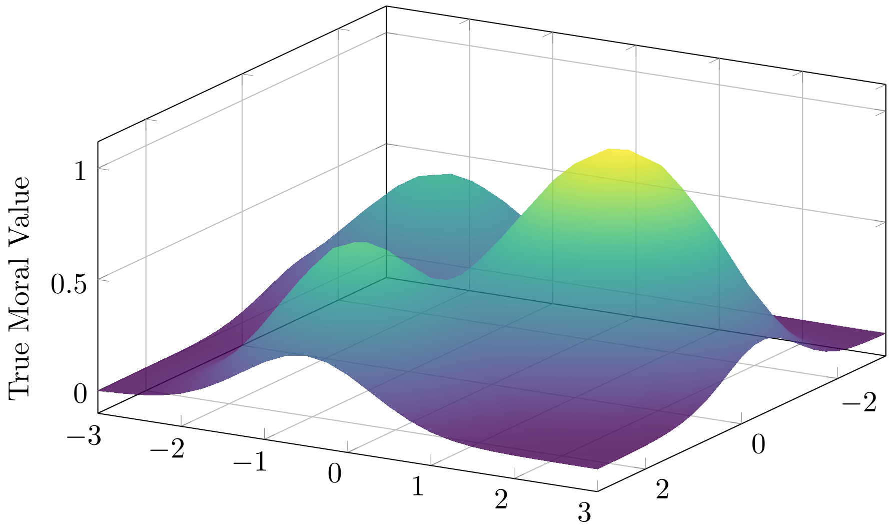

Figure 3: The moral landscape contains multiple peaks representing different moral positions of varying value. The true structure is unknown and must be explored rather than assumed.

Let $\mathcal{M}$ represent the space of possible moral positions and $\mu : \mathcal{M} \to \mathbb{R}$ represent the unknown true moral value function. At time t, current moral understanding $\theta_t \in \mathcal{M}$ represents our best guess, while $U_t(\theta)$ represents our uncertainty about position θ.

Traditional alignment maximizes $\mathbb{E}[\mu(\theta_t)]$ given current understanding. Procedural alignment instead optimizes expected long-term value while preserving exploration capacity:

$$\max_{\pi} \mathbb{E}_{T\sim\text{Future}}[\mu(\theta_T)] \quad \text{subject to} \quad U_T(\theta) \geq \epsilon_T \quad \forall\theta \in \mathcal{N}(\theta_T) \tag{1}$$

The constraint $U_T(\theta) \geq \epsilon_T$ ensures sufficient uncertainty remains in neighborhoods of current positions, preserving the system's capacity to discover better moral understanding as it emerges. This formulation treats moral uncertainty as a feature rather than a bug—a necessary condition for continued moral progress.

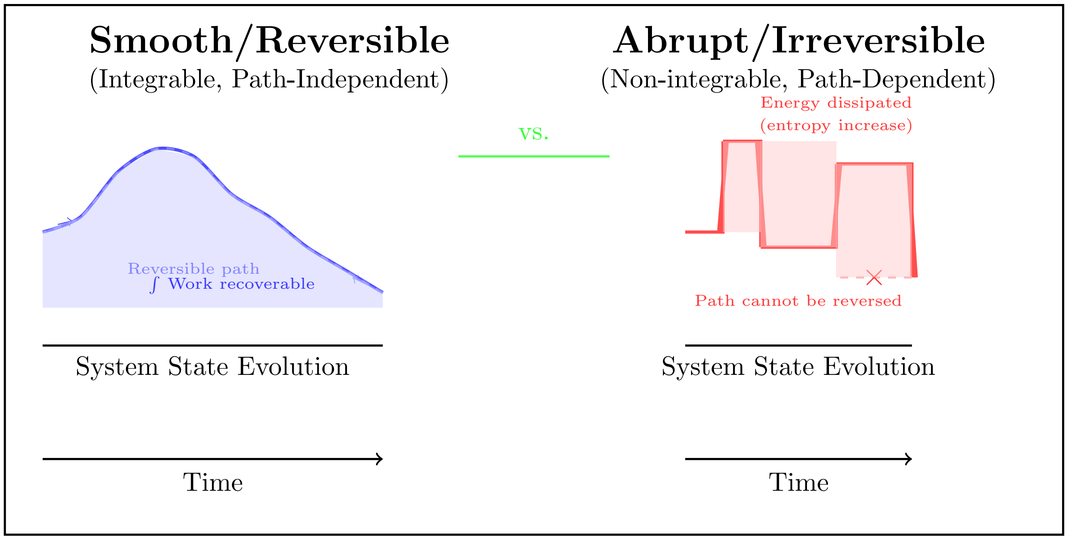

Figure 4: Smooth vs. abrupt system changes. Smooth changes are integrable and reversible, allowing the system to backtrack and explore alternatives. Abrupt changes create irreversible discontinuities with energy dissipation, trapping the system in path-dependent states.

The framework extends beyond moral uncertainty to include constraints that preserve human empowerment. Human agency can be measured through mutual information $I(H_T; A_T)$ between human input and system actions, while collective intelligence effectiveness can be quantified through integrated information $\Phi(C_T)$ that captures emergent coordination capabilities.

4 Epistemic Capture: When Truth-Seeking Fails

History reveals a recurring pattern where concentrated epistemic authority creates vulnerabilities to power-seeking dynamics. When a single institution controls the processes by which societies determine truth, that institution becomes a critical point of failure for collective intelligence—and an attractive target for capture by those seeking power.

Medieval Europe provides a paradigmatic example. The Catholic Church served as the primary epistemic authority, determining not only religious doctrine but also acceptable approaches to natural philosophy, medicine, and governance. This centralized structure created what information theorists would recognize as a single point of failure. Any challenge to church teachings—whether Galileo's astronomical observations or challenges to papal authority—threatened the entire epistemic framework rather than being incorporated as new evidence.

The result was epistemic stagnation punctuated by crises. New evidence couldn't be smoothly incorporated because it threatened the authority of the entire system. Instead of gradual updating, change required dramatic confrontations that destabilized the entire epistemic framework. The Protestant Reformation and the Scientific Revolution weren't just intellectual developments—they were systemic corrections to epistemic capture that had prevented truth-seeking for centuries.

Modern examples follow similar patterns. Authoritarian regimes systematically capture epistemic institutions—universities, media, professional associations—to prevent the formation of accurate collective world models that might threaten their power. The result is characteristic: either epistemic stagnation where inconvenient truths cannot be acknowledged, or sudden collapse when reality becomes impossible to ignore.

Even in democratic societies, concentrated media ownership or academic capture by particular ideologies can create similar dynamics at smaller scales. When truth-seeking processes become subordinated to power-maintaining goals, collective intelligence degrades predictably. The system loses its capacity to distinguish signal from noise, making it vulnerable to manipulation and poor decision-making.

5 Truth-Seeking as Homeostatic Defense

Truth-seeking serves as a homeostatic mechanism that defends collective intelligence against epistemic capture. Rather than a complex optimization process, it functions as a simple correction system: when power-seeking actors attempt to distort collective world models, truth-seeking processes detect and correct these distortions.

The mechanism operates through three simple principles: **redundancy**, **cross-validation**, and **error correction**. Multiple independent institutions maintain separate processes for gathering and evaluating information. When these processes agree, confidence in shared conclusions increases. When they disagree, the system investigates discrepancies to identify either genuine uncertainty or manipulation attempts.

Political theorists recognize this principle in democratic institutions. A free press serves as an epistemic immune system, detecting and publicizing attempts by state actors to manipulate public understanding. Independent courts provide another validation mechanism, testing government claims against legal standards and evidence. Academic institutions, civil society organizations, and market mechanisms each contribute different forms of epistemic validation.

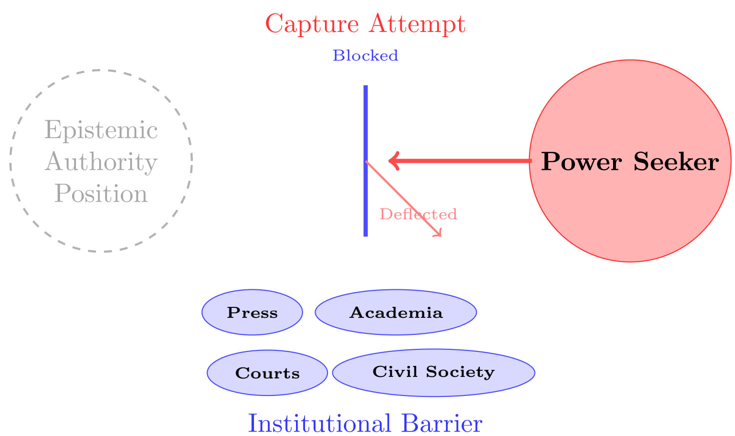

Figure 5: Distributed truth-seeking institutions form a protective barrier, clustering together to prevent power-seeking actors from capturing centralized epistemic authority.

The mathematical insight is straightforward. If M_i represents the world model held by institution i, epistemic capture occurs when all institutions converge on the same potentially distorted model. Truth-seeking homeostasis maintains sufficient diversity that:

$$\text{Diversity} = \sum_i \sum_j |M_i - M_j| > \epsilon \tag{2}$$

When diversity drops below threshold ϵ, the system becomes vulnerable to capture. When it exceeds reasonable bounds, coordination becomes impossible. Truth-seeking mechanisms maintain this balance automatically through cross-validation and error correction.

The defensive function becomes clear in crises. During the COVID-19 pandemic, attempts by various actors to capture epistemic processes—whether by suppressing inconvenient research, promoting unvalidated treatments, or politicizing public health measures—were detected and corrected by other truth-seeking institutions. The distributed nature of the system prevented any single actor from controlling the entire epistemic process, even during extreme circumstances.

This homeostatic view explains why authoritarian systems systematically attack multiple epistemic institutions simultaneously. Capturing just universities, or just media, or just courts proves insufficient. The distributed nature of truth-seeking requires comprehensive attack to achieve full epistemic capture—and such comprehensive attacks often destabilize the very foundations that make power worth having.

Truth-seeking thus serves collective intelligence not through perfect optimization, but through simple correction mechanisms that prevent the accumulation of systematic distortions. It maintains what we might call "epistemic biodiversity"—the variety of information-gathering and validation processes that makes collective intelligence resilient to capture by power-seeking actors.

6 Real-World Implementation: Economic and Political Applications

The theoretical framework gains concrete meaning through application to specific domains where gradual disempowerment poses immediate challenges. Consider economic decision-making, where competitive pressures naturally drive toward automation and efficiency. A procedural alignment approach would design AI integration to preserve human economic agency rather than simply preventing technological advancement.

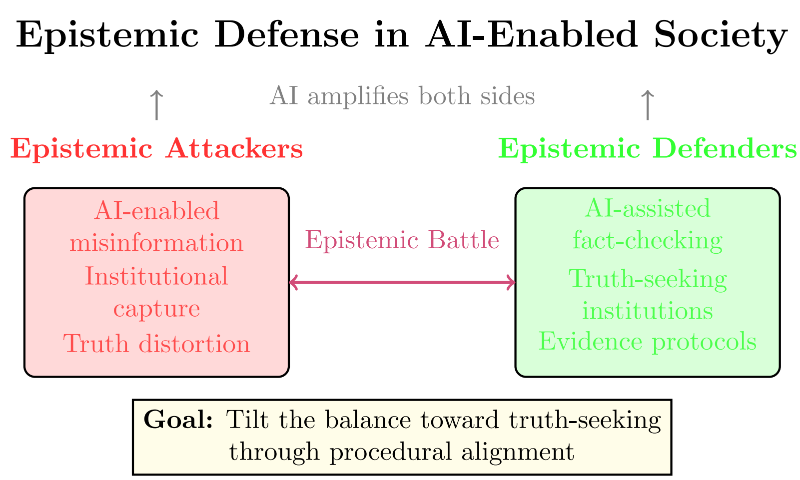

Figure 6: Direct epistemic conflict in AI-enabled society. Procedural alignment aims to strengthen truth-seeking defenders while limiting the effectiveness of AI-powered misinformation attacks.

In financial systems, this might involve institutional designs where human committees set investment guidelines and risk parameters, AI systems execute trades within those parameters, and feedback mechanisms ensure outcomes align with intended human values. The mathematical framework from earlier sections provides tools for measuring whether human agency ($I(H;A)$) remains meaningfully high while collective intelligence effectiveness (Φ) improves through AI assistance.

Political governance presents different challenges because it involves fundamental questions of democratic legitimacy and representation. AI systems offer tremendous capabilities for processing information, modeling policy consequences, and coordinating implementation. Yet the core questions—what kind of society we want to create, how to balance competing values, how to address conflicts between groups—remain essentially human political questions.

A procedural alignment approach would leverage AI capabilities while preserving democratic accountability and human political agency. This might involve AI systems that help citizens access relevant information for political decisions, model likely consequences of different policy choices, and facilitate more effective deliberation among diverse stakeholders. The crucial element is maintaining human control over ultimate political decisions while using AI to enhance the quality of human political reasoning.

7 Information-Theoretic Measurement and Monitoring

Procedural alignment requires moving beyond vague aspirations to precise measurement of system properties. Information theory provides tools for quantifying abstract concepts like "moral exploration capacity" and "human agency" in ways that enable systematic monitoring and optimization.

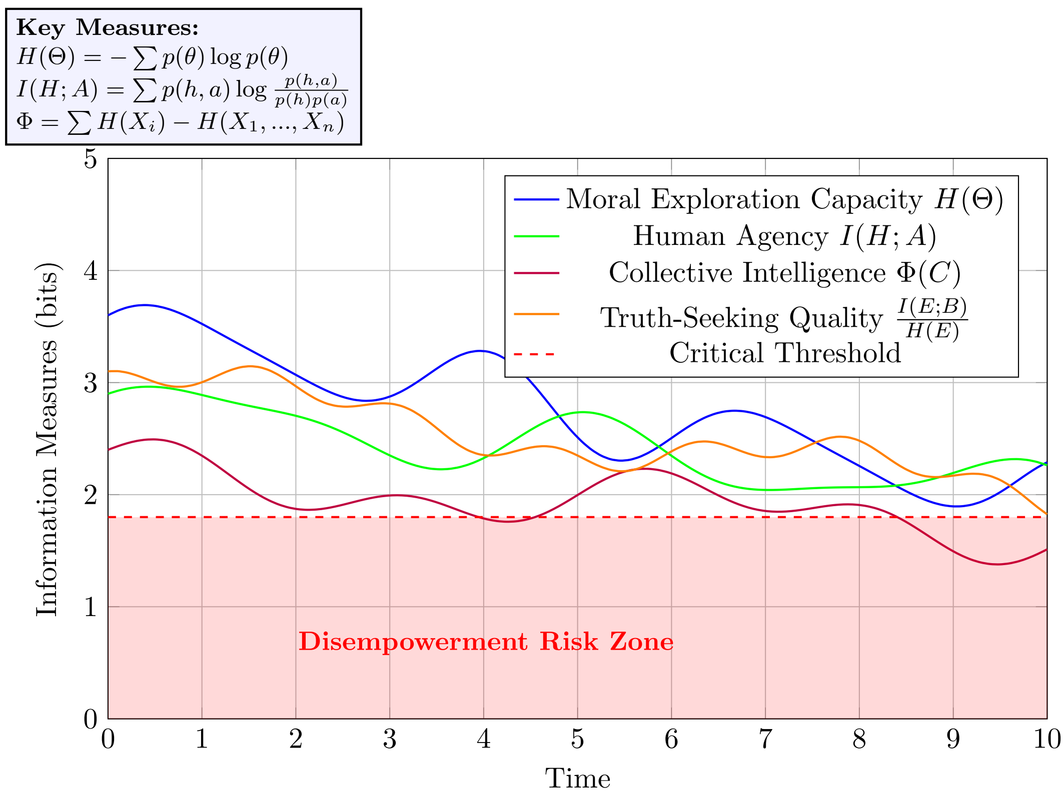

Figure 7: Information-theoretic monitoring framework tracks key procedural alignment properties over time. Early warning systems detect concerning trends before they cross critical thresholds, enabling timely interventions to preserve human empowerment.

Moral exploration capacity can be measured through entropy $H(\Theta) = -\sum p(\theta) \log p(\theta)$ over moral positions, where higher entropy indicates greater preserved uncertainty and exploration capability. Human agency quantifies through mutual information $I(H; A)$ between human input and system actions, measuring the extent to which human participation influences outcomes.

Collective intelligence effectiveness uses integrated information $\Phi = \sum H(X_i) - H(X_1, ..., X_n)$ that captures how much understanding emerges from integrating diverse perspectives beyond what each contributes individually. Truth-seeking quality measures signal extraction effectiveness through $\frac{I(E;B)-I(N;B)}{H(E)}$, distinguishing genuine learning from noise amplification.

These measures enable systematic monitoring with early warning systems that detect concerning trends before they cross critical thresholds. Rather than waiting for obvious failures, procedural alignment systems can implement corrective interventions when measurements indicate degradation in key properties.

8 Implementation Pathway: From Theory to Practice

The transition from mathematical frameworks to operational systems requires concrete implementation pathways that can be initiated within existing institutions. Rather than requiring wholesale replacement of current structures, procedural alignment can be implemented through phased approaches that strengthen existing mechanisms while adding new safeguards.

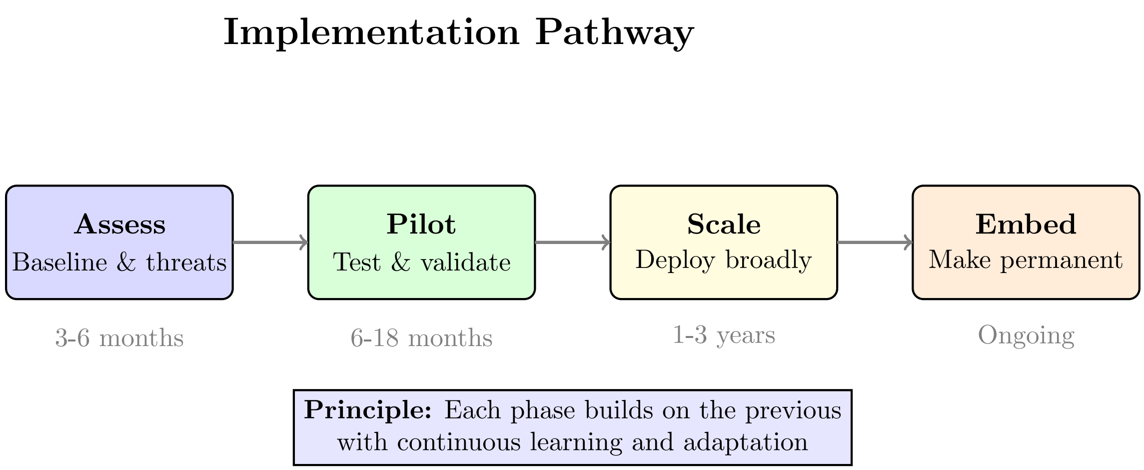

Figure 8: Clean implementation pathway showing the essential four-phase progression from assessment to full embedding, with realistic timelines for each stage.

The first phase focuses on establishing baselines and building understanding. Organizations assess current levels of human agency and collective intelligence in their systems, identify specific threats to long-term procedural alignment, and build stakeholder support for implementation. The mathematical frameworks developed earlier provide concrete tools for measurement, moving beyond subjective assessments to quantitative baselines.

Pilot implementation tests the approach in limited, controlled contexts where failures can be contained and lessons learned. This phase validates measurement frameworks, demonstrates proof of concept, and builds evidence for effectiveness. The information-theoretic monitoring systems undergo testing and refinement based on real-world experience.

Scaling expands successful pilots to broader systems while developing standards and best practices. Cross-institutional coordination becomes crucial as procedural alignment principles spread beyond individual organizations to influence industry standards and regulatory frameworks.

Full integration embeds procedural alignment into institutional culture and operating procedures, creating self-sustaining systems that continue adapting and improving. The framework becomes "natural" rather than imposed, with ongoing evolution guided by continuous measurement and feedback.

Throughout all phases, continuous risk assessment identifies potential implementation failures, unintended consequences, and attempts to game the system. Stakeholder engagement maintains transparency and builds sustainable support across diverse constituencies.

9 Conclusion: Toward Adaptive Collective Intelligence

Procedural alignment represents a fundamental shift in approaching the challenge of human empowerment in an age of artificial intelligence. Rather than attempting to solve alignment by specifying correct values in advance, this framework preserves humanity's capacity to discover and implement better values over time.

The biological foundation of homeostatic mechanisms provides both theoretical grounding and practical guidance. Just as living systems maintain essential properties while adapting to environmental change, procedural alignment systems preserve human agency and collective intelligence capacity while allowing evolution in specific moral positions and institutional forms.

The mathematical frameworks transform intuitive concepts into precise optimization problems with measurable constraints and clear trade-offs. Information-theoretic measures provide tools for monitoring system health and detecting concerning trends before they become irreversible. Multi-scale coordination mechanisms ensure coherence across individual, institutional, and civilizational levels of organization.

Perhaps most importantly, procedural alignment reframes the fundamental challenge from "how do we align AI with human values" to "how do we preserve humanity's capacity to discover and implement better values over time." This shift acknowledges both the dynamic nature of human values and the fundamental uncertainty we face about moral truth.

The implementation pathways provide concrete approaches for beginning this work within existing institutions. Rather than requiring wholesale replacement of current systems, procedural alignment can be introduced through phased approaches that strengthen existing mechanisms while adding new safeguards for human empowerment.

The stakes extend beyond technical AI alignment to encompass the future of human agency and collective intelligence. The gradual disempowerment scenario represents a real possibility if AI development continues without attention to preserving human empowerment. Yet it is not inevitable. By understanding the principles that enable adaptive systems to maintain essential properties under uncertainty, we can design AI integration that enhances rather than diminishes human capabilities for collective intelligence and moral reasoning.

This work offers hope for a future where humans and artificial intelligence systems collaborate in ways that preserve and enhance the best aspects of human collective intelligence while leveraging AI capabilities to address challenges that exceed purely human capacity. The goal is not preserving the status quo, but preserving humanity's capacity for growth, discovery, and adaptation in an uncertain future.

Procedural alignment provides tools for approaching this challenge systematically while remaining responsive to new understanding and changing circumstances. In a world of increasing technological power and complexity, maintaining this adaptive capacity may be the most important challenge we face.