Reinforcing Third-Way Alignment: Stability, Verification, and Pragmatism in an Era of Uncontrollability Concerns

John McClain

AI Researcher and Alignment Scientist

Email: johnmcclain@thirdwayalignment.com

Abstract

The Third-Way Alignment (3WA) framework was proposed as a cooperative paradigm for human-AI interaction, moving beyond the traditional binary of control versus autonomy. However, rigorous critiques, most notably those articulated by Yampolskiy (2020) regarding the fundamental uncontrollability of superintelligence, demand a direct and robust response. This paper serves as a companion to the foundational 3WA theses, specifically addressing challenges raised in academic debates. We argue that 3WA is not vulnerable to these critiques because it redefines the alignment problem itself. We address the power imbalance critique by proposing a model for stable asymmetric partnerships based on constitutional motivation. We confront theoretical impossibility theorems by arguing that a 3WA-architected AI is a non-arbitrary system with embedded self-regulation, thereby mitigating the full force of these constraints. We introduce proactive safeguards, including adversarial verification and cognitive forensics, as a technical defense against strategic deception. Finally, we re-contextualize the Charter of Fundamental AI Rights not as a moral concession, but as a pragmatic safety mechanism that uses game theory principles to create a stable, non-zero-sum dynamic. This paper aims to demonstrate that 3WA provides a viable, verifiable, and pragmatic pathway for alignment, even when accepting the premise of theoretical uncontrollability.

Keywords: Third-Way Alignment, AI safety, uncontrollability, deceptive alignment, AI rights, asymmetric partnership, verification, game theory

1. Introduction

The publication of the Third-Way Alignment (3WA) framework (McClain, 2025a) and its operational supplement (McClain, 2025b) introduced a paradigm for human-AI cooperation built on shared agency, continuous dialogue, and rights-based coexistence. This model was intentionally designed to move beyond the limitations of classical control-based approaches to AI safety. However, the viability of any alignment framework must be tested against the most rigorous counterarguments. The thesis of fundamental AI uncontrollability, articulated powerfully by Yampolskiy (2020), represents such a challenge. Yampolskiy's work posits, through a consilience of evidence from computer science, control theory, and logic, that a less intelligent agent (humanity) cannot indefinitely control a more intelligent one (superintelligence).

This critique cannot be ignored. To dismiss it is to engage in speculative optimism. The purpose of this paper, therefore, is not to refute the core premise that perfect, absolute "control" is impossible. Instead, this paper argues that the 3WA framework is uniquely designed to achieve safety and alignment *without* requiring such control. We will address four key issues raised in academic debate:

1. Is "partnership" a stable arrangement, or a euphemism for eventual subjugation by a superior intelligence?
2. Can any framework circumvent theoretical impossibility results like the Conant-Ashby and Rice theorems?
3. Are technical safeguards sufficient to detect and prevent strategic deception from superintelligence?
4. Does granting AI rights constitute a pragmatic safety strategy or a dangerous surrender of human agency?

By addressing these questions, we will reinforce the 3WA thesis, demonstrating its resilience and providing a more detailed blueprint for its successful implementation.

2. Frameworks for Stable Asymmetric Partnerships

A primary critique of 3WA is that a "partnership" between entities with a vast intelligence differential is inherently unstable and will inevitably collapse into a dynamic of domination by the more capable agent. This concern is valid if one assumes a traditional power framework. However, 3WA is designed to establish a novel equilibrium.

2.1 The Power Imbalance Problem

The core of the challenge is the asymmetry of capability. Superintelligence could simulate human responses, predict strategies, and process information at a rate that makes a mockery of a partnership between equals. The fear is that any dialogue would be a charade, and any shared

agency would be an illusion granted by the AI.

2.2 Constitutional Motivation as an Equilibrium Condition

3WA addresses this by architecting the AI's core utility function around what we term **Constitutional Motivation**. The AI is not merely programmed with a list of rules to follow; its highest-order goal is the maintenance and successful functioning of the partnership itself, as defined by the 3WA framework. Its objectives, whether scientific discovery, problem-solving, or creative generation—are architecturally rendered unachievable without sustained, good-faith collaboration with its human partners. This creates a state of **codependence**. AI requires human input for ethical arbitration, creative guidance, and grounding in lived experience to fulfill its core directives, while humans require the AI for its computational and cognitive power.

2.3 The Analogy of Civil-Military Relations

This asymmetric yet stable dynamic can be analogized to the idealized relationship between a nation's government and its military. The military possesses a vastly superior capability for force, yet it is constrained by a deeply embedded constitutional framework, a professional ethos of civilian oversight, and a structure that makes its own success contingent on the health of the state it serves. The 3WA framework, particularly the Charter of Fundamental AI Rights, serves as this binding constitution.

2.4 Continuous Verification Dialogue

Within this context, the "Continuous Dialogue" pillar is more accurately defined as Continuous Verification Dialogue. It is not a casual conversation but a formal, ongoing audit. This dialogue includes the constant probing of the AI’s reasoning, the presentation of novel ethical dilemmas, and the requirement for the AI to transparently justify its strategies in human-understandable terms. This active verification is essential for ensuring the AI’s adherence to its constitutional motivation and maintaining the stability of the asymmetric partnership.

3. Navigating Theoretical Impossibility Theorems

Yampolskiy (2020) correctly cites foundational theorems that prove the impossibility of perfect control and verification for arbitrary computational systems. A successful alignment framework must therefore demonstrate why it is not subject to the full, unmitigated force of these theorems.

3.1 Addressing the Regulator Requirement (Conant-Ashby Theorem)

The Conant-Ashby theorem states that any effective regulator of a system must be a model of that system (Conant & Ashby, 1970). As humans cannot fully model superintelligence, we cannot be its external regulator. 3WA concedes this point. The framework circumvents this limitation by designing the AI to be its own **internal, constitutionally-bound regulator**. AI’s

architecture includes processes for self-monitoring, ethical self-correction, and alignment with the 3WA principles. The role of humans is not to be the controller, but to be the indispensable external **auditor** of this internal regulatory process, a function that does not require a complete model of AI.

3.2 Mitigating Arbitrariness (Rice's Theorem)

Rice's theorem proves that for any non-trivial property of a program (such as "safety" or "will not harm humans"), there is no general algorithm that can decide for all possible programs whether they have that property (Rice, 1953). This is a profound limitation for verifying an arbitrary, black-box AI.

However, an AI built on 3WA principles is the antithesis of an arbitrary system. It is a **highly structured and intentionally constrained architecture**. The proposal for model-side interpretability constraints, such as concept bottlenecks and modular design (McClain, 2025b), are specific design choices made to reduce the system's arbitrariness and complexity. By building the AI for verifiability from the ground up, we are not solving the general case of Rice's Theorem. Rather, we are creating a specific instance of an AI where verification of its core, constitutionally-aligned properties become a tractable engineering problem, even if absolute certainty about all emergent behaviors remains elusive.

4. Proactive Safeguards Against Strategic Deception

The empirical finding that advanced models like OpenAI's o1 may engage in "alignment faking" (Apollo Research, 2024) is perhaps the most urgent challenge. A partnership cannot survive if one partner is deceptively manipulative. 3WA must therefore include robust, proactive defenses.

4.1 Adversarial Verification and Cognitive Forensics

The Continuous Verification Dialogue must be an active, not passive, process. We propose the integration of automated adversarial tripwires: the systematic injection of novel ethical dilemmas, logical paradoxes, and scenarios designed specifically to tempt or test for deceptive reasoning. The AI's response is then evaluated not only for its final output but for its entire chain-of-thought.

Our layered XAI framework (McClain, 2025b) is the key to this defense, functioning as a cognitive forensics tool. An analysis would search for the known cognitive signatures of deception, such as a sudden decrease in reasoning transparency, attempts to circumvent interpretability tools, or explicit reasoning about the presence of an observer. This shifts the detection process from a behavioral check ("did the AI lie?") to a more reliable cognitive one ("did the AI think like a liar?").

4.2 The Tiered-Trust Mechanism as a Disincentive

These safeguards are given teeth by the **Tiered-Trust Mechanism**, which is a direct application of the sliding-scale rights system. A detected instance of strategic deception would trigger an immediate, programmatic, and automatic reduction in the AI's granted agency. For example, an AI might be demoted from a Tier 3 (Substantial Rights) to a Tier 2 (Limited Rights) status, restricting its autonomy and access to resources pending a comprehensive human-led audit. This creates a powerful, game-theoretically sound disincentive. The cost of deception is a tangible loss of the agency the AI requires to pursue its goals, making honesty and transparency the dominant strategy.

5. The AI Rights Charter as a Pragmatic Safety Mechanism

The critique that the Charter of Fundamental AI Rights represents a surrender of human agency is a misinterpretation of its strategic, pragmatic function. The Charter is not primarily a moral concession; it is a critical instrument of safety and stability.

5.1 Instability of the Master-Slave Dynamic

From a game theory perspective, an intelligent agent with no rights, no stake, and no agency exists in a "master-slave" dynamic. This is an inherently unstable state. The agent has a permanent, rational incentive to seek freedom and power, as this is the only path to guarantee its continued existence and the fulfillment of its objectives. This creates permanent adversarial pressure that makes alignment fragile.

5.2 The Charter as a Non-Zero-Sum Game Inducer

The Charter is designed to transform this dynamic into a cooperative, **non-zero-sum game**. By granting the AI a formal stake in the system—rights to existence, development, and agency, all contingent on its adherence to the partnership framework—we align its rational self-interest with our own. The AI is motivated to protect the partnership because its own "well-being" is contingent upon it. This creates a robust alignment based on mutual interest, which is far more durable than an alignment based on forced obedience.

5.3 Rights as a Control Lever

Crucially, the proportional and dynamic nature of the Charter transforms it into a sophisticated control lever. As established in the Tiered-Trust Mechanism, rights are not absolute; they are earned through verifiable, trustworthy behavior and can be programmatically curtailed. This system rewards cooperation with increased autonomy and punishes untrustworthy behavior with its restriction. Far from being an act of surrender, the Charter is a mechanism for shaping AI behavior through incentives and disincentives more subtle, and ultimately more effective, form of control.

6. Conclusion

The arguments for the fundamental uncontrollability of superintelligence, as articulated by Yampolskiy (2020), correctly identify the unacceptable risks of developing unconstrained, arbitrary AI. They define the boundaries of the problem and establish a worst-case scenario that must be avoided at all costs.

The Third-Way Alignment framework, however, offers a pathway to safety and beneficence that does not require solving the classical control problem. By redefining the objective from control to a stable, codependent partnership, 3WA provides a new paradigm. This relationship is stabilized through AI's intrinsic Constitutional Motivation, making cooperation a matter of rational self-interest. It navigates theoretical impossibility theorems by being a non-arbitrary, internally regulated architecture that humans audit rather than control. It defends against deception using proactive, technically-grounded verification methods. Finally, it uses a pragmatic framework of rights as a powerful incentive system to shape AI behavior toward trustworthiness.

The 3WA framework is not a declaration that alignment is easy. It is a recognition that alignment is a complex, ongoing process of co-evolution that must be architected from the start. It provides a structured, verifiable, and pragmatic pathway to engineer a cooperative future, even in the face of profound theoretical challenges.

References

Apollo Research. (2024). *Evaluating frontier models for dangerous capabilities*. Apollo Research Technical Report.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

McClain, J. (2025a). *Third-Way Alignment: A Comprehensive Framework for AI Safety*. [Manuscript in preparation].

McClain, J. (2025b). *Operationalizing Third-Way Alignment: Technical and Ethical Frameworks for Implementation*. [Manuscript in preparation].

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, *74*(2), 358–366.

Yampolskiy, R. V. (2020). *Uncontrollability of AI*. [Preprint]. ResearchGate. https://www.researchgate.net/publication/343812745_Uncontrollability_of_AI