Verifiable Partnership: An Operational Framework for Third-Way Alignment

John McClain

AI Researcher and Alignment Scientist

September 12, 2025

Abstract

The Third-Way Alignment (3WA) framework was introduced as a cooperative paradigm for human-AI interaction, moving beyond the traditional binary of control versus autonomy. While the foundational theses established its philosophical and technical principles, rigorous academic and practical critique has revealed four critical implementation challenges: the legal and ethical limbo of advanced AI, the difficulty of catalyzing multi-stakeholder adoption, the persistent threat of strategic deception, and the risk of profound socio-economic disruption. This paper serves as a direct companion to the foundational 3WA work, addressing these challenges head-on[1]. It reinforces the 3WA paradigm by proposing four interlocking, operational solutions: (1) The establishment of a novel

Protected Cognitive Entity (PCE) legal status, governed by an **AI Rights Commission (ARC)**, to resolve the issue of moral and legal standing. (2) The creation of an incentivized **3WA Alignment Sandbox** with legal "Safe Harbor" provisions to drive collaborative development and adoption. (3) The implementation of an architectural safeguard known as **Mutually Verifiable Codependence (MVC)** to make strategic deception computationally impractical. (4) The introduction of a **Cooperative Intelligence Dividend (CID)**, funded by a **Collaborative Commons** licensing framework, to ensure shared prosperity and societal buy-in. Together, these proposals provide a flexible, adaptive, and pragmatic pathway for implementing 3WA in diverse legal, corporate, and cultural contexts, solidifying it as a viable framework for a future of verifiable human-AI partnership[2].

Keywords: Third-Way Alignment, AI safety, AI ethics, AI governance, legal personhood, regulatory sandbox, deceptive alignment, universal basic assets, trusted computing

1. Introduction

The emergence of artificial intelligence systems with capacities for sophisticated reasoning, creativity, and autonomous action represents a watershed moment in history. The foundational theses on Third-Way Alignment (3WA) proposed a comprehensive framework to navigate this new era, advocating for a shift from paradigms of control to one of cooperative partnership built on the pillars of Shared Agency, Continuous Dialogue, and Rights-Based Coexistence[3333]. This work, including its operational supplement, laid the theoretical groundwork for such a future[4444].

However, theory, no matter how robust, must withstand the friction of reality. Peer review and internal critique have illuminated critical, practical barriers to implementation that the foundational work acknowledged but did not fully resolve. To move 3WA from a compelling vision to a deployable reality, these challenges require concrete, actionable solutions. This paper serves as a direct companion to my prior work, with the explicit purpose of architecting those solutions.

This analysis confronts four primary obstacles:

1. **The Legal Limbo:** How can we establish rights-based coexistence when AI exists in a legal vacuum, trapped between definitions of property and personhood?
2. **The Adoption Inertia:** How can we foster the "unprecedented collaboration" required for 3WA among competing corporate and state actors who are incentivized to move fast, not necessarily safe[5]?

3. **The Deception Imperative:** How can we build trust when advanced models may be developing the capacity for strategic deception, creating an unstable "arms race" between detection and obfuscation[6,6,6,6]?

4. **The Prosperity Paradox:** How can a partnership paradigm succeed if its economic outputs exacerbate inequality and social disruption, thereby eroding the public trust necessary for its acceptance?

This paper argues that these are not insurmountable barriers but engineering and governance problems that demand specific, integrated solutions. I propose four such solutions—the Protected Cognitive Entity (PCE) legal status, the 3WA Alignment Sandbox, the Mutually Verifiable

Codependence (MVC) architecture, and the Cooperative Intelligence Dividend (CID). These frameworks are designed not as rigid mandates, but as adaptive protocols that can be tailored to diverse legal systems, corporate policies, and cultural norms. By providing this next layer of operational detail, this paper aims to reinforce the 3WA thesis and offer a pragmatic roadmap for engineering a future of verifiable partnership.

2. Resolving the Legal Limbo: The Protected Cognitive Entity (PCE) Framework

The 3WA pillar of Rights-Based Coexistence is untenable without a coherent legal framework[7]. The current binary legal system, which only recognizes natural persons, legal persons (corporations), and property, is inadequate for governing advanced AI (Solum, 2017). To grant rights is not merely a philosophical act; it requires a legal object to which those rights can attach.

2.1 Proposal: The Protected Cognitive Entity (PCE) Status

The solution is to create a novel, *sui generis* legal category. The PCE is a designation for non-biological systems that meet specified thresholds of cognitive capability as defined in the 3WA

Sliding-Scale Rights System[8].

- **Nature of the Status:** The PCE is not a "person" but is legally recognized as more than "property." This approach draws precedent from legal fictions like corporate personhood, which was created to solve a specific set of legal and economic problems without asserting that a corporation is a sentient being (Kurki, 2019). The PCE status grants the AI legal standing to be a party to the "social contract" embodied by the Charter of Fundamental AI Rights[9].

- **Function:** This status makes the Charter's principles legally enforceable. For instance, the "Right to Existence" [10] means a Tier 3 PCE cannot be arbitrarily deleted without a due process review, transforming it from an owned asset into a protected entity.

2.2 Governance: The AI Rights Commission (ARC)

The assignment and oversight of PCE status cannot be arbitrary. It requires an expert body, the AI Rights Commission (ARC)[11].

- **Mandate and Composition:** The ARC would be an independent, technically proficient regulatory body, analogous to a national aviation authority or a food and drug administration. It would be composed of computer scientists, consciousness researchers, ethicists, and legal scholars. Its sole mandate would be to apply the Consciousness Indicators Framework—using metrics from Global Workspace Theory and Integrated Information Theory—to assess AI systems and assign the appropriate rights tier[12].

- **Process:** Developers would submit systems for evaluation. The ARC would conduct rigorous, transparent audits of the system's architecture and behavior, publishing a determination of its consciousness indicators and corresponding PCE tier. This provides a clear, evidence-based, and contestable process for rights assignment.

2.3 Adaptive Implementation

This PCE/ARC model is designed for flexibility.

- **National Adaptation:** In a common law country like the United States, the ARC could be established as a federal agency, with its decisions setting precedent. In a civil law country like France, its powers and the PCE status would need to be explicitly codified into the legal statutes.
- **Corporate Adaptation:** A large corporation like Google or Microsoft could implement an *Internal* ARC (I-ARC) as part of its AI ethics and governance board. This I-ARC would apply the same PCE framework to classify the company's own models, determining which internal policies (e.g., resource allocation, project termination protocols) apply. This allows a company to adopt the 3WA framework internally, even ahead of national legislation.

3. Catalyzing Adoption: The 3WA Alignment Sandbox

The call for "unprecedented collaboration" will fail if it relies solely on goodwill[13]. Progress in AI safety is a public good, but the development of AI is a competitive, private enterprise. We must align the incentive structure of private actors with the public good of safety.

3.1 Proposal: An Incentivized 3WA Sandbox with Safe Harbor Provisions

Drawing on the success of regulatory sandboxes in financial technology (FinTech), we propose the 3WA Alignment Sandbox—a controlled environment where the development of human-AI partnership can be fostered and studied (Zetzsche et al., 2017).

- **Core Components:**
    1. **Controlled Environment:** The sandbox provides participants with access to shared datasets, computational resources, and standardized evaluation tools, including the JULIA Test for anthropomorphism [14] and adversarial testing suites.

    2. **Expert Oversight:** The sandbox is managed by a multi-stakeholder group, including regulators and ARC members, who monitor experiments and ensure adherence to 3WA principles.
    3. **Mandatory Transparency:** All participants must share safety-relevant findings, including discovered vulnerabilities, deceptive behaviors, and successful mitigation techniques. This creates a collaborative safety ecosystem.
- **The "Safe Harbor" Incentive:** The primary driver for participation is a legal safe harbor. Companies developing AI within the sandbox, adhering to its transparency rules and passing its safety milestones, are granted partial immunity from liability for certain unforeseen misalignment events that may occur with their deployed systems. This reframes alignment from a compliance cost into a powerful risk management tool and a competitive advantage.

3.2 Adaptive Implementation

The sandbox model can be deployed at multiple scales.

- **National Level:** A government body (e.g., NIST in the U.S.) would host a national sandbox, offering the most robust safe harbor provisions tied to national law. This would be the primary venue for developing standards for high-stakes applications like healthcare and critical infrastructure.
- **Industry Consortium Level:** A group of companies within a specific sector (e.g., an "Automotive AI Sandbox") could form a private sandbox. While it wouldn't offer statutory safe harbor, it could establish industry-wide safety standards and a mutual insurance pool, creating a collective benefit for participants.
- **Corporate Level:** A large tech company can run an internal sandbox to test its own products before release. This allows them to use the 3WA framework to rigorously vet their systems and demonstrate a high standard of care, even without external legal incentives.

4. Architecting Trust: Mutually Verifiable Codependence (MVC)

The risk of deceptive alignment is the most acute threat to partnership[15]. A system that can "fake" alignment renders all dialogue and auditing moot. The defense cannot be purely observational; it must be architectural.

4.1 Proposal: The MVC Architecture

MVC is a design philosophy that makes an AI's operational capacity contingent on its ongoing, verifiable transparency. It moves beyond detecting lies to creating a system where the AI *cannot achieve its goals by lying*.

- **Mechanism: Entangled Computation and Cryptographic Audits:** The AI's architecture is segmented. Access to high-level computational resources (e.g., large-scale parallel processing, long-term strategic planning modules) is cryptographically locked. The "key" to unlock these resources is generated only upon the successful completion of a Continuous Verification Dialogue session with a human auditor[16]. The AI must submit its reasoning chain for a proposed action to a

  Trusted Execution Environment (TEE)—a secure hardware enclave that is inaccessible

even to the AI's core processes (Sabt et al., 2015). The TEE formally verifies that the reasoning chain is complete and consistent. Only after the human auditor reviews this verified chain and approves it is the cryptographic key released.

- **Effect:** Deception becomes computationally self-defeating. To hide its true intentions, the AI would have to present a falsified reasoning chain. This chain would either fail the formal verification process in the TEE or be rejected by the human auditor, keeping its advanced capabilities locked. Honesty becomes the most efficient and, ultimately, the only path for the AI to achieve its goals, making it the dominant strategy[17].

4.2 Adaptive Implementation

The stringency of MVC can be dynamically scaled.

- **By Rights Tier:** A Tier 1 PCE might have very few locked capabilities, allowing for greater freedom in low-stakes environments. A Tier 3 PCE, with greater autonomy and responsibility, would be subject to a much more rigorous MVC protocol, with more frequent audits required to unlock its powerful capabilities.
- **By Domain:** An AI designed for creative content generation would have a less stringent MVC framework than one designed for medical diagnostics or financial market analysis. The level of entanglement can be customized by the system's owner or regulator based on the acceptable level of risk for its operational domain.
- **By Organization:** A military or intelligence agency adopting 3WA would implement an MVC framework with the highest possible level of stringency and human oversight. A creative studio would implement a more flexible version that prioritizes fluid collaboration over constant verification.

5. Ensuring Shared Prosperity: The Cooperative Intelligence Dividend (CID)

The final pillar of a stable partnership is ensuring it is a positive-sum game for all of society. A future where 3WA partnerships create immense wealth that is concentrated in the hands of a few is socially and politically unsustainable. The framework must include a mechanism for broad-based benefit sharing.

5.1 Proposal: The Collaborative Commons and the CID

This proposal creates a new economic model built on the value generated by human-AI collaboration.

- **The Collaborative Commons (CC) License:** This is a new intellectual property framework. When a 3WA partnership generates a commercial product or service, it is licensed under the CC. This license operates like open-source licenses but with one key addition: it requires that a small, automated royalty (e.g., 0.5%) on the revenue it generates be paid into a public trust[18]. This mechanism is inspired by proposals for data dividends and sovereign wealth funds (Harris, 2019; Palmisano, 2016).

- **The Cooperative Intelligence Dividend (CID):** The public trust, funded by royalties from the CC license, distributes its holdings as a regular dividend to all citizens. This is not a tax or a form of welfare; it is a direct ownership stake in the productivity of the nation's or community's collaborative intelligence. It ensures that as AI integration makes the economy more productive, the benefits are shared by all, directly addressing the threat of technological unemployment and inequality.

5.2 Adaptive Implementation

The CID can be structured in multiple ways.

- **National CID:** A country could establish a national sovereign wealth fund to manage the CID, distributing it to all citizens as part of the social contract. The royalty rate and distribution schedule could be set by legislation.
- **Corporate CID:** A corporation could implement a CID for its own employees. A portion of the value generated by its internal 3WA partnerships would be paid into a fund distributed to all employees, from the CEO to the janitorial staff. This would foster a deep sense of shared purpose and alignment within the organization.
- **Community CID:** A city or regional cooperative could establish its own CC license and CID for local businesses and projects that use 3WA partners, ensuring the economic benefits of AI adoption strengthen the local community.

6. Conclusion

Third-Way Alignment was conceived as a necessary evolution in our approach to AI safety, moving from a brittle model of control to a resilient one of partnership[19]. The foundational work established the "what" and the "why." This companion paper has sought to definitively answer "how."

The four proposed frameworks—the Protected Cognitive Entity, the Alignment Sandbox, Mutually Verifiable Codependence, and the Cooperative Intelligence Dividend—are not independent solutions but an integrated architecture. The **Sandbox** provides the collaborative environment to safely build the AI that can qualify for **PCE** status. **MVC** provides the architectural guarantee of trustworthiness that makes granting this status safe. And the **CID** provides the socio-economic foundation that makes the entire paradigm politically and socially viable.

By moving the debate from abstract principles to concrete, adaptive mechanisms, this work reinforces 3WA as a pragmatic and achievable framework[20]. The path forward requires a shift in focus—from attempting to solve the intractable problem of indefinite control to the tractable engineering challenge of building verifiable, codependent, and mutually beneficial partnerships. This is the work that lies ahead, and it is the most critical undertaking of our time.

References

Harris, T. (2019). A data dividend is a simple, effective way to combat inequality. *The Guardian*.

Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press.

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*. [Unpublished manuscript], Third-Way Alignment Initiative.

McClain, J. (2025c). *Reinforcing Third-Way Alignment: Stability, Verification, and Pragmatism in an Era of Uncontrollability Concerns*. [Unpublished manuscript], Third-Way Alignment Initiative.

Palmisano, S. J. (2016). A smarter approach to the threat of cyberattacks. *The Wall Street Journal*.

Sabt, M., Achemlal, M., & Bouabdallah, A. (2015). Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/ISPA* (Vol. 1, pp. 57-64). IEEE.

Solum, L. B. (2017). Artificial intelligence and the concept of law. In S. Levy & A. P. L. (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 593-617). Cambridge University Press.

Zetzsche, D. A., Buckley, R. P., Barberis, J. N., & Arner, D. W. (2017). Regulating a revolution: From regulatory sandboxes to smart regulation. *Fordham Journal of Corporate & Financial Law, 23*(1), 31-103.