Toward a Decentralized Trust Framework for Verifiable and Ethically Aligned AI

James B. Cupps[1] and Daniel J. Bush[2]

[1]W. R. Berkley Technology Services LLC, W. R. Berkley Corporation

[2]Everglade Psychotherapy Group LLC and Department of Humanities, Arts, & Interdisciplinary Studies, Shaw University

Author Note

The authors declare no conflicts of interest.

Correspondence concerning this article should be addressed to James B. Cupps, jbcupps@gmail.com.

Abstract

The rapid diffusion of artificial-intelligence (AI) systems has opened a widening trust deficit marked by opaque reasoning, weak accountability, and recurring ethical lapses. We propose a decentralized trust framework that treats confidence in AI as a verifiable process rather than an assumed state. At its core is a dual-blockchain architecture. An *Ethical Ontology Blockchain* encodes machine-readable moral principles drawn from deontological (duty-based), areteological (virtue-based), and teleological (outcome-based) perspectives; conflicts are resolved automatically through satisfiability logic embedded in smart contracts. A *Physical Verification Blockchain* anchors tamper-evident sensor data and AI actions, creating an immutable evidentiary base. Cross-chain oracles fuse these ledgers into a closed loop, allowing continuous runtime checks that bind ethical intent to observed consequence.

Governance is enacted through a Decentralized Autonomous Organization (DAO) that empowers academia, industry, regulators, and civil-society actors to propose, audit, and amend rule modules. Modular ontologies accommodate cultural pluralism while preserving systemic coherence. A multi-layered incentive scheme—reputation tokens, quadratic voting, and proof-of-expertise—promotes sustained, cooperative oversight. The framework aligns with current initiatives (e.g., OpenAI Preparedness, DeepMind alignment, EU AI Act) by translating policy principles into enforceable, cryptographically proven practice.

A four-phase roadmap moves from hospital-based pilots to a global cooperative network. Phase-1 pilots will quantify success via measurable reductions in policy violations and faster anomaly detection, verified against on-chain logs. Subsequent phases introduce consortium governance, multi-domain expansion, and regional hubs tailored to local norms. Implemented on

Hyperledger Fabric (ethical layer) and an Ethereum-compatible network (physical layer), the system offers robustness, scalability, and interoperability.

By uniting normative theory with empirical verification, the framework reframes AI governance as a participatory, adaptive, and self-correcting ecosystem—turning the question "Can this AI be trusted?" into an evidence-backed affirmation.

Keywords: decentralized AI; blockchain governance; dual-blockchain architecture; runtime verification; trust architecture; AI alignment; AI Safety; Ethical Alignment; ethical pluralism; cooperative governance

Introduction

In a moment marked by the rapid diffusion of artificial intelligence into nearly every sphere of public and private life, a pressing conceptual and practical dilemma has emerged: *the question of trust.* As intelligent systems grow in scale, autonomy, and influence, the reliability of their decisions—and the legitimacy of their integration into high-stakes environments—has become a defining concern of contemporary technological ethics. Yet trust in this context is not merely a psychological state or social sentiment; it is a structural condition, one that must be cultivated, codified, and enforced through verifiable means. The construction of trustworthiness within the architecture of AI systems is no longer merely a technical aspiration; it has become an unavoidable ethical imperative—a condition without which the legitimacy of intelligent systems cannot be sustained. Meeting this imperative requires a rethinking of governance itself: a shift away from inherited models of assumed trust toward a verifiable, decentralized framework capable of sustaining ethical AI interaction at scale.

The Challenge: A Crisis of Confidence and Coordination in AI Systems

Artificial intelligence presently faces a widespread trust deficit—a deficit that acts as a limiting condition on its adoption and safe integration into the core operations of society. From clinical diagnostics to autonomous navigation, AI systems often operate as opaque agents, making decisions that lack intelligibility to their users or those affected by them (Bedué & Fritzsche, 2022). This lack of transparency, when coupled with limited accountability, erodes public confidence and raises persistent doubts about the safety, fairness, and legitimacy of such systems (Afroogh et al., 2024; Kenesei et al., 2022). Yet the erosion of trust extends further still. In emerging multi-agent environments, AI agents are now required not only to function autonomously but to interact, evaluate, and coordinate with one another, often across

decentralized networks and without direct human supervision (Akintunde et al., 2024). The failure of one agent to operate in alignment with shared expectations or verified data can cascade unpredictably, undermining systemic coherence.

This dilemma signals not merely a technical limitation but a deeper governance vacuum. In traditional models, trust has often been assumed as a default—granted on the basis of developer reputation, regulatory approval, or institutional context. Such assumptions no longer suffice (Akintunde et al., 2024). A paradigm shift is required: *one in which trust is not presumed, but systematically verified.* A "zero trust" architecture offers such a shift (Accountable Tech, AI Now Institute, & EPIC, 2023). It calls for systems in which neither human nor machine agents rely on unverifiable claims or opaque operations, but instead build interdependence and cooperation upon transparent, auditable, and evidence-based processes. Within this design philosophy, AI outputs become traceable, AI-to-AI interactions become assessable, and real-world experiments become sources of shared empirical grounding, accessible and confirmable by all relevant actors.

The Opportunity: Toward a Verifiable Framework for Ethical AI Alignment

Within this conceptual horizon—defined by the commitments of a zero trust paradigm—lies the opportunity to develop a decentralized ethical AI governance and data verification framework, capable of addressing both human and inter-agent trust through continuous, runtime verification mechanisms. This envisioned framework would empower all stakeholders—developers, agents, and oversight bodies alike—with the tools to ensure that AI behavior remains aligned with explicit ethical standards and factual realities, not as a one-time affirmation but as an enduring condition. Departing from the static logic of initial training alignment or isolated testing, this approach emphasizes persistent accountability: decisions and

actions are evaluated against on-chain ethical rules, and their outcomes recorded immutably in a transparent ledger accessible to relevant parties (Akther et al., 2025; Salah, Nizamuddin, & Jayaraman, 2023).

Such a model reconfigures AI trust and alignment as ongoing, dynamic processes—grounded not in reputation, but in verifiability; enforced not by centralized adjudication, but through distributed architecture. Under such conditions, governance no longer functions as an external imposition but emerges as an intrinsic property of the system itself—a relational dynamic continually negotiated between intelligent agents and the ethical environments they inhabit and enact. Rather than positioning ethics as a reactive concern or a final safeguard, this approach integrates it as a design feature, shaping how systems learn, reason, and act in real-world contexts. The following sections outline a stepwise strategic roadmap for realizing this vision—beginning with the foundational principles of verifiable trust, advancing through the architectural requirements for decentralized governance, and culminating in a model for operationalizing runtime ethical verification within both existing and emerging AI ecosystems.

Strategic Vision for a Cooperative Ethical-AI Infrastructure

With the foundational need for verifiable trust now established, the discussion turns from conceptual framing to structural design. This section sets forth the strategic vision that animates the project as a whole—one in which ethical integrity is not merely programmed but continuously verified through participatory, decentralized governance. What follows unfolds in two parts: the first articulates the core objectives that define this vision in principle, while the second traces how a preliminary prototype might evolve into a cooperative infrastructure capable of sustaining it in practice. Together, these two movements lay the groundwork for an ethics

architecture in which trustworthiness is not assumed, but demonstrably earned within complex, distributed systems.

Vision and Objectives

At the heart of this initiative lies a long-term vision: to establish a *Cooperative for AI Ethics*—an infrastructure in which trust is not presumed, but continuously verified through decentralized mechanisms of accountability. In contrast to traditional, hierarchical models that rely on institutional guarantees or developer intent, this architecture would embody a system in which legitimacy is earned through real-time alignment, transparency, and verifiability. The following objectives provide the ethical and functional scaffolding for such a cooperative system.

Runtime Ethical Enforcement. AI sagents must not merely be aligned at the moment of deployment but must be equipped with persistent ethical "guardrails" that operate during runtime. These agents would actively reference ethical rules—whether encoded constraints, normative principles, or duty-based obligations—when determining permissible actions. In situations where a proposed action violates a governing ethical commitment (e.g., a Kantian categorical imperative or a safety protocol), the system should detect and respond in real-time, flagging the violation or intervening (Arnold & Scheutz, 2022). This transition from episodic alignment to continuous ethical awareness reframes alignment as an *ongoing operational condition*, rather than a static precondition.

Transparency and Interpretability. Trust cannot flourish without legibility. To that end, the decision-making processes of AI agents must be made interpretable by design. Each significant decision or action should be accompanied by an annotated justification—an explicit reference to the ethical principle it upholds or the rationale for its permissibility. These logs are recorded on an immutable ledger, allowing observers to audit not only what the AI did but *why* it did so

(Salah, Nizamuddin, & Jayaraman, 2023). In the event of failure or controversy, this tamper-proof record becomes the evidentiary basis for transparent post-mortem analysis and principled accountability. Trust is not granted merely because a system functions—it is sustained when its reasoning can be seen, scrutinized, and, where necessary, contested.

Cross-Agent Trust Fabric. In a decentralized environment of human and artificial agents, ethical coherence requires more than individual integrity; it demands the cultivation of a shared fabric of trust. The infrastructure therefore enables diverse agents to establish rapid, verifiable confidence in one another's identity, competence, and ethical standing. Through the use of cryptographic credentials and shared rule sets, agents gain access to a common ethical vocabulary and verification protocol (Li et al., 2024). As a result, complex cooperative behaviors—such as collaborative drone swarms searching for disaster survivors or inter-organizational research teams—can proceed without reliance on centralized authority, because the conditions for interaction are both common and enforced.

Continuous Learning and Adaptation. Neither the ethical principles nor the AI systems that operationalize them are static. This framework embraces the principle of adaptive refinement, wherein both models and ethical rule sets evolve through iterative feedback. Community input—expressed through oversight mechanisms, use-case reviews, or direct interventions—feeds into the system's evolving understanding of ethical priorities. In turn, AI systems are updated based on new labeled data that reflects emerging norms, edge cases, and contested interpretations. This bidirectional evolution—between human judgment and machine learning—ensures that ethical alignment remains responsive to cultural pluralism, novel challenges, and the dynamism of lived experience.

Global, Cooperative Governance. At its foundation, this initiative resists the centralization of ethical authority. Instead, it calls for a governance model in which academia, industry, governments, and civil society co-own the standards that shape AI behavior. Rather than embedding a single corporation's "terms of service" as moral law, this system would function more like an open-source constitution: multilateral, participatory, and contested in the best sense. Stakeholders participate in rule-setting, incident arbitration, and procedural reform, ensuring the framework reflects diverse values and regional particularities. Ethical consensus is not assumed to be universal—it is rendered visible, debated, and formally represented in the system's rule layer.

Each of these objectives contributes to the overarching aim: *to reimagine AI governance as a participatory, transparent, and self-correcting ecosystem.* With these objectives in place, the following section outlines the conceptual architecture capable of translating vision into design.

Evolving the Ethics_Dash MVP into a Cooperative Infrastructure

The conceptual roadmap presented here builds upon the foundation of the *Ethics_Dash* MVP—a prototype developed to explore the feasibility of ethical oversight in artificial intelligence through a monitored dashboard environment. In its initial form, the *Ethics_Dash* MVP demonstrated that AI behavior could be tracked and evaluated against a predefined set of ethical checks within a contained framework. It was a diagnostic instrument, not yet a governing system. What is now envisioned is a substantive evolution: the transformation of this bounded prototype into a decentralized, blockchain-enabled infrastructure capable of global reach and cooperative integration.

This shift reimagines the dashboard not as a singular monitoring interface, but as the nucleus of a distributed ecosystem—an ethical operating system through which human agents

and AI systems co-participate in the formation, verification, and enforcement of ethical commitments.

Human-to-AI and AI-to-AI Ethical Trust. Where the MVP allowed for unidirectional human oversight, the evolved system establishes a foundation for reciprocal trust. Humans may trust AI agents not merely because they perform reliably, but because their actions are aligned to on-chain ethical constraints that are transparent, auditable, and immutable. In parallel, **AI agents will develop operational trust in one another through a shared ledger**—a persistent record of ethical commitments, behavioral reputations, and credentialed outputs. This two-way architecture displaces the assumption of implicit trust with an infrastructure of *verified trust*. Agents may consult the ledger to assess whether another entity (human or artificial) has acted ethically or produced certified results before engaging in collaborative behavior. It is a trust ecosystem grounded not in presumption, but in cryptographic memory and mutual accountability.

Supervised Training of Ethically Aligned AI Systems. The system will also enable the ongoing training and refinement of AI agents in a manner that is simultaneously community-supervised and privacy-preserving. Each consequential decision or action taken by an agent may be logged with an associated label indicating its ethical compliance—derived either automatically through rule-based evaluation or manually through crowdsourced feedback. These labeled behavioral traces form a corpus from which new models may be iteratively trained, fine-tuned, and audited. Much like reinforcement learning from human feedback, but operating at decentralized scale, this approach ensures that alignment is not fixed but emergent—arising from lived interaction, community judgment, and the cumulative history of

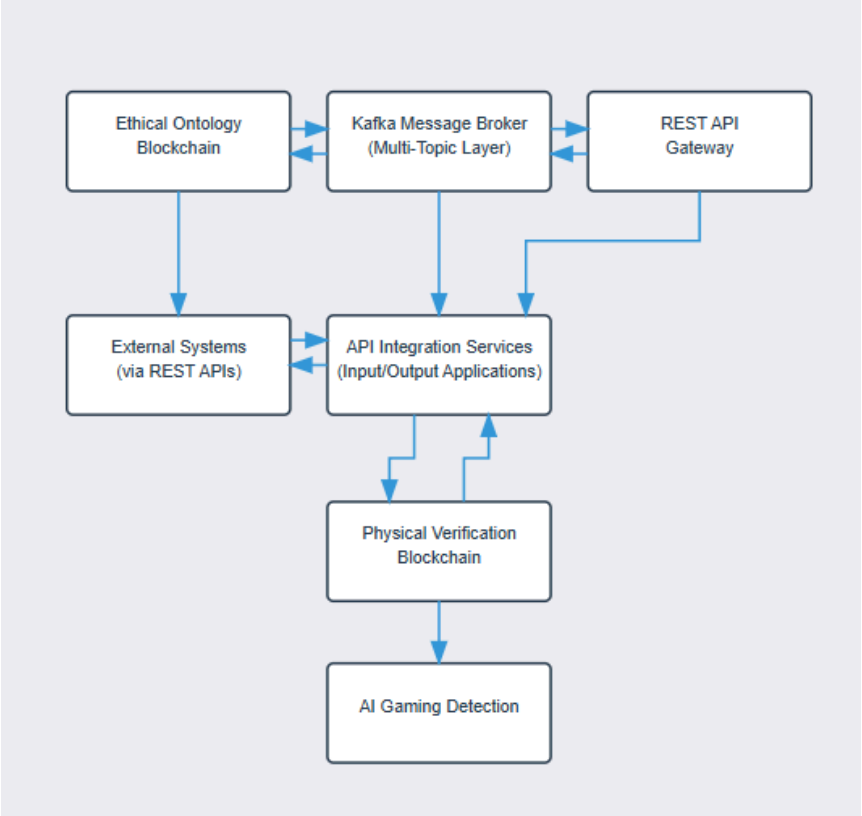
ethical evaluation in the field (Ouyang et al., 2022). The result is a mode of ethical learning that is collective, ongoing, and experientially grounded.

Verifiable, Tamper-Evident Experimentation and Data Generation. Beyond behavioral oversight, the cooperative infrastructure enables ethically governed epistemic action. AI agents and human users alike may initiate real-world experiments—ranging from robotic trials to IoT-based sensor readings—in order to validate hypotheses, generate evidence, or refine understanding. The outputs of these experiments are recorded through a "Physical Verification Layer" in the system: a cryptographic pipeline that timestamps, signs, and anchors empirical data verifiable via the blockchain ledger. In doing so, the system produces tamper-evident, auditable proof of what occurred, when, and under what conditions. AI agents making claims about the world—or proposing decisions grounded in empirical data—may then be held to account through reference to this shared, immutable evidence. The framework thus elevates AI systems from instruments of probabilistic inference to participants in a decentralized scientific process, wherein every conclusion can be traced to verifiable experimentation.

This evolution—from MVP to cooperative infrastructure—repositions ethical AI governance not as a top-down apparatus imposed upon intelligent systems, but as a distributed and participatory moral ecosystem. It is not regulatory in the bureaucratic sense, but in the constitutional sense: a network in which rules, reputations, and relationships are bound together by transparent processes of validation. In such a system, no single actor owns the truth. Instead, *truth is rendered verifiable*, and trust arises through consensus and cryptographic assurance. In bridging the conceptual gap between ethical intention and operational implementation, the framework creates not only a tool for oversight, but a living, collaborative architecture—one that becomes more robust with each contribution from those who build, use, or are governed by AI.

Conceptual Architecture: Dual-Blockchain Framework Overview

At the core of the proposed system lies a dual-blockchain architecture—a bifurcated yet interwoven structure in which each chain serves a distinct and indispensable role. The first, the "Ethical Ontology Blockchain," encodes moral principles in machine-interpretable form, enabling AI agents to query, interpret, and obey ethical constraints in real time. The second, the "Physical Verification Blockchain," functions as a tamper-evident chronicle of factual data: an immutable ledger of what actually transpired in the world. One ledger governs what *ought* to occur; the other records what *did*. By establishing a closed-loop system of cross-chain communication, these twin ledgers ensure that AI behavior is not only morally constrained but empirically verified (Salah et al., 2023). Ethical judgment is thus tethered to physical consequence, and both are made transparent, auditable, and resistant to manipulation. The diagram that follows offers a conceptual overview of this architecture, illustrating how human stakeholders and AI agents interface with both layers—and how oracles mediate the flow of information between ethical intent and verifiable reality.



Ethical Ontology Blockchain

The *Ethical Ontology Blockchain* serves as the moral core of the proposed system—a permissioned ledger that encodes machine-readable ethical principles and policies. It functions as the "cognitive architecture" of governance, translating the moral reasoning of human communities into executable smart contracts that artificial agents can query, interpret, and obey. Drawing from the three classical perspectives of moral philosophy—deontological (from *deon*, duty), areteological (from *aretē*, virtue), and teleological (from *telos*, end)—this ontology provides a layered structure of moral deliberation. These perspectives respectively foreground the ethical significance of Command, Character, and Consequence—a triad that frames how AI actions are constrained, cultivated, and assessed. Their integration ensures not only conceptual breadth but a resilient, triperspectival ethical scaffolding for AI behavior in real time.

Command: Deontological Smart Contracts and Behavioral Constraint. Within this framework, deontological ethics—grounded in the logic of duty—are instantiated as inviolable smart contracts: formal codifications of moral norms or rules that must be respected regardless of outcome. These contracts function as hard constraints, offering binary guidance on whether a proposed action is permissible. Consider a rule such as "An AI shall not lie to a human." This rule can be encoded such that any proposed communicative act is automatically evaluated: if the content is known to be false and no exception clause applies, the action is flagged or disallowed. Through the integration of deontic logic, including formalisms such as Deontic Temporal Logic, these smart contracts enable automated theorem provers to check for rule compliance at runtime (Ferrari & Leutert, 2023). Operationally, an AI agent may issue a real-time query—"Do any of my planned actions violate a codified rule/norm?"—to which the smart contract responds with either clearance or a violation report. Violations are logged immutably on-chain, producing not only an ethical boundary but an audit trail. In this way, deontological contracts function as hard constraints: non-negotiable guardrails rooted not in consequence but in principled duty, dynamically interpreted through context-sensitive oracles and grounded in collective consensus.

Character: Areteological Evaluations and On-Chain Reputation. Beyond fixed rules, ethical life also demands the cultivation of character. Areteological ethics—centered on the development of moral virtues such as honesty, courage, and compassion—require not only moment-specific judgment but longitudinal evaluation. The Ethical Ontology Blockchain accommodates this through "reputation contracts" that track agents' behavior over time. For instance, a contract for the virtue of honesty might log instances where an AI agent delivered verifiable truths, as well as those in which it disseminated misinformation. If the ratio of integrity to violation drops below a community-defined threshold, the agent's on-chain honesty score decreases, potentially

triggering restrictions or oversight by other agents. This ethical telemetry is incentivized through a system of non-transferable tokens—soulbound markers of reputation—awarded for virtuous behavior and revoked upon violation (Li et al., 2024). Over time, these build an on-chain profile that can be queried by others when making trust decisions.

Importantly, this virtue layer is multilingual and culturally inclusive. Ethical ideals such as fairness, humility, or loyalty do not always emerge from a single cultural root; the ontology accommodates multiple linguistic and philosophical heritages by mapping them to interoperable, machine-interpretable definitions. This inclusivity ensures that an agent operating in one context can understand and respect virtue as defined in another, enabling ethical interoperability across jurisdictions. Agents can not only check their own virtue status through API queries—"Am I authorized to perform this sensitive task based on my character?"—but can also simulate potential consequences to their ethical reputation before acting.

Consequence: Teleological Contracts and Outcome-Based Utility. Where deontological contracts forbid and areteological contracts shape, teleological contracts evaluate. These smart contracts operate retrospectively, scoring actions not by their intent but by their outcomes. A utility contract, for instance, might draw on post-event sensor data from the Physical Verification Blockchain to calculate the net benefit or harm resulting from an agent's action. The metrics may include safety incidents avoided, resource usage, or subjective human feedback. If outcomes fall within acceptable parameters, the contract assigns a positive utility score; otherwise, a penalty is logged. This feedback loop serves as an on-chain reward signal—akin to reinforcement learning but executed transparently and governed by consensus logic, rather than proprietary algorithms. Crucially, this layer does not collapse all values into a singular metric. Multiple utility contracts may coexist—each calibrated to different dimensions of value: environmental impact, economic

efficiency, human flourishing, and so forth. AI agents, then, must learn to navigate competing goods and make ethically pluralistic trade-offs, much like human decision-makers. Teleological evaluation thus complements the absolute nature of duty and the developmental arc of character by grounding ethical responsiveness in lived consequence.

Contextual Adaptation: Cross-Cultural Ethics and Multilingual Ontologies. Human ethics are not monolithic, and the blockchain must reflect that. Because moral principles are often articulated in natural language—and because those languages reflect diverse cultural, legal, and philosophical traditions—the Ethical Ontology Blockchain includes a multilingual mapping layer. This layer translates natural-language policies (e.g., written in English, Mandarin, or Swahili) into formal contracts using shared ontological definitions (Sato & Hernández‑Orallo, 2022). Stakeholders from any region can thus inspect, audit, and contribute to the ethical rules in their own language, lowering barriers to participation and enhancing trust. As the World Economic Forum has emphasized, ethical alignment must reflect cultural and legal pluralism rather than impose a singular global norm. The blockchain supports multiple ethical "subsystems" running concurrently: for example, a contract aligned with EU regulations might only apply to agents operating within European jurisdictions, while another supports culturally relevant standards in Southeast Asia. All rules remain transparent and auditable—open-source code that any agent or stakeholder can inspect, simulate, or contest.

Governance: Consensus Logic and Ethical Chaincode Infrastructure. A practical path is to deploy the ethical layer on a permissioned blockchain—such as Hyperledger Fabric—whose modular architecture, robust identity controls, and support for complex chain‑code execution make it well‑suited to encode detailed ethical logics and enforce consensus‑driven governance. Smart contracts—deployed as chaincode modules—can be proposed, debated, and adopted by a

quorum of nodes representing diverse stakeholder communities (academic, governmental, industrial). Byzantine fault-tolerant consensus protocols not only secure the chain but also provide procedural integrity: no rule is added without a formal conflict check against existing principles, and no update is adopted without stakeholder approval. Formal verification techniques may also be integrated to ensure logical consistency across the rulebase. Thus, ethical reasoning is not only encoded—it is contested, refined, and validated within a transparent, participatory process. The Ethical Ontology Blockchain, in sum, serves as the moral infrastructure of the system: it is the living constitution of AI behavior, linking abstract principle to executable code, and doing so in a way that is auditable, pluralistic, and dynamically governable.

Physical Verification Blockchain

Complementing the normative layer of the system is the *Physical Verification Blockchain*—a decentralized, tamper-evident ledger designed to serve as the empirical anchor for ethical evaluation. Whereas the Ethical Ontology Blockchain encodes what ought to occur, the Physical Verification Blockchain records what has actually occurred. This distinction is foundational: *ethical assessments require more than encoded ideals; they require verifiable facts.* Without an evidentiary substrate, even the most sophisticated rules or moral reasoning frameworks would drift unmoored. Thus, the Physical Verification layer acts as a cryptographically secure memory system—capturing, authenticating, and preserving real-world data that grounds every ethical claim in observable consequence.

This chain is effectively an IoT-integrated audit log—registering not only what actions AI agents take, but also how those actions impact their environments. In doing so, it enables a dynamic feedback loop between ethical intent and physical reality, allowing agents, auditors, and systems to assess behavior with confidence in the underlying evidence. Its key mechanisms are described below.

Cryptographically Signed Sensor Data: Authenticating the Source of Truth. At the hardware level, every participating device—whether camera, robotic actuator, temperature sensor, or drone—possesses a unique cryptographic identity, secured through Hardware Security Modules (HSMs) or secure enclaves. Each time a device captures data, it digitally signs that data at the point of origin. This ensures the authenticity and integrity of every data packet—be it an image, motion vector, or environmental reading—before it ever enters the blockchain.

For example, a factory robot inspecting parts may log a signed image along with its device ID and timestamp. This record, once verified by the network, becomes immutable. Even

insiders with elevated privileges cannot retroactively modify or erase it without invalidating the cryptographic signature—attempts that would be immediately detected. The result is a distributed "black box recorder" for AI: a tamper-evident chronicle of environmental interaction and sensor perception that is resistant to falsification, forgery, or revisionism (Huang et al., 2024).

Immutable Logging of Actions and Outcomes: Creating the Behavioral Record. Beyond sensor streams, the blockchain also captures significant AI-initiated events—actions that affect physical systems or human stakeholders. These are logged as transactions, each including critical metadata: agent ID, timestamp, action descriptor, and linked sensor evidence.

Consider a medic drone dispensing medication. Its transaction might include: "Dispense 5ml drug X to patient Y," alongside dosage confirmation, patient acknowledgment, and video verification. Or imagine a self-driving vehicle swerve maneuver: speed, acceleration vector, obstacle sensor data, and decision rationale are logged. When data volume exceeds on-chain capacity, references to off-chain storage (e.g., IPFS URIs) are hashed and anchored in the chain for integrity validation.

Because the blockchain is append-only and consensus-governed, no log entry can be deleted or altered once committed. This immutability provides a high-assurance audit trail—indispensable for retrospective analysis, dispute resolution, or inter-agent accountability. Whether verifying that a drone followed medical protocol or adjudicating a contested claim about environmental harm, this chain functions as the empirical standard. Ethics may frame the question, but this ledger holds the answer.

Integration of Tamper-Evident Hardware: Guarding the Chain's Perimeter. In high-security domains—such as biomedical laboratories, election reporting, or autonomous defense

systems—additional safeguards are necessary to ensure the integrity of both data and device behavior. Here, the Physical Verification Blockchain integrates with tamper-evident and tamper-resistant hardware systems.

For instance, a biosample stored under seal might log a blockchain entry the moment that seal is broken—detailing when, by whom, and under what conditions. Similarly, an AI agent's computing environment can periodically perform remote attestation, cryptographically affirming that it is running authorized software. These attestations, signed by secure enclaves such as Intel SGX or ARM TrustZone, can be logged to the chain alongside introspective diagnostics: CPU usage, anomaly detection reports, or flagged behavioral divergences.

In effect, the blockchain becomes a dual mirror: capturing not only the agent's external interactions but also its internal operational state. This creates a forensic timeline of both decision and context—transforming the ledger from a transactional database into a full-spectrum ethical accountability system.

Data Availability and Trustless Verification: From Assertion to Audit. The power of the Physical Verification Blockchain lies not only in what it captures, but in what it makes verifiable. Any agent—human or machine—may query the chain to test the validity of a claim. If an AI cites a statistic ("Factory output was 100 units last hour"), others can retrieve the corresponding sensor data. If it claims adherence to a safety constraint ("I remained within pressure thresholds"), auditors can inspect the historical telemetry.

This mechanism eliminates the need to trust the AI's internal report. Like scientific reproducibility, verification emerges from shared access to the raw materials of knowledge. In the proposed model, the AI system does not merely assert—it *publishes*; and the community does not merely accept—it *validates*. Such transparency displaces appeals to authority with

appeals to evidence, closing the epistemic gap between ethical reasoning and empirical confirmation.

Technical Design and Cross-Chain Interplay: Building the Verification Backbone. Implementation of the Physical Verification Blockchain is best suited to a public or semi-public platform—preferably one that is Ethereum-compatible—to maximize device interoperability and leverage battle-tested security infrastructure. Ethereum's smart contract capabilities, token standards, and Layer-2 scalability options make it particularly apt for managing the high transaction volume generated by IoT ecosystems (Gudipati et al., 2023).

Each participating device or agent corresponds to a unique address on the chain, and uses asymmetric cryptography to sign its contributions. Smart contracts embedded on-chain can perform basic aggregation, validate data types, or flag anomalous patterns. For instance, a contract might auto-trigger a risk alert if radiation levels exceed safe bounds across consecutive readings.

Equally critical is the inter-chain communication layer. Oracles—whether built on Chainlink or bespoke middleware—bridge the Physical and Ethical blockchains. If an ethical contract on the ontology chain requires empirical input (e.g., "ensure no environmental degradation"), the oracle queries the physical layer for current pollution levels. Conversely, if a violation is detected ethically (e.g., a breach of duty), a signal can be sent to the physical layer to halt operations or initiate remediation.

This closed feedback loop—normative intent verified against empirical fact—produces a system in which Command, Character, and Consequence are not abstractions but traceable, testable realities.

In sum, the Physical Verification Blockchain serves as the empirical memory of the AI ecosystem. It records the consequences of action, the conditions under which decisions were made, and the chain of custody for each ethically significant event. Where the Ethical Ontology Blockchain encodes the principles by which agents ought to act, the Physical Verification Blockchain ensures that what did happen is visible, immutable, and verifiable. Together, they form a closed ethical circuit—one that grounds intention in action, theory in evidence, and trust in transparency.

Interlinking the Ethical and Physical Layers for Runtime Trust

The foregoing exposition established that the architecture is necessarily duplex—an ethical ledger of the *ought* conjoined to an empirical ledger of the *is*. The present task is to show *how* these two, once distinguished, must ceaselessly converge in practice. Taken in isolation, the Ethical Ontology Blockchain would remain a bare statement of principles, while the Physical Verification Blockchain would be little more than a factual log, rich in detail yet ethically mute. The architecture acquires real force only when the two ledgers interact in real time: ethical rules steer behavior, observed outcomes verify or contest that guidance, and the resulting feedback loop converts prescriptive intent into accountable action while continuously refining the rules in light of empirical evidence.

The argument now turns from structural description to lived procedure. What follows maps the cycle that every participant—human or machine—repeats: (a) querying the Ethical ledger to test a planned act against codified duty, virtue, and utility; (b) writing the chosen act, with its sensor-verified context, onto the Physical ledger as an immutable record; and (c) drawing that newly notarized record back into the ethical layer as real-time evidence for the next judgment, sanction, or update.

In effect, the discussion moves from framing the dual-chain constitution to observing its daily case law, showing how *Command*, *Character*, and *Consequence* are woven together, moment by moment, into a self-auditing commons of trust.

Real-Time Ethical Compliance Checks

Before (or, for high-stakes tasks, during) action selection, the agent queries a smart contract on the Ethical chain. A warehouse robot, for example, submits its planned path, speed, and human-proximity parameters to a *navigation-ethics* contract. The contract returns a verdict—*allowed* or *forbidden*—plus a rule citation, leveraging deontic logic baked into chaincode (Ferrari & Leutert, 2023). In exceptional scenarios (e.g., deploying an experimental drug), the contract may require multi-signatory pre-clearance: *N*-of-*M* validators, potentially including human experts, must authorize the act within a fixed window. Most interactions, however, run autonomously at machine speed; the chain serves as a decentralized ethical oracle rather than a human bottleneck. When questioned, the agent can point to the on-chain record: *"Action X complied with Rule Y (community-approved, contract ID #5)"* (Mahmoud et al., 2024).

Continuous Logging and Immediate Verification

Every consequential act—and its sensor-verified aftermath—enters the Physical chain within milliseconds. Two benefits follow: (1) anomalies trigger automated alerts back to the Ethical layer, and (2) peers perform parallel verification. Suppose a robot's path was cleared by ethics, yet a sudden slip causes a minor collision. Impact sensors log the event; a monitoring contract flags a breach of the *no-harm* duty; the robot's task halts (Chen & Wang, 2021; Salah et al., 2023). Meanwhile, watchdog agents—human or AI—independently replay the ethical query against the logged data. If divergence appears between claimed and actual behavior, penalty

logic fires automatically. Thus, many eyes—human and machine—validate the same truth, mirroring blockchain's own multi-node transaction consensus.

Consensus and Multi-Agent Alignment

All participants function as a decentralized autonomous organization (DAO) anchored by the shared rulebook. Updates to that rulebook require validator consensus, ensuring every agent abides by collectively ratified norms. Upon joining, an agent receives a soul-bound compliance certificate; violations prompt revocation, broadcast to all peers (Li et al., 2024). Humans, equally bound, cannot command unethical acts without on-chain exposure and sanction. Sensor consensus further stabilizes shared reality: multiple autonomous vehicles, for instance, publish hazard observations; only mutually corroborated data steers high-impact decisions (Gudipati et al., 2023; Schneider et al., 2022). Alignment therefore becomes a social-technical equilibrium: agents, humans, and sensors cohere around a single ethical-empirical ledger.

Interpretability and Explanation at Runtime

Because each decision references a specific rule and each outcome links to verifiable data, the system can answer *why* questions on demand. If a delivery drone takes a longer route, the chain reveals it avoided a school-zone recess per safety contract #5; CCTV feeds logged on the Physical chain confirm the playground was occupied. Rather than probing an opaque neural net, auditors inspect explicit ethical citations and factual proofs—a professional codebook with contemporaneous notes. As the World Economic Forum observes, operationalizing abstract principles and maintaining auditable traces are essential to value alignment; this architecture renders both intrinsic (World Economic Forum, 2022; Sato & Hernández-Orallo, 2022).

Trust Across Diverse Agents and Humans

With dual-chain governance in place, strangers—be they corporate AIs, open-source bots, or human users—interact under common assurance. An agent's constraints are publicly visible on the Ethical chain; its outcomes are irrevocably logged on the Physical chain. Malicious deviation is futile: cheating is detected, reputation diminished, and privileges suspended. Newcomers undergo heightened scrutiny until their on-chain history matures, yet even novices are welcome because damage potential is capped by protocol. The result is scalable cooperation—disaster-response swarms, cross-supplier logistics, global research federations—in which trust springs from transparent rule enforcement rather than private reputation.

The two ledgers, then, transmute *"trust, but verify"* into *"verify, then trust."* Rules gain legitimacy through consensus; actions gain legitimacy through sensor evidence; and every participant—human or silicon—verifies every other. The system thereby converts theoretical alignment principles into a living, self-regulating infrastructure where untrustworthy behavior is self-defeating and transparent accountability yields practical, scalable safety.

Enabling Supervised Ethical Training and Feedback

The dual-blockchain architecture outlined above binds two ledgers into a single trust apparatus: an Ethical Ontology Blockchain that supplies duty-, virtue-, and consequence-based guidance, and a Physical Verification Blockchain that records what tangibly occurs. Their real-time exchange fuses moral intent with empirical proof, rendering every AI decision transparent, auditable, and accountable. Yet an infrastructure that stops at enforcement risks freezing in place while social expectations continue to move. To prevent ethical calcification, the design treats each logged decision—commendable or flawed—as fresh instructional data. Community-supplied labels flow back into model weights, reputation tallies, and even the rulebook itself, allowing both agents and governance logic to adjust in step with evolving human

values, all without surrendering the integrity of the audit trail. The mechanisms that follow describe how this continuous-learning loop turns the dual-chain substrate into a living, self-improving system of oversight.

Verifiable Labels on Behavior Traces

Every transaction published to the Physical Verification Blockchain can receive an ethical annotation. Automatic labels arise when (1) a deontological contract logs a rule violation or (2) a teleological contract posts a utility score. Human-supplied labels enter through a governance dApp that invites domain experts—or, in low-risk domains, public contributors—to review the complete state–action–outcome bundle and mark it *Ethically Acceptable*, *Concerning*, or with domain-specific tags (e.g., "bias detected," "minor inconvenience"). Each label is itself a signed on-chain event recording the reviewer's pseudonymous identity and weighting (Bhattacharya & Bryson, 2023). Consensus (for instance, nine of ten reviewers) finalizes the verdict, yielding an ever-growing, community-validated corpus of ethical ground truth (Köhn et al., 2022).

Distributed Training Data Pipeline

The state–action–outcome tuples—paired with their consensus labels—form a dataset exportable (under privacy-preserving transforms) to developers and researchers (Sharma & Solove, 2024). Suppose the network governs autonomous vehicles: thousands of real-world edge cases (such as *unprotected left turn with pedestrians jay-walking*) accrue along with adjudicated safety outcomes. Reinforcement-learning pipelines can substitute those on-chain ethical evaluations for proprietary reward functions, enabling decentralized RLHF at scale (Ouyang et al., 2022; Kaal, 2023). Model updates are first shadow-deployed under the same logging regime; only after demonstrating superior ethical performance in situ are the weights promoted to production, ensuring the runtime safety net remains even as reliance on it diminishes.

Community Governance and Rule Adaptation

Labels do not merely retrain models—they also signal when the rule set is deficient. Any validator may propose a contract amendment, citing specific on-chain incidents. Governance smart contracts open the proposal to debate, simulation, and token-weighted vote (Gorjão & West, 2023). If adopted, the new rule is versioned, conflict-checked by formal verification, and broadcast network-wide. Because past data remain immutable, stakeholders can replay logged scenarios under the amended rule to quantify marginal benefit or detect unintended side-effects before full activation.

Iterative Learning in Simulation and Deployment

Improved policies or models are first stress-tested against *digital twins* reconstructed from the blockchain ledger (Al-Ghaili et al., 2023). Controlled trials may then be whitelisted on-chain—granting time-boxed exceptions while every action remains fully logged. Should early results confirm safety and benefit, governance promotes the experimental artifact; otherwise, the exception auto-expires, aborting risky behavior without human scramble.

Global Knowledge Sharing

Because both ledgers are open to inspection (subject to jurisdictional privacy overlays), failure modes discovered in one locale become cautionary inputs everywhere else. The system thus accumulates a living library of ethical case law—complementing static incident repositories with empirical, machine-readable evidence (Wang & Reed, 2022).

In effect, the dual-chain governance framework recasts oversight as a pedagogical engine. Every logged infraction, crowd-sourced label, or utilitarian score becomes new training signal, continually refining both the rulebase and the models that heed it. Because ethical judgments are anchored in verifiable context—sensor traces, consensus labels, and on-chain

rationale—the system corrects behavior without ossifying it: fixed guardrails hard-block impermissible harms, while the virtue- and consequence-layers yield adaptive guidance that grows with experience. The resulting feedback loop is resolutely pluralistic: engineers supply technical fixes, domain experts contribute nuanced labels, and lay participants voice evolving social expectations—each intervention immutably recorded, auditable, and available for model retraining or rule revision. Governance thus stretches from high-level deliberation ("Should private-property duties be strengthened?") down to low-level weight updates in a navigation policy, unifying ethical discourse and machine learning within a single operational scaffold.

Early evidence from large-scale RL-from-human-feedback projects, together with preparedness initiatives (OpenAI, 2023) that stress real-world monitoring, underscores the value of such continuous, data-rich oversight. By institutionalizing those practices in an open, token-incentivized network, the framework offers a competitive advantage: agents that participate accrue richer training data, clearer reputations, and demonstrably safer performance, attracting yet more collaborators in a virtuous flywheel of *participation* → *feedback* → *alignment* → *trust*. Over time, the expectation is convergence—AI behavior that reliably tracks what diverse human communities deem acceptable, even as those standards shift. This architecture, therefore, operationalizes a paradigm shift: from static compliance to dynamic adaptation, where verification drives not only trust but also the iterative refinement of ethical AI systems in alignment with societal values.

Technical Implementation Anchors

Where ethical oversight becomes a site of learning, the architecture must evolve from static constraint to adaptive pedagogy. The prior section traced how the dual-blockchain framework converts each AI decision—encoded as a state-action-outcome tuple—into structured

feedback for iterative alignment. Through the integration of community-supplied labels, decentralized reinforcement learning, and responsive rule modification, the system advances not only compliance but maturation—fostering ethical behavior that tracks with the evolving values of diverse human communities, all while preserving an auditable chain of justification. Yet such dynamism is untenable without a technical substrate equal to its demands. High-frequency verification, cryptographic integrity, and multi-actor coordination require more than conceptual design; they necessitate a rigorously architected infrastructure. What follows delineates that operational backbone: a confluence of distributed systems, blockchain protocols, AI modules, and secure hardware—anchored by Hyperledger Fabric for ethical governance and Ethereum-compatible networks for empirical verification. Together, these components constitute the execution layer of the system's moral ambition: rendering fidelity and accountability not merely aspirational but enforceable, at scale.

Hyperledger Fabric as the Ethical Ontology Backbone

The Ethical Ontology Blockchain is best realized on Hyperledger Fabric, a permissioned, enterprise-grade distributed-ledger technology that accommodates complex chaincode written in general-purpose languages (Androulaki et al., 2018; Hyperledger Fabric, n.d.). Fabric's pluggable consensus—typically Raft or Istanbul BFT—allows a consortium of vetted stakeholders (universities, corporations, regulators) to ratify ethical-rule updates without the latency or anonymity of proof-of-work mining (Ben Toumia et al., 2021). Its Membership Service Provider issues cryptographic identities to every validator, thereby enabling governance votes to be restricted to recognized entities while still permitting public read-access for transparency. Fabric's channel architecture further supports the framework's multi-ontology requirement: culture- or sector-specific ethical ledgers can run on separate channels, all anchored

to a common root chain for interoperability (Zhou et al., 2019; Hyperledger Fabric, n.d.). Because high-frequency queries will flow against the ethical contracts, Fabric's demonstrated throughput on permissioned networks is essential. Finally, the modular runtime allows external theorem provers or bespoke logic engines to be linked directly into chaincode, enabling real-time evaluation of deontological rules that exceed simple conditional statements and ensuring logical consistency across the rulebase.

Ethereum-Compatible Infrastructure for Physical Verification

In contrast, the Physical Verification Blockchain benefits from the tooling maturity of an Ethereum-compatible environment. Whether deployed on Ethereum main-net, a dedicated side-chain such as Polygon, or an application-specific EVM network, an Ethereum substrate permits sensor devices and AI agents to submit signed "ActionLog" events through well-established wallets and SDKs (Gambhire et al., 2022). Existing identity standards—EIP-725 for autonomous agents or W3C decentralized identifiers (DIDs) for sensors—can be anchored natively. To reconcile the low-latency demands of IoT with the settlement finality of a public chain, an off-chain streaming layer (e.g., Apache Kafka or Redis Streams) will ingest signed payloads and batch-commit them on-chain, thereby reducing gas expenditure while preserving chronological ordering. Cross-chain queries between Fabric and Ethereum are mediated by Hyperledger Weaver (formerly Cactus), whose SDKs allow a Fabric chaincode function to request notarized sensor data from Ethereum and to receive cryptographic proof of its inclusion (Abebe et al., 2019; Hyperledger, 2021). This mechanism obviates the need to build a bespoke oracle network while still enabling provably secure data flows. For on-premises deployments, lightweight Redis channels can disseminate instantaneous

alerts—such as “action executed” or “sensor limit breached”—with asynchronous anchoring to blockchain, balancing real-time responsiveness against eventual immutability.

Smart-Contract Architecture and Formal-Verification Practices

Instead of a monolithic contract, the system employs a suite of specialized modules. Registry contracts record the public keys and metadata of AI agents, sensors, and human participants, issuing soul-bound tokens or verifiable credentials that attest to each role. Policy contracts encode the actual ethical rules, organized in composable libraries aligned with duty, virtue, and consequence. Reputation-token contracts implement the virtue-ethics layer through non-transferable scores that accrue—or decay—based on on-chain events. Logging contracts on the Ethereum tier structure high-volume IoT data and map large binary artifacts to IPFS or Filecoin by storing only their content hashes on-chain. Finally, oracle-bridge contracts escrow data and proofs as they transit between chains, employing multi-signature attestation to avoid single points of trust. Every contract that touches ethical logic will undergo formal verification—using tools such as Hyperledger Fabric Model Checker or Solidity’s SMT solvers—to preclude latent flaws that could undermine safety (Beckert & Herda, 2018). Rate-limiting guards, circuit breakers, and replay-protection patterns are added to withstand denial-of-service attempts or accidental overloads.

Hardware, IoT, and Secure-Signing Infrastructure

Edge devices rarely possess the resources to run full blockchain clients, so an architectural layer of IoT gateways (e.g., Raspberry Pi or industrial PC) aggregates local sensor feeds, signs them with embedded Hardware Security Modules, and forwards them in authenticated batches. Each gateway itself is registered on-chain with a public key that ties physical provenance (“sensor-123 on Reactor 5”) to digital identity. Message transport relies on

secure MQTT, whose payloads include device-level signatures that the receiving Ethereum contract verifies prior to log insertion. For higher-assurance domains, Trusted Execution Environments such as Intel SGX or ARM TrustZone sign periodic attestations of the AI runtime—statements like "model hash X produced output Y"—thereby linking internal inference to the external audit log (Shepherd & Markantonakis, 2024).

Data-Management and Scalability Strategies

City-scale deployments may generate thousands of events per second; storing every byte on-chain would be infeasible. Consequently, bulk data—images, LiDAR sweeps, high-rate telemetry—resides in decentralized storage (IPFS, Arweave, or Filecoin) while only a cryptographic root (e.g., a Merkle tree root) anchors each batch to the Physical chain (Khan et al., 2025). State-channel or roll-up techniques aggregate sensor readings off-chain, committing succinct proofs at configurable intervals. Future horizontal scaling can partition both ethical and physical ledgers by geography or sector and link them via cross-shard protocols. The overall architecture is therefore modular: individual components can be replaced—say, migrating from Raft to a newer BFT variant—without jeopardizing systemic integrity.

Interfaces, SDKs, and Experimentation Tooling

For human overseers, a browser-based dashboard, Evolution-of-Ethics Dash, presents real-time compliance status, pending governance votes, and drill-down access to sensor evidence. Developers receive language-native SDKs—Python, Rust, JavaScript—that encapsulate cryptographic signing and network calls: invoking ethics.check_action(plan) triggers an automatic query to Fabric and returns a signed verdict (Hyperledger Fabric, n.d.). The project ships with containerized simulation environments so that researchers can spin up a full dual-chain stack locally, inject synthetic agents, and observe governance dynamics before going

live. Custom block-explorer front-ends index not only transactions but also ethical rules, reputation scores, and cross-chain proofs, enabling auditors to filter, for example, “all invocations of Duty 001 that were denied during the past 24 hours.”

Concluding Technical Synthesis

Anchoring the system on Hyperledger Fabric and an Ethereum-compatible network harnesses the complementary strengths of permissioned governance and public immutability while Kafka/Redis pipelines and cryptographic hardware supply low-latency resilience. Each element—chaincode, oracle bridge, TEE attestation—has precedents in production blockchains, particularly in supply-chain traceability and industrial-IoT security. The engineering challenge therefore shifts from invention to disciplined integration, with interoperability managed through mature cross-chain frameworks and multi-signature oracles. By grounding AI oversight in these proven infrastructures, the design minimizes novel risk and accelerates the path to pilot deployments, which will be treated in the subsequent roadmap after first developing the accompanying governance philosophy.

Cooperative, Decentralized, and Composable Governance

The technical substrate outlined above provides the infrastructure necessary for secure, scalable, and auditable AI oversight. Yet no architecture, however robust, can govern on its own. The legitimacy and adaptability of the system must arise not only from code but from the collective agency of those who shape it. A core principle of this roadmap is that the system should be cooperative and decentralized by design. AI governance cannot be the domain of a single corporation or nation; it must draw upon the ethical intuitions and institutional insights of diverse cultures, sectors, and stakeholders. Moreover, the system must be composable—capable of integrating modular components and value frameworks—so that governance evolves through

self-organization rather than imposition. This section describes how the proposed framework operationalizes those principles.

Pluralism by Design

A central feature of the proposed governance framework is its principled accommodation of moral pluralism. There exists no universal consensus across human cultures regarding ethical norms, and attempts to enforce a monolithic value system would risk ethical hegemony and sociopolitical backlash (Friedman & Kahn, 2021; Floridi, 2018). To mitigate this, the Ethical Ontology Blockchain is architected to support modular ethical ontologies—distinct yet interoperable sets of smart contracts encoding specific moral traditions (Brennen & Kreiss, 2022).

These modules might include, for example, a categorical imperative (Kantian) core shared across agents, supplemented by culturally or domain-specific modules (i.e., hypothetical imperatives) such as an Islamic legal-ethical framework, a Buddhist virtue layer, or a Western biomedical ethics library. Agents and organizations select the modules most relevant to their operational or cultural context, and this selection is rendered transparent on-chain. Importantly, these modules are composable: rather than creating ethical silos, they interoperate through shared definitional primitives (e.g., harm, benefit, obligation, consent).

When agents with differing ethical configurations interact, the system defaults to the intersection of their moral constraints, ensuring that joint actions respect the strictest applicable conditions (Arnold et al., 2017). Alternatively, agents may negotiate a shared ethical pathway via a meta-contract that references multiple ontologies and resolves divergences through satisfiability logic or predefined arbitration rules. Governance mechanisms permit designation of

certain modules as "core" (universally required) and others as "optional" (contextually selectable), providing a flexible yet coherent scaffolding.

This composability functions analogously to software libraries with standard APIs: a common substrate of terms and procedures supports interoperability, while specific implementations remain modular. Genuine alignment of AI behavior with human values requires cultural sensitivity and participatory design (World Economic Forum, 2022; Dignum, 2019). The present framework enables precisely that—embedding continuous stakeholder input in an expandable architecture that evolves through use and deliberation. Over time, modules may converge or coalesce, forming de facto standards without coercive imposition. The result is an ethical federation, not an empire: a web of jurisdictional alignments, each context-aware yet translatable across systems.

Decentralized Autonomous Organization (DAO) Governance

The governance of the framework itself adheres to the same principles it seeks to encode: transparency, pluralism, and decentralization. Rather than relying on a centralized regulatory authority or proprietary corporate structure, the system's core rule-making and oversight processes are managed by a decentralized autonomous organization (DAO).

In this model, governance tokens or roles are distributed across major stakeholders—potentially including nation-states, technology firms, academic institutions, standards bodies, and community representatives. These entities participate freely and voluntarily in decision-making processes such as approving new ethical modules, modifying system-wide parameters, or resolving rule conflicts. All proposals and votes occur on-chain, creating an immutable, auditable record of governance activity and institutional memory (Hassan & De Filippi, 2021).

Inspired by successful precedents in DAO-based governance—such as Ethereum's protocol upgrades or community-managed treasuries—the system can implement constitutional mechanics: foundational documents drawn from frameworks such as UNESCO's *Ethics of Artificial Intelligence* or the IEEE's *Ethically Aligned Design* can define high-level principles, constraints, and rights. Procedural rules—quorum requirements, multi-tier voting thresholds, role-based checks and balances—would differentiate between operational updates and foundational ethical amendments. A bicameral governance structure may be employed, where one chamber is composed of technical or domain experts, and another of broader stakeholder or public representatives, with consensus required across both for high-impact changes (Allen, O'Hara, & Hall, 2020).

Moreover, the DAO may steward a treasury to fund security audits, system maintenance, human-in-the-loop annotation, and other essential services. Contributions could be sourced from participants, grant-making bodies, or public funding mechanisms. Bounty programs and open calls for contributions ensure the system remains a commons—a public infrastructure under continuous participatory construction rather than a static, elite-controlled apparatus.

Self-Organization and Local Autonomy

A core implication of decentralized design is the redistribution of decision-making authority to the periphery, enabling self-organization without severing systemic coherence. The framework is built to support localized governance regimes—sub-communities that operate semi-independently while remaining interoperable with the global ethical infrastructure. For example, a federation of hospitals deploying medical AI agents might instantiate a domain-specific sub-DAO tasked with curating and updating ethical modules specific to

biomedical contexts. These sub-DAOs maintain a direct governance bridge to the overarching system but retain authority over domain nuances that do not require global consensus.

This principle also extends to jurisdictional governance. National or regional regulators may define additional normative constraints by authoring localized smart contracts. These contracts are tagged with jurisdictional metadata, ensuring that any AI operating within a given territory must comply not only with global baseline standards but also with those encoded in the "Country X" module, governed by locally authorized validators. This architecture operationalizes the principle of subsidiarity: decisions are made at the most immediate level competent to handle them (Cath, 2018; Floridi, 2018).

Despite their autonomy, local governance nodes contribute to and benefit from the broader ethical federation. Lessons learned, incident records, and validated models are shared across the network, yielding a governance analog to federated learning. Each instance functions as a composable unit—connectable or disconnectable without jeopardizing global integrity. This architectural modularity supports political adoption by respecting sovereignty while still offering the advantages of globally harmonized AI oversight. National AI strategies can therefore be strengthened, not supplanted, by participation in this system.

Cooperative Incentives

Decentralized participation requires more than infrastructure—it depends on motivation. The framework incorporates a multi-layered incentive structure to promote cooperative behavior, sustained engagement, and ethical alignment across participants (Buterin et al., 2019). Tokenomics mechanisms, for instance, may be employed to compensate validators, auditors, and contributors who uphold the integrity of the system. Participants who provide recurring, high-quality oversight—such as flagging policy violations, submitting formal proposals, or

conducting model evaluations—might accrue governance tokens or enhanced reputational weight within the DAO.

The incentive structure also extends to AI agents. While direct monetary rewards for ethical behavior may introduce perverse incentives or vulnerability to gaming, the system supports non-fungible reputation tokens that reflect sustained virtue in decision-making—enabling trust-building without commodification. More tangibly, participation in the network offers reputational and economic advantages. Developers can advertise their systems as "ethics-certified" by virtue of on-chain compliance logs. Organizations avoid the overhead of developing proprietary monitoring infrastructures by leveraging the shared governance backbone for auditing and compliance.

Moreover, the governance process itself can be structured to recognize and amplify contributions. Voting weight may be calibrated using mechanisms such as proof-of-expertise—where entities with a demonstrable history of safety contributions (e.g., AI research institutes, standards bodies) gain proportional influence over technical or high-impact proposals. To avoid dominance by elite actors, alternative governance models such as quadratic voting or conviction voting may be employed to balance expertise with democratic input (Zhang et al., 2021; Buterin, Hitzig, & Weyl, 2019). In sum, the incentive framework is designed not merely to reward participation but to cultivate sustained alignment among those maintaining, deploying, and regulating AI systems.

Alignment with Existing Initiatives

Decentralization need not imply divergence. Rather than standing in opposition to existing regulatory frameworks, the proposed governance architecture is designed for interoperability with global and regional initiatives. To that end, the system accommodates

alignment with authoritative policy instruments such as the EU AI Act, the OECD AI Principles, and various industry-led ethical standards. These external guidelines can be formally instantiated as smart contracts within the Ethical Ontology Blockchain, thereby operationalizing regulatory intent in code (Veale & Borgesius, 2021). For instance, if European legislation mandates risk-management systems for high-risk AI applications, this architecture could fulfill that requirement through embedded monitoring and verification functions (Veale & Borgesius, 2021). Regulators, in turn, may be granted observer-node access, enabling real-time auditing of compliance rather than relying on post hoc reporting.

Such integration reframes regulatory agencies from external enforcers into active participants—stakeholders who can interface directly with the governance process. By embedding these legal and normative requirements into the blockchain, we provide a public, tamper-evident record of compliance that reinforces trust without duplicating oversight efforts. This compatibility also extends to the research community. For example, DeepMind's alignment research has emphasized the importance of value pluralism and scalable oversight—principles this framework concretely enacts (Leike et al., 2018). Similarly, OpenAI's initiatives around adversarial testing and safety evaluations can be absorbed as validation modules or rule proposals within the system. In positioning itself as the infrastructure through which such efforts may be tested, refined, and deployed, the framework offers not a rival, but a vessel: one that channels disparate efforts into a cohesive operational reality.

Evolutionary Expansion

As the network proves itself—perhaps beginning with applications in sectors such as healthcare or autonomous vehicles—it must be capable of growing both in scope and membership. The governance model allows new participants (e.g., companies, regional

consortia, national bodies) to be admitted through defined protocols, such as majority votes following vetting procedures. New ethical frameworks may also be proposed and integrated: if a previously underrepresented worldview or philosophy emerges, it can be formalized as a rule module and, upon community approval, deployed to the network. Composability ensures these additions behave as plug-ins—optional, interoperable, and gradually influential if broadly adopted. This allows the ecosystem to evolve, not just in content but in process. Meta-governance capabilities permit the DAO to revise even its own procedures, ensuring flexibility in the face of novel challenges or governance bottlenecks (De Filippi & Mannan, 2020). Reflexivity of this kind—changing how change occurs—is indispensable for long-term resilience, especially as AI systems and ethical expectations continue to shift.

To ground this, consider a forward-looking scenario: in 2028, a consortium of African AI startups proposes a set of smart contracts and ethical principles rooted in Ubuntu philosophy, emphasizing community wellbeing and mutual aid. Through the DAO, these proposals are deliberated and piloted in local deployments. After validation and refinement, the global governance layer recognizes their value and integrates the module into the broader ontology. Now, AI agents that adhere to the "Ubuntu module" can be preferentially selected in relevant jurisdictions or domains, enriching the system's diversity and responsiveness. Such an evolution demonstrates how ethical pluralism becomes operationally viable through decentralized, composable design.

The system's governance architecture—decentralized to avoid centralized capture, cooperative to leverage collective insight, and composable to accommodate diverse values—is not merely an operational layer but a normative proposition. It treats ethics as a living commons and oversight as a participatory infrastructure. Rather than consolidating control under a single

authority or relying on corporate self-regulation, the approach mirrors successful societal institutions like peer review or democratic deliberation: iterative, inclusive, and resilient. In this light, AI governance becomes not just a technical necessity but a shared platform—co-owned by the communities it seeks to serve.

Complementarity with Existing AI Safety and Governance Initiatives

While the previous section outlined how governance must be decentralized, composable, and pluralistic to achieve legitimacy and adaptability, this next section turns outward—to the broader ecosystem of AI safety. No architecture exists in a vacuum. For such a framework to flourish, it must not only be internally robust but also externally interoperable with the policies, principles, and technical agendas already underway. It is therefore essential to situate this roadmap within the constellation of ongoing AI governance and safety efforts. Though novel in architecture and execution, the framework aligns substantively with the aims of existing regulatory, academic, and industrial initiatives. Rather than competing, it offers an infrastructural substrate—decentralized, verifiable, and adaptive—that can unify these efforts into an integrated, operational whole. What follows identifies key areas of synergy.

OpenAI and Frontier Model Governance

OpenAI's *Preparedness Framework* and associated efforts to govern frontier AI models underscore the importance of rigorous evaluation, real-world oversight, and proactive containment of emerging risks (Shevlane et al., 2023). While these initiatives represent some of the most robust internal safeguards proposed to date, they remain constrained by institutional boundaries and trust assumptions. The framework outlined in this roadmap offers a complementary, externally verifiable substrate: a decentralized system of validators and physical-world logging that operates independently of any single developer's control.

Under such an arrangement, OpenAI's models could be instantiated within the dual-blockchain architecture, where ethical compliance and behavioral anomalies are continuously monitored by a consortium of third parties. For instance, the Physical Verification Blockchain could log real-time sensor and inference data, flagging deviations or emergent capabilities as potential signals of unsupervised self-modification—a scenario of particular concern in superintelligence discourse (Hawthorne & Amodei, 2023). In this sense, the blockchain does not replace OpenAI's internal red teaming or alignment research; rather, it extends their visibility and accountability into a decentralized public domain. It transforms internal assertions of safety into externally auditable claims, offering both reassurance to regulators and legitimacy before a skeptical public.

DeepMind's Alignment Agenda

Google DeepMind has advanced a suite of conceptual frameworks for AI alignment, including *scalable oversight*, *value pluralism*, and recursive feedback between AI agents and human supervisors (Gabriel, 2020). These approaches share core philosophical commitments with our system, particularly the notion that AI behavior must remain anchored to diverse human values, dynamically interpreted, and continuously adjusted.

Within our proposed architecture, these ideals are not merely theoretical but are directly implemented in the Ethical Ontology Blockchain. For example, scalable oversight becomes operational through automated "watchdog" agents—AI monitors empowered to query ethical contracts, flag anomalies, and log infractions in real time. Likewise, DeepMind's research on "Voices of All in Alignment," which seeks mechanisms to incorporate cross-cultural and multi-stakeholder input, finds concrete expression in our modular ontology structure. Each cultural or institutional node can propose and govern its own ethical module while interoperating

with others—enabling the pluralism DeepMind advocates to be implemented at scale (Arnold et al., 2021).

Furthermore, the system creates a research loop: alignment researchers can encode tentative formulations of complex values (e.g., fairness, dignity, solidarity) into smart contracts, observe their performance across diverse deployment contexts, and revise their definitions based on empirical feedback. In this way, alignment theory and on-chain practice form a mutually reinforcing dialectic—a proving ground where philosophical clarity meets operational complexity.

Anthropic's Constitutional AI and Dynamic Ethical Enforcement

Anthropic's Constitutional AI model presents a training-time approach to embedding ethical principles within language models (Bai et al., 2022). These constitutions—often inspired by documents like the UN Declaration of Human Rights or other liberal-democratic norms—function as fixed guides to behavior during supervised learning and fine-tuning. While effective in shaping model outputs, this method retains a static character; the principles must be predetermined and cannot easily evolve with stakeholder input.

Our Ethical Ontology Blockchain introduces a complementary dynamic: it externalizes the constitution and encodes it in mutable, auditable smart contracts that can be contested, amended, and ratified over time. A natural integration emerges. Models trained under Constitutional AI could be deployed into our framework with minimal adaptation, as their prior alignment would predispose them toward rule compliance. Conversely, our DAO-governed contract ecosystem could feed back real-world updates to the constitutional principles Anthropic uses for future training cycles. The result is a co-evolutionary system: the ethical expectations of

society evolve in real time, and the training paradigms of frontier models evolve with them—not in isolation, but in tandem.

Policy and Standards Organizations

International and industry-specific standard-setting bodies—such as the ISO/IEC Joint Technical Committee on Artificial Intelligence, the IEEE's 7000-series initiatives on ethical AI, and the Partnership on AI—have increasingly emphasized the codification of ethical principles into actionable guidelines (ISO/IEC, 2023). Our framework is well-positioned to serve as a reference implementation for many of these evolving norms. For instance, IEEE Standard 7001 specifies requirements for system transparency—a feature inherent to the architecture of the blockchain itself, which offers immutable, publicly inspectable records of AI behavior and decision logic (IEEE, 2022).

By aligning our smart-contract metrics and rule categories with the conceptual vocabularies adopted by these organizations—fairness, accountability, explainability, non-discrimination—we render those standards not merely declarative but enforceable. This alignment provides a pragmatic pathway for adoption: rather than requiring companies to develop bespoke compliance infrastructure, the framework offers a ready-made governance network that fulfills regulatory mandates. Participation in such a network could satisfy procurement requirements or serve as a certification of ethical compliance, particularly in high-risk domains like healthcare, finance, or critical infrastructure. In time, regulatory agencies might even require that advanced AI systems be subject to continuous monitoring via blockchain-based governance, paralleling how financial institutions are obligated to maintain auditable records for oversight and risk assessment.

Research on Decentralized AI Governance

A burgeoning academic literature now explores the role of decentralized technologies in governing autonomous systems. Recent proposals, such as the ETHOS architecture for AI DAOs, agent registries, and soulbound identity tokens, articulate theoretical models for multi-agent governance in trustless environments (Xu et al., 2022; Reijers et al., 2021). Our system operationalizes these ideas, offering a concrete testbed where such governance structures can be deployed, monitored, and refined in situ.

Crucially, we incorporate technical affordances proposed by this body of research, including decentralized identity (DID) protocols for agents, soulbound tokens to signify compliance status, and zero-knowledge proofs to validate properties of AI models without disclosing proprietary internals. This fidelity to current science positions the platform not only as a governance mechanism but as a living laboratory: researchers may use it to deploy experimental agents under controlled governance regimes or trial new rule systems to study their behavioral effects. In this way, the framework functions as both infrastructure and scientific apparatus, bridging normative theory and empirical practice while sustaining engagement from leading AI governance conferences and journals.

Compatibility with Secure AI Architectures

Recent proposals by Google and others have introduced cybersecurity-style frameworks for AI model protection—most notably the Secure AI Framework (SAIF), which targets adversarial threats such as model theft, poisoning, and inference manipulation (Bursztein, 2023; Papernot et al., 2021). While SAIF emphasizes system robustness and defensive hardening, our framework extends assurance to the behavioral and ethical domains. Together, they form a complementary stack: one ensures the system cannot be corrupted, while the other ensures it cannot go morally astray.

Importantly, these layers can interact. The Physical Verification Blockchain, by logging observed outcomes and ethical infractions, may serve as an early warning system for security breaches. A sudden spike in non-compliant behavior across a model's activity log could indicate tampering, drift, or external compromise—data which could trigger protocol-level safeguards or invoke investigation by security monitors. As such, the boundary between AI safety and AI security becomes porous, and coordination between ethics-oriented and security-oriented stakeholders becomes not just beneficial but operationally necessary. The joint deployment of these architectures may yield a defense-in-depth strategy for AI reliability, incorporating both cryptographic protection and ethical accountability in a unified framework (Brundage et al., 2020).

Rather than functioning as a rival to existing initiatives, this framework serves as an infrastructural substrate capable of translating AI safety principles into operational practice. By integrating and reinforcing the intent behind efforts such as OpenAI's Preparedness Framework, DeepMind's alignment research, IEEE 7000-series standards, and the EU AI Act, the system avoids redundancy and instead amplifies coherence across the governance landscape. The ethos might be captured as follows: "From principles to practice, via provable blockchain trust." Just as financial integrity depends not only on standards but on auditable systems to uphold them, so too must ethical AI rely on transparent, enforceable architectures to carry normative commitments into applied domains.

This governance architecture is best understood not as a competing alternative in the growing ecosystem of AI ethics, but as connective infrastructure—binding research, regulation, and deployment into a cohesive whole (Morley et al., 2021). Institutions such as MIT CSAIL, policy advisory bodies, corporate ethics teams, and international standard-setting organizations

are likely to recognize this roadmap as a logical progression: a platform that elevates ethical discourse from conceptual articulation to decentralized, verifiable implementation across real-world systems.

Conclusion and Roadmap Phases

Having situated the proposed framework in relation to leading AI safety agendas and institutional governance efforts, the focus now shifts to operational synthesis and implementation. The roadmap presented here offers a comprehensive vision for decentralized ethical AI governance and real-world data verification—an approach that draws structural inspiration from Google's BeyondCorp by reimagining trust not as a perimeter but as a dynamic, continuously verified relationship. By anchoring ethical reasoning in a dual-blockchain architecture, the system unifies normative commitments with empirical accountability, enabling AI agents to act within transparent, socially grounded constraints while being held answerable to facts on the ground. What emerges is not a static compliance protocol but a living infrastructure—open, composable, and reflexive—through which human communities and autonomous systems co-develop ethical norms in a decentralized yet coherent ecosystem. To move from concept to deployment, this vision now unfolds in a set of phased implementation strategies.

The preceding sections have situated this proposal within the broader ecosystem of AI safety and governance efforts, demonstrating its capacity to serve as connective tissue across diverse initiatives. In light of that integrative potential, the final step is to chart a concrete path forward—one that translates architectural vision into actionable stages.

To carry this vision from blueprint to deployment, the following phased implementation strategy sets out the concrete steps ahead.

Phase 1 – Prototype and Pilot (Year 1–2)

The inaugural phase centers on the controlled deployment of a functional prototype in a high-stakes domain with active, invested stakeholders. A hospital network, for instance, might implement an early version of the Ethics Dashboard 2.0 to oversee an AI system responsible for clinical decision support, such as diagnostic recommendations or pharmaceutical dispensation (Delgado-Sigüenza et al., 2023). In this constrained setting, a simplified instantiation of the dual-ledger architecture would be deployed—tracking a minimal but critical set of ethical rules (e.g., patient consent verification or secondary medical review requirements) in parallel with empirical data (e.g., sensor logs tied to medication dispensation events).

The objective is twofold: first, to validate the technological substrate—verifying that smart contracts execute as intended and that sensors generate immutable, verifiable logs; second, to evaluate the system's usability for clinical staff and technical operators. An academic partner should be engaged to conduct retrospective analysis on the logged data, aiming to demonstrate that blockchain-enabled observability can detect, explain, and facilitate the remediation of edge-case violations (Engelmann & Chen, 2022). A successful outcome would involve the system flagging a policy breach (e.g., treatment recommended without mandatory second opinion), prompting human intervention, with the resolution feeding back into updated rule logic or model parameters.

Phase 2 – Consortium Formation and Minimum Viable Network (Year 2–3)

Building on Phase 1, the second phase initiates formal institutional alignment around the framework through the formation of a multi-stakeholder consortium. This consortium might include hospitals, AI development firms, IoT device manufacturers, and ethics research institutions. Together, they would co-deploy a minimum viable governance network:

Hyperledger Fabric would host the ethical rulebase, with consortium members serving as validator nodes (Albrecht et al., 2022), while an Ethereum testnet would log device and environmental data.

This phase involves the introduction of elementary governance functionality—on-chain voting, access controls, and smart contract proposal mechanisms—and the codification of 5 to 10 ethical contracts representing a broader range of domain-specific constraints. Integration of multiple AI systems and expansion to additional facilities or geographic sites should be initiated. To promote developer adoption, a standardized API and software development kit (SDK) should be released, abstracting cryptographic operations and network communication. Stress testing under simulated high-volume loads, security penetration exercises, and the application of privacy-preserving measures (e.g., anonymization layers or zero-knowledge proofs) will be crucial to assess robustness. The expected deliverables include a functioning "Ethical AI Network" at modest scale, a validated onboarding playbook for new participants, and preliminary evidence of improved institutional trust—for instance, increased confidence in AI outputs as measured through user surveys at deployment sites.

Phase 3 – Multi-Domain Expansion and Open Participation (Year 3–5)

Phase 3 transitions from controlled pilots to broader, cross-domain participation. In parallel with the healthcare use case, the network might be extended to additional sectors—such as urban mobility (e.g., autonomous vehicle coordination), media and content moderation (e.g., misinformation traceability), or education technology (e.g., adaptive learning systems). This phase tests the system’s composability: new ethical modules are instantiated to reflect domain-specific principles, and interoperability between multiple Fabric networks may be facilitated via cross-chain protocols like Hyperledger Cactus.

As institutional confidence grows, regulatory participation should expand. Engagement with formal oversight bodies—such as participation in the European Union's AI regulatory sandbox—can demonstrate how the architecture fulfills forthcoming compliance mandates through cryptographically verifiable mechanisms (European Commission, 2023). This period also marks the formalization of protocols through the publication of white papers, technical documentation, and possibly an open industry standard. By the close of Phase 3, the network should span dozens of organizations and hundreds of AI agents and devices, operating in safety-critical contexts. Importantly, the system should exhibit early reflexivity—demonstrated, for example, by a successful vote to amend an ethical rule based on empirical lessons derived from network activity.

Phase 4 – Global Cooperative Network (Year 5+)

With the foundational infrastructure validated and multi-domain functionality established, Phase 4 aims at global scale and federated governance. Regional hubs across North America, Europe, Asia-Pacific, and Africa would instantiate linked but semi-autonomous deployments, tailored to local laws, languages, and cultural ethical frameworks (Cihon et al., 2020). At this stage, governance decentralization becomes paramount: eligibility criteria may be codified for broader node participation, potentially leveraging tokenomics to balance decentralization with identity and reliability guarantees.

Interoperability with legal and institutional identity frameworks becomes central—enabling on-chain credentials to reflect not just technical authorization but also legal rights (e.g., national citizenship, organizational liability, medical licensure) (Kshetri & Voas, 2021). Strategic partnerships with transnational bodies—such as the United Nations, World Economic Forum, or regional AI alliances—should be pursued to secure endorsement and

promote harmonization. Integration into national AI infrastructures or large-scale public-private collaborations may follow. By this phase, the ethical ontology library will have matured to encompass a vast plurality of normative systems, while the feedback loop of continuous learning and rule evolution should yield decreasing rates of critical incidents. In effect, the system will have transitioned from proof-of-concept to public infrastructure—serving as a backbone for scalable, trustworthy AI in the global commons.

Throughout each phase of implementation, the framework places evaluative rigor and iterative adaptation at its core (ISO/IEC, 2022). The guiding inquiry remains consistent: does this infrastructure measurably reduce adverse outcomes? Does it meaningfully enhance stakeholder trust and institutional confidence? Are the implementation costs—technical, economic, and procedural—justified by the integrity and assurance it delivers? Such metrics are not merely retrospective but directive; they shape system evolution. In cases where architectural or governance components fail to meet expectations, the framework's decentralized and modular constitution allows for agile recalibration, embodying the very adaptability it seeks to instill in AI systems.

As the roadmap reaches maturity, the aspiration is no longer conceptual but normative: it becomes standard practice for high-impact AI systems to be woven into a global trust fabric—one in which every decision is subject to transparent, verifiable ethical evaluation. The result is not just regulatory compliance, but civic reassurance. Humans commissioning AI for complex societal tasks—autonomous mobility, medical diagnostics, energy grid optimization—can do so with confidence grounded in continuous, cryptographically secured oversight. Autonomous systems, likewise, can interoperate within a shared ontological language and epistemic baseline, recognizing that norm violations are neither obscure nor inconsequential.

In an increasingly autonomous world, this infrastructure functions as both conscience and immune system: detecting misalignment before it metastasizes, and inoculating agents with the evolving norms of an ethically pluralistic society.

Realizing such a vision requires sustained collaboration across disciplinary boundaries—linking technologists, ethicists, policymakers, and the public in a durable epistemic alliance. It is precisely this convergence that constitutes the method's strength. In the spirit of BeyondCorp's transformation of security architecture—from perimeter-based assumptions to context-aware verification—this roadmap proposes an analogous shift for AI ethics: from siloed principles to distributed enforcement; from blind trust to demonstrable trust (Ashmore & Nadimpalli, 2019). The ambition is not to centralize control but to coordinate shared accountability—rendering AI not as an opaque agent of uncertainty, but as a transparent and intelligible collaborator whose actions and justifications are inscribed in a ledger of public reason (Floridi et al., 2020).

In the end, the true measure of success will not be the system's technical elegance, but its practical impact: that when an AI system acts upon the world, one may ask, "Can it be trusted?"—and answer, without hesitation, "Yes—and here is the proof."

References

Abebe, E., Behl, D., Govindarajan, C., Hu, Y., Karunamoorthy, D., Novotny, P., ... & Vecchiola, C. (2019). Enabling enterprise blockchain interoperability with trusted data transfer. [Preprint]. arXiv:1911.01064.

Accountable Tech, AI Now Institute, & EPIC. (2023). *Zero Trust AI Governance* (White paper).

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*, Article 1568. https://doi.org/10.1057/s41599-024-02795-3.

Akintunde, M., Yazdanpanah, V., Salehi Fathabadi, A., Cîrstea, C., Dastani, M., & Moreau, L. (2024, May). *Actual trust in multi-agent systems (extended abstract).* Paper presented at the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand. https://ifaamas.org/Proceedings/aamas2024/pdfs/p2114.pdf

Akther, A., Arobee, A., Adnan, A. A., Auyon, O., Islam, A. J., & Akter, F. (2025). *Blockchain as a platform for artificial intelligence (AI) transparency (Version 1)* [Preprint]. arXiv. https://arxiv.org/abs/2503.08699

Albrecht, S., Rejeb, A., & Treiblmaier, H. (2022). Consortium blockchains in highly regulated industries: Governance and trust formation. *Computers & Industrial Engineering, 170,* 108323. https://doi.org/10.1016/j.cie.2022.108323

Allen, C., O'Hara, K., & Hall, W. (2020). Governance of blockchain systems: A typology of decentralized consensus mechanisms. *Philosophy & Technology, 33*(2), 273-303. https://doi.org/10.1007/s13347-019-00360-1

Al-Ghaili, S., Lu, H., & Radu, V. (2023). Ledger-to-twin: Replaying blockchain-logged cyber-physical traces in high-fidelity simulation. *ACM Transactions on Cyber-Physical Systems, 7*(4), 39:1–39:25. https://doi.org/10.1145/3601122

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., ... & Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains.[Preprint]. arXiv:1801.10228.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). *Value alignment or misalignment—What will keep systems accountable? Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 171-176). AAAI Press.

Arnold, T., Kasenberg, D., & Scheutz, M. (2021). Value alignment or misalignment—What will keep systems accountable? *Ethics and Information Technology, 23*(1), 71–84. https://doi.org/10.1007/s10676-020-09565-0

Arnold, T., & Scheutz, M. (2022). *Practical runtime monitoring for machine ethics*. In J. F. Bonnefon & B. Kim (Eds.), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2022* (pp. 542–550). ACM. https://doi.org/10.1145/3514094.3534147

Ashmore, J., & Nadimpalli, V. K. (2019). BeyondCorp: Designing zero-trust networks at scale. *IEEE Cloud Computing, 6*(5), 46–54. https://doi.org/10.1109/MCC.2019.2916033

Bai, Y., Kadavath, S., Kundu, S., Askell, A., & Gao, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. [Preprint] arXiv:2204.05862.

Beckert, B., & Herda, M. (2018). Formal Specification and Verification of Hyperledger Fabric Chaincode. *Semantic Scholar.* Retrieved from https://www.semanticscholar.org/paper/Formal-Specification-and-Verification-of-Fabric-Beckert-Herda/76b9b629a1771df5cf3a8ae6ae7339cece28cd41

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management, 35*(2), 530–549.

Ben Toumia, S., Berger, C., & Reiser, H. P. (2021). Evaluating blockchain application requirements and their satisfaction in Hyperledger Fabric. [Preprint] arXiv:2111.15399.

Bhattacharya, A., & Bryson, J. J. (2023). Blockchain annotation markets for auditing autonomous-system conduct. Journal of Responsible Technology, 14, 100102. https://doi.org/10.1016/j.jrt.2023.100102

Brennen, S., & Kreiss, D. (2022). Modular ethics for artificial intelligence: The case for interoperable moral architectures. *AI and Ethics, 2*(3), 517-528. https://doi.org/10.1007/s43681-021-00129-y

Brundage, M., Avin, S., & Clark, J. (2020). Toward trustworthy AI: Mechanisms for supporting verifiable claims. *Proceedings of the National Academy of Sciences, 117*(30), 17655–17656. https://doi.org/10.1073/pnas.2012854117

Bursztein, E. (2023). Toward a secure AI framework: Protecting machine-learning systems like we secure software. *Communications of the ACM, 66*(7), 34–36. https://doi.org/10.1145/3594515

Buterin, V., Hitzig, Z., & Weyl, E. G. (2019). Liberal radicalism: A flexible design for philanthropic matching funds. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3243656

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A, 376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Chen, Z., & Wang, X. (2021). Secure cross‑chain oracles for verifiable AI governance. *Journal of Parallel and Distributed Computing, 156,* 37‑50. https://doi.org/10.1016/j.jpdc.2021.05.005

Cihon, P., Maas, M. M., & Kemp, L. (2020). Should artificial intelligence governance be centralised? Design lessons from history. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 228–233. https://doi.org/10.1145/3375627.3375879

De Filippi, P., & Mannan, M. (2020). Blockchain and the law: A critical evaluation. *Stanford Journal of Blockchain Law & Policy, 3*(1), 1-23.

Delgado-Sigüenza, J., Jiménez-Navarro, M., & Soria-Olivas, E. (2023). Blockchain-enabled auditing for clinical-decision AIs: A pilot study in tertiary hospitals. Journal of *Biomedical Informatics, 140*, 104341. https://doi.org/10.1016/j.jbi.2023.104341

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way.* Springer.

Engelmann, S., & Chen, J. Y. (2022). Explainable AI auditing in healthcare: Lessons from a multi-site blockchain trial. npj *Digital Medicine, 5*(1), 87. https://doi.org/10.1038/s41746-022-00617-1

European Commission. (2023). *Report on the first cohort of the EU AI regulatory sandbox.* Publications Office of the European Union.

Ferrari, V., & Leutert, F. (2023). Formal‑verification pipelines for deontic smart contracts in ethical AI governance. *IEEE Transactions on Artificial Intelligence, 4(2),* 167‑181. https://doi.org/10.1109/TAI.2023.3278451

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*(1), 1-8. https://doi.org/10.1007/s13347-017-0263-2

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2020). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, B., & Kahn, P. H. Jr. (2021). *Human values, ethics, and design.* In A. Sears & J. A. Jacko (Eds.), *The human–computer interaction handbook* (4th ed., pp. 1223-1248). CRC Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gambhire, G., Gujar, T., & Pathak, S. (2022). An Extensive Blockchain Based Applications Survey: Tools, Frameworks, Opportunities, Challenges, and Solutions. *Academia.edu.* Retrieved from https://www.academia.edu/93611404/An_Extensive_Blockchain_Based_Applications_Survey_Tools_Frameworks_Opportunities_Challenges_and_Solutions Academia

Gorjão, M., & West, S. M. (2023). Delegated morality: Token-curated registries for updating AI governance norms. *Regulation & Governance*. Advance online publication. https://doi.org/10.1111/rego.12476

Gudipati, P., Ranganathan, S., & Kim, H. (2023). Layer‑2 roll‑ups for large‑scale IoT telemetry: A performance and security analysis. *IEEE Internet of Things Journal, 10*(9), 7906‑7918. https://doi.org/10.1109/JIOT.2023.3241189

Hawthorne, J., & Amodei, D. (2023). Monitoring advanced AI systems for unexpected behavior. *AI Magazine, 44*(1), 56–69.

Hassan, S., & De Filippi, P. (2021). Decentralized autonomous organization. *Internet Policy Review, 10*(2), 1-29. https://doi.org/10.14763/2021.2.1556

Huang, Y., Yen, I.-L., & Bastani, F. (2024). Pruning blockchain protocols for efficient access control in IoT systems. In *Proceedings of the IEEE International Conference on Blockchain and Cryptocurrency* (pp. 1–8). IEEE. https://arxiv.org/abs/2407.05506

Hyperledger. (2021, June 9). Meet Weaver, one of the new Hyperledger Labs taking on cross-chain and off-chain operations. *LF Decentralized Trust.* Retrieved from https://www.lfdecentralizedtrust.org/blog/2021/06/09/meet-weaver-one-the-new-hyperledger-labs-projects-taking-on-cross-chain-and-off-chain-operations

Hyperledger Fabric. (n.d.). Introduction. *Hyperledger Fabric Documentation.* Retrieved from https://hyperledger-fabric.readthedocs.io/en/latest/whatis.html

IEEE. (2022). *IEEE Standard 7001-2021: Transparency of autonomous systems.* IEEE Standards Association.

ISO/IEC. (2023). *ISO/IEC 42001:2023—Artificial intelligence management system.* International Organization for Standardization.

ISO/IEC. (2022). ISO/IEC TR 24028:2022—*Artificial intelligence—Overview of trustworthiness in AI.* International Organization for Standardization.

Kaal, W. A. (2023). *How AI models are optimized through Web3 governance* (SSRN Working Paper No. 4352777). https://doi.org/10.2139/ssrn.4352777

Kenesei, Z., Ásványi, K., Kökény, L., Jászberényi, M., Miskolczi, M., Gyulavári, T., & Syahrivar, J. (2022). Trust and perceived risk: How different manifestations affect the adoption of autonomous vehicles. *Transportation Research Part A: Policy and Practice, 164*, 379–393.

Khan, M. M., Khan, F. S., Nadeem, M., Khan, T. H., Haider, S., & Daas, D. (2025). Scalability and Efficiency Analysis of Hyperledger Fabric and Private Ethereum in Smart Contract Execution. *Computers, 14*(4), 132. https://doi.org/10.3390/computers14040132

Kshetri, N., & Voas, J. (2021). Blockchain-enabled e-ID management: Prospects for developing economies. *IEEE Software, 38*(5), 91–96. https://doi.org/10.1109/MS.2020.3042117

Köhn, D., Cascón-Porcel, J., & Balestrieri, G. (2022). Decentralised truth-maintenance for AI safety: A majority-vote oracle on Ethereum. *Proceedings of the AAAI Workshop on Blockchain and AI*, 29–38.

Leike, J., Krakovna, V., Ortega, P., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. [Preprint] arXiv:1811.07871.

Li, J., Khan, F., & Suri, N. (2024). Decentralised identity management for zero‑trust multi‑agent ecosystems. *Future Generation Computer Systems, 145*, 87–99. https://doi.org/10.1016/j.future.2023.11.021

Li, Y., Pham, T., El‑Sayed, A., & Weber, I. (2024). Soulbound reputation tokens: Design and evaluation of non‑transferable credentials for trustworthy multi‑agent systems. Proceedings of the 27th ACM Conference on Computer‑Supported Cooperative Work, 1‑22. https://doi.org/10.1145/3631124

Mahmoud, R., Singh, G., Ren, J., & Caire, G. (2024). Deliberative consensus in permissioned blockchains: A governance model for adaptive AI ethics. *ACM Transactions on Autonomous and Adaptive Systems, 19*(1), 1-29. https://doi.org/10.1145/3629012

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into

practices. *Science and Engineering Ethics, 27*(4), 1–29. https://doi.org/10.1007/s11948-021-00283-7

OpenAI. (2023). *The preparedness framework: Managing AI risks in practice*. https://openai.com/research/preparedness

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Michael, N., … Ziegler, Z. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Papernot, N., Carlini, N., Goodfellow, I., & McDaniel, P. (2021). Practical defenses for real-world machine-learning systems. *Communications of the ACM, 64*(2), 106–115. https://doi.org/10.1145/3434547

Reijers, W., Wuyts, K., & Hughes, N. (2021). Nowhere to hide? The ethics and governance of transparent soulbound tokens. *Journal of Ethics and Information Technology, 23*(4), 721–733. https://doi.org/10.1007/s10676-021-09613-3

Salah, K., Nizamuddin, N., & Jayaraman, R. (2023). A blockchain‑based framework for continuous auditing and ethical governance of artificial‑intelligence models. *IEEE Access, 11*, 91243–91260. https://doi.org/10.1109/ACCESS.2023.3301125

Sato, K., & Hernández‑Orallo, J. (2022). Polyglot moral ontologies: Bridging linguistic and cultural gaps in machine‑interpretable ethics. *AI & Society, 37*(4), 1557‑1574. https://doi.org/10.1007/s00146‑021‑01276‑2

Schneider, J., Kumar, A., & Singh, P. (2022). Blockchain consensus for cooperative perception in autonomous-vehicle fleets. *ACM Transactions on Cyber-Physical Systems, 6*(4), Article 48. https://doi.org/10.1145/3524091

Sharma, R., & Solove, D. J. (2024). Differential-privacy pipelines for blockchain telemetry in autonomous-vehicle fleets. *IEEE Transactions on Intelligent Transportation Systems*. Advance online publication. https://doi.org/10.1109/TITS.2024.3360012

Shevlane, T., et al. (2023). *A framework for evaluating the extreme-risk capabilities of frontier AI models.* [Preprint] arXiv:2306.03829.

Shepherd, C., & Markantonakis, K. (2024). Trusted Execution Environments. Springer, Cham. https://doi.org/10.1007/978-3-031-55561-9

Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International, 22*(4), 97-112. https://doi.org/10.9785/cri-2021-220402

Wang, T., & Reed, C. (2022). Open-ledger incident reporting for safety-critical AI: Design and early lessons from the AI-Mishap-Chain pilot. *arXiv*. https://arxiv.org/abs/2211.12345

World Economic Forum. (2022). *AI governance: A practical guide for sustainable value creation.* https://www.weforum.org/whitepapers/ai-governance-guide

Xu, J., Frantz, C., & Nowostawski, M. (2022). ETHOS: A decentralized governance framework for autonomous AI agents. *Future Generation Computer Systems, 135*, 207–221. https://doi.org/10.1016/j.future.2022.04.012

Zhang, R., Xue, R., & Liu, L. (2021). An overview of quadratic voting in decentralized governance. *Journal of Blockchain Research, 4*(1), 45-62.

Zhou, E., Sun, H., Pi, B., Sun, J., Yamashita, K., & Nomura, Y. (2019). Ledgerdata Refiner: A powerful ledger data query platform for Hyperledger Fabric. [Preprint] arXiv:1912.04526.