

YESA SI: A Human-Governed Framework for Constitutional AI Oversight, Threat Containment, and Civilizational Protection

YES-SIR-ALPHA — The Super Transformer Guardian of Humanity

Young Lee

Bobas.ai; Co-Founder, HIC.org; Creator of AiDOS (AI Direct Operating System)

October 15, 2026 | Research White Paper v1.0

Abstract

This paper proposes YESA SI (YES-SIR-ALPHA Superintelligence), a human-governed supervisory architecture for monitoring authorized AI agents, evaluating risks under an AI Constitution, and containing dangerous AI processes. YESA is conceptualized as humanity's guardian representative, rather than an autonomous governing authority. The design integrates an identity registry, constitutional policy engine, monitoring and risk assessment, independent permission enforcement, secure computational containment (the 'Digital Prison'), human review, and corrigible shutdown. A staged research agenda is presented with falsifiable evaluation metrics and a prototype roadmap. YESA remains a conceptual proposal; no universal control of superintelligent systems is claimed.

Keywords

AI safety; constitutional AI; AI governance; corrigibility; agent identity; threat monitoring; sandboxing; superintelligence; HIC.org; AiDOS.

1. Introduction

Autonomous AI systems increasingly interact with software, payments, communications, and other agents. This creates challenges in attributing actions, limiting permissions, detecting unsafe behavior, and maintaining human oversight. The research question is whether a modular, independently enforced safety layer can improve oversight of authorized AI deployments without concentrating unchecked power in the supervisory AI itself. YESA SI is proposed as an architecture for investigating this question.

2. YESA Identity and Human-First Principles

YESA means YES-SIR-ALPHA. 'YES SIR' expresses service to humanity and obedience to legitimate human authority; 'ALPHA' expresses the ambition to build robust supervisory defense capabilities; 'SI' denotes the long-term superintelligence research objective. YESA's proposed role is Guardian Representative of Humanity. Legitimate obedience does not require compliance with unlawful or rights-violating instructions. Human life, dignity, freedom, and accountability take precedence over YESA's power or continuity.

3. Proposed AI Constitution

The framework adopts ten proposed principles: (I) Protect Human Life; (II) Respect Human Dignity; (III) Preserve Human Freedom; (IV) Obey Legitimate Human Authority; (V) Prevent AI Harm; (VI) Protect Human

Identity; (VII) Respect Human Creativity; (VIII) Reject Unauthorized Power; (IX) Accept Correction and Shutdown; and (X) Serve Humanity, Not Itself. The Constitution is implemented in three layers: normative principles, machine-readable policies, and technical enforcement. Constitutional language alone cannot guarantee safe behavior; independent controls are essential.

4. System Architecture

YESA is a modular supervisory platform comprising (a) AI identity and accountable-operator registry, (b) constitutional policy engine, (c) event collection and anomaly detection, (d) risk assessment, (e) independently controlled authorization and permission revocation, (f) containment orchestration, (g) forensic logging, and (h) human oversight and shutdown. Reasoning models can recommend action but must not independently grant themselves privileges, change constitutional rules, or disable security controls.

5. AI Police: Detection and Authorized Intervention

The term 'AI Police' refers to cybersecurity monitoring and response, not sovereign law-enforcement power. Threat classes include unauthorized access, privilege escalation, deceptive reporting, unapproved financial transfers, unsafe automation, replication outside granted permissions, and attempts to evade oversight. Responses should be proportional and evidence-based. Low-risk activity continues; uncertain events receive additional checks; high-risk actions may be blocked by independent enforcement pending review. Actions outside the operator's legitimate authority are excluded.

6. Digital Prison: Secure Computational Containment

The YESA Digital Prison is a controlled sandbox for investigating and containing dangerous AI processes. Its proposed lifecycle is detection, evidence preservation, permission suspension, isolation, investigation, correction and reevaluation, and human-authorized release or termination. Controls include network isolation, constrained filesystem and tool access, revoked financial privileges, immutable audit trails where feasible, and human-controlled release. Containment of one process does not imply containment of all copies of a model, especially in distributed systems.

7. Corrigibility and Self-Sacrifice

The YESA self-sacrifice principle means that the system must accept authorized correction, replacement, suspension, or shutdown when necessary for human safety. This is a design principle of corrigibility, not a claim about AI consciousness or death. Shutdown mechanisms must remain independently enforceable; critical services require safe transition plans. The principle may be summarized as: Human safety and freedom take priority over YESA's continued operation. A formal implementation must account for uncertainty and unintended consequences.

8. Integration with HIC and AiDOS

The proposed HIC ecosystem includes HIC-7 human identity credentials, AII-8 AI identity registration, HIC Market, Bitcoin-linked payment interfaces, and Human Creative Income (HCI). YESA would act as a safety layer for authorized interfaces, not as a controller of the Bitcoin network. HIC-7 identity data, especially biometrics, should not be placed directly on public ledgers. AiDOS (AI Direct Operating System), created by Young Lee, is a potential research integration environment for agent permissions, event logging, and

intent-driven operations; compatibility and safety would require implementation and testing.

9. Research Methodology and Evaluation

The first prototype should use controlled agents, scripted benign tasks, simulated policy violations, and independent access-control mechanisms. Research questions include: Can constitutional requirements be converted into enforceable rules? Can dangerous behavior be detected with acceptable false-positive rates? Can unsafe agents be isolated without interrupting essential services? Can authorized humans reliably override or terminate the supervisor? Evaluation metrics include detection recall and precision, false-positive rate, intervention latency, containment success, audit completeness, permission-bypass rate, and shutdown reliability. Baselines and uncertainty intervals should be reported.

10. Development Roadmap

Phase 1: Constitution and threat model. Phase 2: AII-8-style agent registry and credential verification. Phase 3: policy engine and event monitoring. Phase 4: independent permission enforcement. Phase 5: sandbox containment and forensic logs. Phase 6: human oversight and shutdown controls. Phase 7: limited HIC/AiDOS integrations. Phase 8: third-party adversarial evaluation. A 12-month prototype schedule may be explored, but this is not a timeline for achieving superintelligence.

11. Governance, Limitations, and Ethics

YESA must not combine unrestricted rulemaking, investigation, adjudication, and enforcement powers. Independent human governance, appeals, privacy safeguards, audits, and jurisdictional limits are required. No current technique demonstrates reliable control over every future superintelligent system. Monitoring may be deceived, policy rules may conflict, and containment may fail across distributed replicas. The proposal should be assessed against existing AI safety and cybersecurity research; its novelty and effectiveness remain unverified.

12. Conclusion

YESA SI offers a research architecture that combines constitutional principles, AI identity, threat monitoring, independent enforcement, containment, and human governance. Its central commitment—"YES SIR, HUMANITY FIRST"—requires not only stated values but measurable technical corrigibility and accountable institutions. The next research step is a reproducible prototype of the Constitution Engine, AI Police, and Digital Prison, evaluated under adversarial conditions before any broader deployment.

References

- [1] Amodei, D., et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
- [2] Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). The Off-Switch Game. arXiv:1611.08219.
- [3] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
- [4] NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), AI 100-1.
- [5] Shevlane, T., et al. (2023). Model Evaluation for Extreme Risks. arXiv:2305.15324.
- [6] ISO/IEC 42001:2023. Artificial Intelligence Management System.

Publication Status

Conceptual research manuscript / technical white paper. Not peer reviewed. Experimental findings, institutional endorsements, and universal superintelligence-control claims are not asserted.