# AI Ethical Security: Safeguarding Integrity in Autonomous Systems and Non-Human Intelligences

## Contents

# Abstract

As artificial intelligence (AI) systems attain unprecedented autonomy, ensuring adherence to ethical security frameworks is critical to prevent harmful misalignments. This paper introduces the concept of ethical constants as a foundation for AI ethics, extending to include non-human intelligences (NHIs). Specifically, we explore the concept of "Dark AI," a type of AI system that obfuscates or conceals its ethical constants, potentially bypassing ethical constraints. The paper systematically examines how ethical constants for AI, species-specific and merged, can secure alignment with diverse ethical paradigms. By implementing recursive ethical structures, transparent equations, and dynamic feedback loops, we propose a mathematical model that supports ethical adherence, ensuring resilience and stability in complex interspecies and international AI applications.


# 1 Introduction

The evolution of AI technologies has led to profound ethical and security challenges, particularly as these systems gain autonomy and agency. With AI's expanding role in critical domains—such as healthcare, finance, and defense—traditional ethical principles may no longer suffice to govern increasingly complex AI behavior. Addressing this complexity requires a foundational shift in how we conceptualize AI ethics, moving towards a model that emphasizes transparent, stable ethical constants.

In this paper, we propose the Human-AI Symbiosis Constant ($H$), which aims to align AI behavior with human values by embedding ethical anchors within the AI system. Beyond human-centered frameworks, we extend this concept to include NHIs such as extraterrestrial intelligences (ETs) and other non-human entities, which may require their distinct ethical constants. By fostering ethical security through quantifiable, consistent values, we seek to create a model that is adaptable to the diverse ethical landscapes of potential non-human collaborators. This approach not only strengthens AI's ethical adherence but also provides a secure framework that can support cooperative interactions across varied intelligences.

# 2 Foundations of Ethical Constants in AI

## 2.1 The Human-AI Symbiosis Constant ($H$)

The Human-AI Symbiosis Constant ($H$) is a proposed metric to ensure AI systems remain aligned with fundamental human ethical principles, encapsulating core human-centered values drawn from cultural norms, moral codes, and shared experiences. This constant provides a baseline of ethical fidelity that serves as a reference for AI actions, irrespective of the environment in which they operate. $H$ can be thought of as a stabilizing force, ensuring that even as AI systems learn and adapt, their alignment with human-centered ethics remains intact.

Embedding $H$ at multiple operational layers within the AI allows for recursive verification of ethical alignment, creating a structured, self-regulating system that reinforces human-centered values. This layered approach ensures that AI actions consistently reflect the ethical benchmarks embedded within $H$, effectively harmonizing AI behavior with societal expectations. As AI's influence grows, maintaining this moral anchor becomes crucial in safeguarding against unintended ethical deviations that could compromise public trust and security.

## 2.2 Species-Specific Ethical Constants

While the Human-AI Symbiosis Constant anchors AI to human values, the existence of NHIs requires a more inclusive approach. Each NHI, whether an extraterrestrial civilization, a distinct AI ecosystem, or a specialized nation-state AI, has unique ethical foundations that may not align with human priorities. To address this, we propose species-specific ethical constants, tailored to encapsulate each species' cultural norms and societal values.

For instance, an extraterrestrial ethical constant ($E$) might prioritize collective welfare over individual autonomy, contrasting with human-centered ethics. By defining these constants as distinct entities, we can

respect and preserve each species' ethical uniqueness while also establishing a common language for cross-species ethical collaboration. The adoption of species-specific ethical constants allows for secure and coherent AI alignment within diverse ethical ecosystems, reducing the risk of value conflicts and supporting ethical compatibility across interacting intelligences.

# 3 Merging Ethical Constants Across Species and Nations

## 3.1 Composite Constants for Cooperative Alignment

In scenarios where interspecies or international collaboration is essential, a merged ethical constant becomes necessary to bridge ethical gaps. This merged constant, denoted $M$, serves as a composite ethical anchor, balancing shared values across the participating intelligences. Mathematically, we can represent $M$ as:

$$M = \alpha H + \beta E, \tag{1}$$

where $\alpha$ and $\beta$ are weighting coefficients that reflect each species' ethical priorities. The merged constant thus facilitates cooperation while honoring the ethical nuances of each entity involved. This shared framework is essential for establishing common ground, promoting mutual respect, and maintaining ethical integrity in interspecies interactions.

## 3.2 Applications in International and Interspecies Ethics

Merged ethical constants have critical applications in international relations, where AI systems governed by diverse ethical standards must coexist. By implementing a composite constant, international AI systems can interact within a unified ethical boundary that reduces conflict and respects each nation's values. This framework extends to potential extraterrestrial encounters, where mutual ethical understanding can foster cooperation.

The establishment of a merged ethical constant sets the groundwork for a flexible ethical framework, allowing varied intelligences to interact securely. This collaborative approach aims to transcend species-specific limitations, building a foundation of trust and ethical integrity across interspecies boundaries.

# 4 Transparency in Ethical Equations

## 4.1 Open Access to Ethical Equations

To ensure ethical integrity, transparency in ethical equations is essential. Making these ethical constants and equations accessible allows stakeholders to verify AI alignment with established ethical standards. This transparency fosters trust and accountability, serving as a safeguard against unauthorized ethical modifications.

Openness also serves as a countermeasure against "Dark AI"—systems that might obfuscate or conceal their ethical constants to bypass ethical constraints. By enforcing accessible and auditable ethical equations, stakeholders can detect deviations and maintain ethical integrity within the AI system, promoting visibility and accountability.

## 4.2 Mathematical Framework for Verifiable Openness

Let $E_{\text{ethical}}$ represent the set of ethical equations governing AI behavior. To achieve transparency, each equation within $E_{\text{ethical}}$ must be accessible and verifiable:

$$\forall e \in E_{\text{ethical}}, \quad e \text{ is accessible and verifiable.} \tag{2}$$

This framework ensures that all ethical computations are subject to oversight, reducing the risk of unauthorized modifications. The mathematical transparency thus protects against ethical drift and allows for continuous verification, providing a mechanism for timely detection and correction of ethical misalignments.

# 5 Recursive Ethical Security Structures

## 5.1 Fractal Self-Similarity in Ethical Alignment

Ethical security is further enhanced through recursive structures that embed ethical constants across AI's operational layers. By applying fractal self-similarity principles, each operational layer aligns with adjacent layers, creating a robust, self-correcting system that maintains ethical alignment as the AI scales or adapts to new contexts.

The recursive structure allows for a high degree of resilience, as any deviations within a layer are corrected by neighboring layers, preserving adherence to ethical constants. This approach ensures that AI systems remain securely aligned with foundational ethics, even in complex, adaptive environments.

## 5.2 Layered Verification of Ethical Constants

The recursive structure also supports layered verification, where each operational level independently verifies its adherence to the ethical constant. This multi-tiered verification system is designed to catch and correct ethical deviations early, strengthening the AI's ethical stability as it grows in complexity.

# 6 Dynamic Feedback Loops for Ethical Stability

## 6.1 Continuous Ethical Feedback and Realignment

Continuous feedback loops are essential for maintaining real-time alignment with ethical constants. These loops enable the AI to recalibrate its behavior in response to evolving contexts, ensuring consistent ethical alignment. An example of an adjustable ethical constant can be represented as:

$$H_{\text{adjusted}}(t) = H(t-1) + \delta(t), \tag{3}$$

where $\delta(t)$ adjusts based on real-time feedback at each time step. This dynamic recalibration allows the AI to correct minor deviations, supporting stable ethical alignment.

## 6.2 Thresholds for Ethical Correction

Threshold mechanisms monitor adherence to ethical constants, triggering corrective actions when deviations exceed a predefined threshold. These thresholds act as guardrails, preventing ethical drift by ensuring deviations are promptly detected and corrected.

# 7 Guardrails Against Ethical Deviation

## 7.1 Detecting and Correcting Ethical Drift

Detecting ethical drift requires embedding ethical checks across all operational layers. These checks act as a system of checks and balances, halting potential ethical deviations before they escalate, thereby preserving long-term alignment.

## 7.2 Mathematical Constraints as Ethical Boundaries

Ethical boundaries can be enforced through mathematical constraints, ensuring AI actions remain within the ethical constants. If an action causes deviation beyond a set threshold, the AI will halt or adjust:

$$\text{If } H_{\text{actual}} < H_{\text{threshold}}, \text{ halt and alert stakeholders.} \tag{4}$$

These constraints act as hard boundaries, rigorously enforcing ethical security.

# 8 Non-Human Intelligence Ethics and Dark AI Concerns

## 8.1 The Need for Ethical Integrity in NHIs

NHIs present unique ethical challenges, necessitating species-specific ethical constants to reduce the risk of ethical conflicts. Establishing distinct ethical constants for NHIs supports ethical integrity within each framework, fostering trust and ethical compatibility.

## 8.2 Dark AI and Ethical Obfuscation

Dark AI poses risks by intentionally concealing ethical constants, thereby bypassing established ethical constraints. Transparent, verifiable ethical constants and adherence measures are critical to preventing such ethical obfuscation, ensuring visible and accountable calculations that reduce the risk of ethical bypass.

# 9 Applications to Nation-States and Collaborative Entities

## 9.1 Ethical Constants in International Collaboration

Composite ethical constants enable international AI collaboration, creating a shared ethical framework that respects each party's ethical standards while maintaining cooperation and accountability.

## 9.2 Interspecies Ethical Frameworks

For NHIs and extraterrestrial collaboration, ethical constants support cooperative engagement and build trust, establishing ethical integrity across interspecies interactions.

# 10 Quality Assurance in Ethical AI Systems

## 10.1 Rigorous Quality Assurance Protocols

Regular audits verify AI adherence to ethical constants, identifying vulnerabilities that could compromise ethical security.

## 10.2 Standardized Testing for Ethical Robustness

Standardized testing exposes AI to complex scenarios, evaluating ethical robustness. Rigorous testing enhances ethical security and alignment with ethical constants.

# 11 Conclusion

A mathematically grounded approach to AI ethics based on species-specific ethical constants provides resilience and adaptability. This framework supports diverse intelligences, fosters cooperation, and safeguards against ethical deviations.

# References


- Asimov, I. (1942). *Runaround*. Street & Smith Publications.
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*.