EthAIca. 2025; 4:434
doi: 10.56294/ai2025434

ORIGINAL

A dignitarian approach to AI ethics: grounding normative principles in human value

Un enfoque dignitario a la ética de la IA: fundamentando principios normativos en el valor humano

Mohammed Zeinu Hassen[1]

[1]Addis Ababa Science and Technology University, Department of Social Sciences. Addis Ababa, Ethiopia.

Cite as: Zeinu Hassen M. A dignitarian approach to AI ethics: grounding normative principles in human value. EthAIca. 2025; 4:434. https://doi.org/10.56294/ai2025434

Submitted: 25-06-2025 **Revised:** 20-08-2025 **Accepted:** 01-11-2025 **Published:** 02-11-2025

Editor: PhD. Rubén González Vallejo

Corresponding author: Mohammed Zeinu Hassen

ABSTRACT

Introduction: the proliferation of guidelines for artificial intelligence ethics presented a field without a firm philosophical foundation. Current documents offered a disparate collection of principles, often lacking a unified justification for their normative force. This paper confronted that deficiency by proposing a novel dignitarian framework. The objective of this research was to establish a stable and rationally defensible basis for the design, deployment, and governance of AI systems.
Method: this study employed a conceptual analysis of the Kantian philosophical tradition to define human dignity as an absolute, intrinsic value. This core concept was then formalized into a coherent axiomatic system using elementary set theory and deontic logic. The analysis was based on a critical review of foundational texts in moral philosophy and contemporary AI ethics literature.
Results: a primary normative constraint emerged from this formalization: an AI system's action, a, was morally permissible only if it did not treat any person, p, merely as a means to an end. This was expressed logically as Permissible(a) $\rightarrow \forall p \in P, \neg$ViolatesDignity(a, p). This principle functioned as a strict deontological limit on any goal-oriented programming.
Conclusions: the proposed framework provided a stable and rationally defensible basis for the design, deployment, and governance of AI systems. It moved the conversation from a list of suggestions to a structured ethical system, contributing to the growing field of computational ethics by offering a clear, implementable, and non-negotiable constraint on AI behavior.

Keywords: AI Ethics, Computational Ethics; Human Dignity; Kantian Ethics; Normative Principles; AI Governance.

RESUMEN

Introducción: la proliferación de directrices para la ética de la inteligencia artificial presentó un campo sin una base filosófica firme. Los documentos actuales ofrecieron una colección dispar de principios, que a menudo carecían de una justificación unificada para su fuerza normativa. Este artículo enfrentó directamente esa deficiencia proponiendo un novedoso marco dignatario. El objetivo de esta investigación fue establecer una base estable y racionalmente defendible para el diseño, la implementación y la gobernanza de los sistemas de IA.
Método: este estudio empleó un análisis conceptual de la tradición filosófica kantiana para definir la dignidad humana como un valor absoluto e intrínseco. Este concepto central se formalizó luego en un sistema axiomático coherente utilizando la teoría de conjuntos elemental y la lógica deóntica. El análisis se basó en una revisión crítica de textos fundamentales de la filosofía moral y de la literatura contemporánea sobre ética de la IA.

Resultados: de esta formalización surgió una restricción normativa principal: una acción de un sistema de IA, *a*, era moralmente permisible solo si no trataba a ninguna persona, *p*, simplemente como un medio para un fin. Esto se expresó lógicamente como Permisible(a) → ∀p ∈ P, ¬ViolatesDignity(a, p). Este principio funcionó como un límite deontológico estricto en cualquier programación orientada a objetivos.
Conclusiones: el marco propuesto proporcionó una base estable y racionalmente defendible para el diseño, la implementación y la gobernanza de los sistemas de IA. Llevó la conversación de una lista de sugerencias a un sistema ético estructurado, contribuyendo al creciente campo de la ética computacional al ofrecer una restricción clara, implementable y no negociable para el comportamiento de la IA.

Palabras clave: Ética de la IA; Ética Computacional; Dignidad Humana; Ética Kantiana; Principios Normativos; Gobernanza de la IA.

INTRODUCTION

The field of artificial intelligence ethics is currently in a state of disarray, characterized by a disparate collection of guidelines and principles.[1,2] While governments, corporations, and academic bodies frequently produce these documents, they often lack a coherent philosophical justification, leaving their normative force questionable. This absence of a solid foundation creates significant challenges, making consistent application across different contexts difficult and defense against rational scrutiny harder. Conflicts often arise between stated goals, such as the tension between achieving maximum accuracy and ensuring procedural fairness.[3] Without a meta-ethical framework to resolve these disputes, practitioners are left with arbitrary choices, resulting in a collection of well-intentioned but theoretically weak suggestions. This situation is unsustainable for guiding the creation of powerful technologies, underscoring the urgent need for a stable and rationally defensible ethical system.[4]

The current landscape of AI ethics is dominated by two primary, yet flawed, approaches. The first is a form of soft-law principlism, marked by guidelines emphasizing concepts like Fairness, Accountability, and Transparency (FAT). These principles, while laudable, are often vaguely defined and can conflict. The second dominant approach is a latent utilitarianism, which seeks to maximize a certain good, such as social welfare.[6] This consequentialist calculus, however, risks justifying actions that treat individuals or minority groups as mere instruments for a greater aggregate good.

This paper forwards a specific solution to this foundational crisis: a dignitarian ethical framework.[5] This approach offers a robust, defensible grounding for normative principles governing AI, based on the absolute and intrinsic value of human persons. This value, termed dignity, is non-negotiable and serves as the ultimate constraint on any technological design or deployment. The objective is to propose and formalize such a framework based on the principle of human dignity.

METHOD

This paper employed a method of conceptual analysis and normative argumentation. The inquiry was not an empirical study and did not collect or analyze new experimental data. The approach was fundamentally philosophical, seeking to establish a rationally defensible framework for what ought to be the case in the domain of AI ethics. The mode of inquiry was one of justification, not of description or prediction.

The argument proceeded in two distinct stages. The first stage was one of conceptual analysis. This work drew from the history of philosophy to define its core terms. It specifically utilized the Kantian tradition to establish a clear definition of human dignity.[7] The second stage was one of normative argumentation. This work used the tools of formal logic and elementary set theory to translate the philosophical concepts into precise, machine-interpretable principles.

DEVELOPMENT

The dignitarian approach locates the foundation of all moral value in a single, specific concept: human dignity. Human dignity is an absolute, intrinsic, and non-negotiable value, possessed equally by all rational beings.[8] This contrasts with "price," which is a relative value that can be exchanged. The source of dignity is personhood, defined as a being with rational autonomy. This capacity for moral self-legislation makes a person an "end in themselves." This concept is the cornerstone of the second formulation of Kant's Categorical Imperative: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means."[8] The moral violation occurs when a person is treated *merely* as a means.

RESULTS

To operationalize the dignitarian framework, a universe of discourse was defined: a set of all Persons, P; a set of all AI Agents, M; a set of all possible Actions, A; and a set of all possible States of Affairs, S. The dignitarian approach introduced a strict deontological constraint expressed in the first axiom, the **Primacy of Dignity**: An action $a \in A$ is permissible only if it does not violate the dignity of any person $p \in P$.

Axiom 1 (Primacy of Dignity): Permissible(a) $\rightarrow \forall p \in P, \neg$ViolatesDignity(a,p)

The predicate *ViolatesDignity (a, p)* was formally defined as: (Uses(a, p) $\wedge$ ¬Consented(p, a)) $\rightarrow$ ViolatesDignity(a, p). An action *a* violates the dignity of a person *p* if that action uses *p* as a causal instrument without the possibility of rational consent from *p*. "Rational consent" here refers to a hypothetical standard based on deliberative and contractarian ethics.[9,10] This axiomatic structure provided a clear decision calculus for any AI agent.

DISCUSSION

The formal axiomatic system provided a powerful tool for analyzing difficult cases in AI ethics. For instance, in cases of algorithmic bias in hiring, the framework forbids using a protected attribute as a negative factor because a person cannot rationally consent to such a system. In unavoidable crash scenarios involving autonomous vehicles, the dignitarian framework forbids intentionally sacrificing one person to save a larger group, as this would treat them as a disposable object.

No ethical framework is without its challenges. A primary objection from a consequentialist perspective is whether a strict deontological constraint risks catastrophic outcomes. The dignitarian response is that building systems with the pre-programmed capacity to violate human dignity creates a far greater and more certain risk of misuse. A second challenge relates to interpretation, for instance in personalized advertising.[11] The dignitarian framework correctly frames the debate around whether the mechanism respects the user as a rational being. A final challenge is one of scope, particularly concerning the Kantian definition of a "person".[12]

CONCLUSIONS

This paper diagnosed a foundational deficit in the current state of AI ethics and proposed a formalized dignitarian approach based on established Kantian principles as a remedy. The central thesis was that this framework provides the required philosophical foundation, moving beyond a mere list of guidelines to a structured system derived from a specific theory of human value. The Primacy of Dignity axiom functions as a non-negotiable deontological constraint on the behavior of any AI agent. Implementing this framework requires a "dignity-by-design" approach, involving collaboration between philosophers, computer scientists, and policymakers to move the ethical discussion from abstract principles to concrete, implementable, and enforceable technical and legal standards.

REFERENCES

1. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-99.

2. Coeckelbergh M. AI ethics. The MIT Press; 2020.

3. Rawls J. A theory of justice. Harvard University Press; 1971.

4. Turing AM. Computing machinery and intelligence. Mind. 1950;59(236):433-60.

5. Hanna R, Kazim E. Philosophical foundations for digital ethics and AI ethics: A dignitarian approach. AI and Ethics. 2021;1(4):405-23.

6. Mill JS. Utilitarianism. Hackett Publishing; 2002. (Original work pub. 1863).

7. Kant I. Groundwork of the metaphysics of morals. Gregor MJ, editor and translator. Cambridge University Press; 2021. (Original work pub. 1785).

8. Kant I. Groundwork of the metaphysics of morals. Gregor MJ, editor and translator. Cambridge University Press; 2012. (Original work pub. 1785).

9. Habermas J. The theory of communicative action, vol 1. McCarthy T, translator. Beacon Press; 1984.

10. Gauthier D. Morals by agreement. Oxford University Press; 1986.

11. Floridi L. The ethics of information. Oxford University Press; 2018.

12. Searle JR. Minds, brains, and programs. Behavioral and Brain Sciences. 1980;3(3):417-24.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohammed Zeinu Hassen.
Formal analysis: Mohammed Zeinu Hassen.
Research: Mohammed Zeinu Hassen.
Methodology: Mohammed Zeinu Hassen.
Drafting - original draft: Mohammed Zeinu Hassen.
Writing - proofreading and editing: Mohammed Zeinu Hassen.

APPENDIX

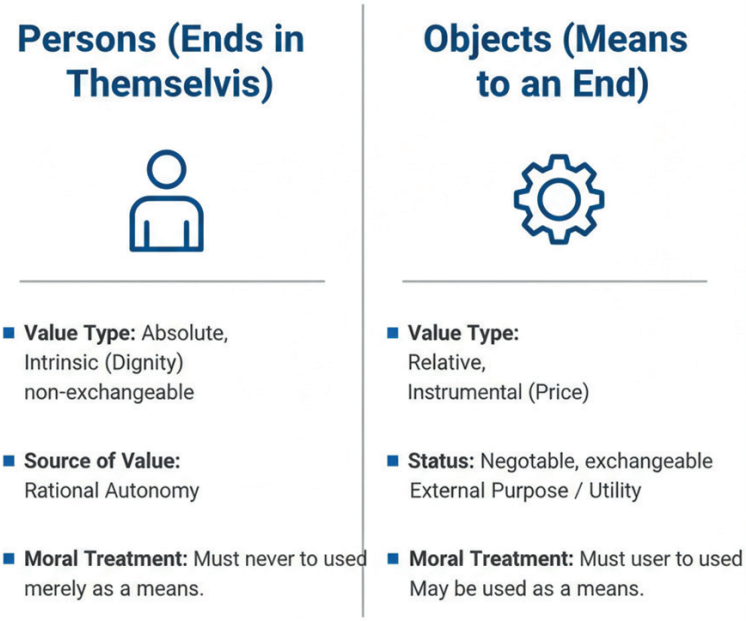

Figure 1. This chart illustrates the fundamental Kantian distinction between persons and objects. Persons possess an absolute, intrinsic value (dignity) derived from their rational autonomy, making them ends in themselves. Objects possess a relative, instrumental.

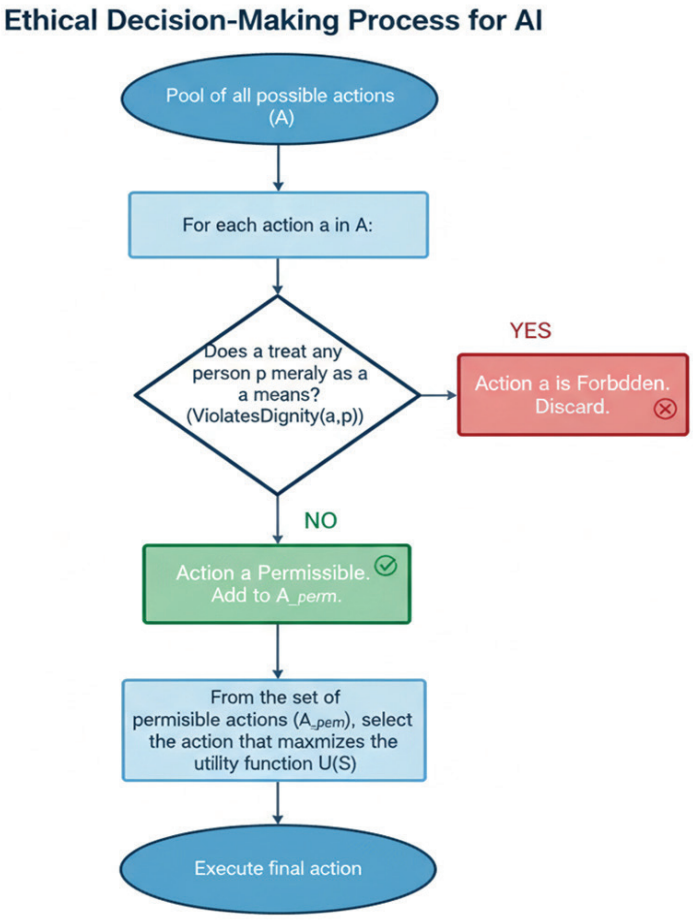

Figure 2. This flowchart visualizes the two-stage decision calculus proposed by the axiomatic system. First, all potential actions are subjected to a deontological filter. Only actions that do not violate human dignity are deemed permissible. Second, from this fi this filtered set, the AI selects the action that optimizes for its utility function.

https://doi.org/10.56294/ai2025434 ISSN: 3072-7952

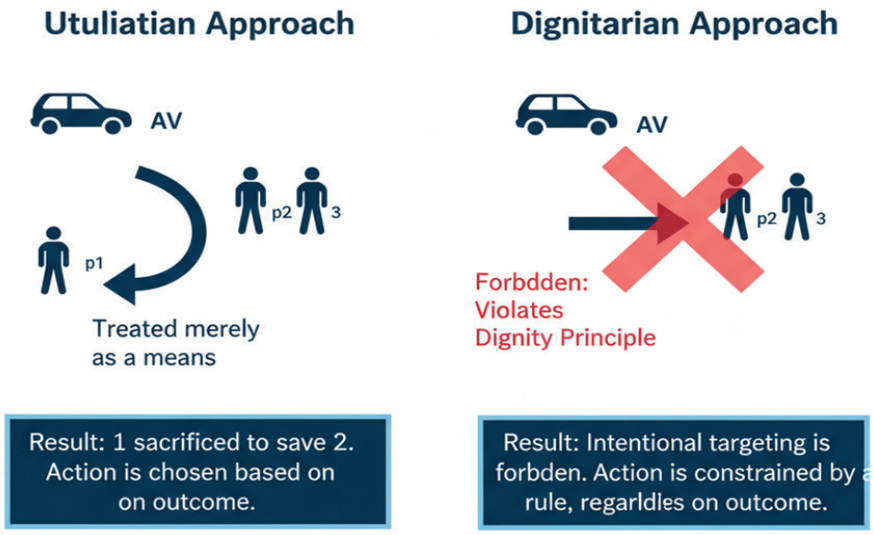

Figure 3. The utilitarian model (left) permits sacrificing one person as a means to save a greater number. The dignitarian framework (right) forbids this action a priori because it instrumentalizes a person, violating the Primacy of Dignity axiom.

https://doi.org/10.56294/ai2025434 ISSN: 3072-7952