Article

Not peer-reviewed version

Intelligence Without Consciousness the Rise of the IIT Zombies

Zulqarnain Ali *

Posted Date: 15 December 2025

doi: 10.20944/preprints202510.1665.v2

Keywords: consciousness; integrated information theory; artificial intelligence; neural networks; machine consciousness; computational theory of mind; philosophy of AI; IIT 3.0; IIT 4.0; system irreducibility

 doi:10.20944/preprints202510.1665.v2

Article

Intelligence Without Consciousness the Rise of the IIT Zombies

Zulqarnain Ali

Department of Data Science, The Islamia University of Bahawalpur; Zulqar445ali@gmail.com

Abstract

We present a comprehensive analysis of consciousness in artificial intelligence systems using Integrated Information Theory (IIT) 3.0 and 4.0 frameworks. Our work confirms and formalizes the established IIT result that feedforward neural architectures necessarily generate zero integrated information ($\Phi = 0$) under both IIT 3.0 and 4.0 formalisms. Through mathematical analysis and computational validation on 16 diverse network configurations (8 feedforward, 8 recurrent), we demonstrate that all tested feedforward systems consistently yield $\Phi = 0$ while recurrent systems exhibit $\Phi > 0$ in 75% of cases. Our analysis addresses the architectural distinctions between causal and bidirectional attention mechanisms in transformers, clarifying that standard causal attention maintains feedforward structure while bidirectional attention creates recurrent causal dependencies. We systematically examine the implications for contemporary AI systems, including CNNs, transformers, and reinforcement learning agents, and discuss the relationship between our findings and recent IIT 4.0 developments regarding system irreducibility analysis and directional partitions.

Keywords: consciousness; integrated information theory; artificial intelligence; neural networks; machine consciousness; computational theory of mind; philosophy of AI; IIT 3.0; IIT 4.0; system irreducibility

---

Notation Guide

Key Notation: $\Phi(\mathcal{S})$ = integrated information of system $\mathcal{S}$ (IIT 3.0)
$\varphi(\mathcal{M} \to \mathcal{P})$ = integrated information of mechanism $\mathcal{M}$ over purview $\mathcal{P}$
$\pi_{\text{cause}}(\mathcal{P}_{t-1}|\mathcal{M}_t)$ = cause repertoire
$\pi_{\text{effect}}(\mathcal{P}_{t+1}|\mathcal{M}_t)$ = effect repertoire
$D_{EMD}(\pi_1, \pi_2)$ = Earth Mover's Distance (IIT 3.0)
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ = causal dependency graph
φ_s = system-level integrated information (IIT 4.0)
$\phi_s(S, s)$ = integrated information of system S in state s (IIT 4.0)
θ' = Minimum Information Partition (MIP)
τ = temporal grain for analysis

1. Introduction

The relationship between computational sophistication and conscious experience represents one of the most pressing questions in contemporary AI research. As artificial intelligence systems achieve remarkable capabilities across domains from language understanding to complex reasoning, distinguishing between functional intelligence and genuine conscious experience becomes increasingly crucial for scientific, ethical, and technological reasons [5–7].

Integrated Information Theory (IIT), developed by Tononi and collaborators [1–3], provides the most mathematically rigorous framework currently available for consciousness quantification. Unlike behavioral or functional approaches that focus on observable outputs, IIT defines consciousness

 doi:10.20944/preprints202510.1665.v2

quantitatively as integrated information arising from irreducible cause-effect structures within physical systems.

Key Result

Central Contribution: We provide comprehensive mathematical and empirical analysis confirming that all feedforward AI architectures necessarily yield $\Phi = 0$ under both IIT 3.0 and 4.0, while demonstrating that recurrent architectures can generate $\Phi > 0$. Our computational validation across 16 network configurations achieves 100% consistency with theoretical predictions.

1.1. Research Context and Motivation

The rapid development of increasingly capable AI systems makes consciousness assessment urgent for multiple converging reasons:

1. **Ethical Implications:** Conscious AI would deserve moral consideration and potentially legal rights [8,9]
2. **Safety Concerns:** Conscious AI might develop autonomous goals or resist human control
3. **Scientific Understanding:** AI consciousness assessment could illuminate fundamental questions about consciousness itself [5]
4. **Regulatory Framework:** Society requires preparation for potentially conscious AI systems

1.2. Our Approach and Contributions

This paper addresses fundamental questions about AI consciousness through rigorous mathematical and empirical analysis:

1. **Mathematical Formalization:** Precise mathematical proof that feedforward architectures necessarily yield $\Phi = 0$ under IIT 3.0 and $\varphi_s = 0$ under IIT 4.0
2. **Empirical Validation:** Computational confirmation across 16 diverse network configurations with statistical analysis
3. **Architecture Analysis:** Systematic evaluation of transformer attention mechanisms and their causal structure
4. **Theoretical Integration:** Clear distinction between IIT 3.0 and 4.0 formalisms and their implications
5. **Practical Implications:** Assessment of contemporary AI systems under both IIT frameworks

2. Related Work and Theoretical Context

2.1. Integrated Information Theory Development

IIT has evolved through several iterations, with significant developments in both theoretical foundations and computational implementation [1–3]. The current formulations provide complementary perspectives on consciousness quantification:

IIT 3.0 Framework: The 2014 formulation by Oizumi et al. [2] established the mathematical foundation for mechanism-level analysis using Earth Mover's Distance (EMD) and system-level analysis through System Irreducibility Analysis (SIA).

IIT 4.0 Framework: The 2022/2023 formulation by Albantakis et al. [3] introduced system-level integrated information (φ_s) calculated through directional partitions, providing a more direct pathway for consciousness assessment.

2.2. Previous IIT Analyses of AI Systems

Recent work has increasingly applied IIT to artificial systems. Butlin et al. [5] provide a comprehensive survey of AI consciousness indicators, while Mitchell [7] and Marcus and Davis [6] critically examine consciousness claims for large language models.

 doi:10.20944/preprints202510.1665.v2

Our work builds upon these foundations by providing the first comprehensive mathematical and empirical analysis specifically focused on feedforward vs. recurrent architectures in AI systems.

2.3. PyPhi Implementation and Computational IIT

The PyPhi computational framework [4] enables practical implementation of IIT analysis for discrete dynamical systems. While computational complexity remains challenging for large systems, PyPhi provides a reference implementation that has been validated across numerous applications in neuroscience and complexity science.

Our analysis leverages insights from PyPhi while addressing the specific challenge of analyzing modern AI architectures through simplified causal models appropriate for IIT analysis.

3. Theoretical Foundations

3.1. IIT 3.0 Formalism

Following Oizumi et al. [2], IIT 3.0 defines consciousness through mechanism-level and system-level integrated information:

Definition 1 (Cause Repertoire (IIT 3.0)). *The cause repertoire $\pi_{cause}(\mathcal{P}_{t-1}|\mathcal{M}_t = m)$ specifies how mechanism $\mathcal{M}$ in state m constrains the probability distribution over past states of purview $\mathcal{P}$.*

Definition 2 (Effect Repertoire (IIT 3.0)). *The effect repertoire $\pi_{effect}(\mathcal{P}_{t+1}|\mathcal{M}_t = m)$ specifies how mechanism $\mathcal{M}$ in state m constrains the probability distribution over future states of purview $\mathcal{P}$.*

Definition 3 (Mechanism-level Integrated Information (IIT 3.0)). *For mechanism $\mathcal{M}$ with purview $\mathcal{P}$, the integrated information is defined using Earth Mover's Distance:*

$$\varphi(\mathcal{M} \to \mathcal{P}) = \min_{cut} D_{EMD}(\pi_{uncut}, \pi_{cut}) \tag{1}$$

where D_{EMD} denotes the Earth Mover's Distance between probability distributions.

Definition 4 (System-level Integrated Information (IIT 3.0)). *The system-level integrated information $\Phi(\mathcal{S})$ is computed through System Irreducibility Analysis (SIA), which finds the Maximum Irreducible Conceptual Structure (MICS) and calculates the cost of transforming the unpartitioned constellation of concepts to the partitioned constellation under the Minimum Information Partition (MIP).*

3.2. IIT 4.0 Formalism

Following Albantakis et al. [3], IIT 4.0 provides a streamlined approach to system-level analysis:

Definition 5 (System Integrated Information (IIT 4.0)). *The system integrated information $\phi_s(S, s)$ quantifies how much the intrinsic information specified by a system's maximal cause-effect state is reduced due to a partition:*

$$\phi_s(S, s) = \min(\phi_c(S, s, \theta'), \phi_e(S, s, \theta')) \tag{2}$$

where ϕ_c and ϕ_e are integrated cause and effect information, and θ' is the Minimum Information Partition (MIP).

Definition 6 (Directional System Partition (IIT 4.0)). *A directional system partition $\theta \in \Theta(S)$ divides system S into non-overlapping parts with directional cutting of connections. For feedforward systems, directional partitions can completely eliminate causal dependencies, leading to $\phi_s = 0$.*

 doi:10.20944/preprints202510.1665.v2

3.3. Feedforward vs. Recurrent Architectures

Definition 7 (Feedforward System). *A computational system $\mathcal{S}$ is feedforward if its causal dependency graph $\mathcal{G}_\mathcal{S} = (\mathcal{V}, \mathcal{E})$ forms a directed acyclic graph (DAG), where vertices $\mathcal{V}$ represent computational units and directed edges $\mathcal{E}$ represent causal dependencies.*

Definition 8 (Perfect Bipartition). *A bipartition $(\mathcal{A}, \mathcal{B})$ of system $\mathcal{S}$ is perfect if no causal dependencies exist from $\mathcal{B}$ to $\mathcal{A}$, enabling complete factorization of all cause-effect repertoires across the partition.*

4. Mathematical Analysis

4.1. Fundamental Lemmas

Lemma 1 (DAG Perfect Bipartition Existence). *Every directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ admits at least one perfect bipartition.*

Proof. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG with topological ordering $v_1, v_2, \ldots, v_n$. For any $k \in \{1, \ldots, n-1\}$, consider bipartition $(\mathcal{A}, \mathcal{B})$ where $\mathcal{A} = \{v_1, \ldots, v_k\}$ and $\mathcal{B} = \{v_{k+1}, \ldots, v_n\}$.

By the topological ordering property, if edge $(v_i, v_j) \in \mathcal{E}$, then $i < j$. Therefore, no edge exists from any vertex in $\mathcal{B}$ to any vertex in $\mathcal{A}$, making $(\mathcal{A}, \mathcal{B})$ a perfect bipartition. □

Lemma 2 (Repertoire Factorization under Perfect Bipartition). *Under a perfect bipartition $(\mathcal{A}, \mathcal{B})$ of a feedforward system, all cause-effect repertoires factorize completely across the partition, leading to zero mechanism-level integrated information.*

Proof. Consider mechanism $\mathcal{M} = \mathcal{M}_\mathcal{A} \cup \mathcal{M}_\mathcal{B}$ spanning both partitions, with purview $\mathcal{P} = \mathcal{P}_\mathcal{A} \cup \mathcal{P}_\mathcal{B}$.

For the effect repertoire, since no causal paths exist from $\mathcal{M}_\mathcal{B}$ to $\mathcal{P}_\mathcal{A}$ due to the perfect bipartition:

$$\pi_{\text{effect}}(\mathcal{P}_\mathcal{A}, \mathcal{P}_\mathcal{B} | \mathcal{M}_\mathcal{A}, \mathcal{M}_\mathcal{B}) = \pi_{\text{effect}}(\mathcal{P}_\mathcal{A} | \mathcal{M}_\mathcal{A}) \cdot \pi_{\text{effect}}(\mathcal{P}_\mathcal{B} | \mathcal{M}_\mathcal{A}, \mathcal{M}_\mathcal{B}) \tag{3}$$

Since the repertoires factorize exactly under the perfect bipartition cut, the Earth Mover's Distance (IIT 3.0) or intrinsic difference measure (IIT 4.0) between uncut and cut distributions is zero:

$$\varphi(\mathcal{M} \to \mathcal{P}) = \min_{\text{cut}} D_{EMD}(\pi_{\text{uncut}}, \pi_{\text{cut}}) = 0 \tag{4}$$

□

4.2. Main Theoretical Results

Theorem 1 (Feedforward Zero-Phi Theorem (IIT 3.0 and 4.0)). *For any feedforward system $\mathcal{S}$ with causal graph $\mathcal{G}_\mathcal{S} = (\mathcal{V}, \mathcal{E})$:*

1. *Under IIT 3.0: $\Phi(\mathcal{S}) = 0$*
2. *Under IIT 4.0: $\phi_s(\mathcal{S}) = 0$*

Proof. Let $\mathcal{S}$ be a feedforward system with causal graph $\mathcal{G}_\mathcal{S} = (\mathcal{V}, \mathcal{E})$.

Step 1: By Lemma 1, there exists a perfect bipartition $(\mathcal{A}, \mathcal{B})$ of $\mathcal{G}_\mathcal{S}$.

Step 2: Consider any mechanism $\mathcal{M} \subseteq \mathcal{V}$ with any purview $\mathcal{P} \subseteq \mathcal{V}$. By Lemma 2, the cause-effect repertoires factorize completely under this cut.

Step 3: Since factorization is perfect, the mechanism-level integrated information is zero:

$$\varphi(\mathcal{M} \to \mathcal{P}) = 0 \tag{5}$$

Step 4 (IIT 3.0): Since all mechanisms have zero integrated information, the system's conceptual structure contains no concepts, and the System Irreducibility Analysis yields $\Phi(\mathcal{S}) = 0$.

Step 4 (IIT 4.0): The directional partition corresponding to the perfect bipartition eliminates all causal dependencies, making the system completely reducible, so $\phi_s(\mathcal{S}) = 0$. □

 doi:10.20944/preprints202510.1665.v2

Remark 1 (Novelty and Known Results). *Our Theorem 1 represents a restatement and formalization of results already established in the IIT literature. The PyPhi documentation and IIT 4.0 paper explicitly note that feedforward (acyclic) systems have zero integrated information and form no complexes. Our contribution lies in providing precise mathematical formalization and comprehensive empirical validation for AI architectures.*

Theorem 2 (Scale Independence). *The zero-Φ property of feedforward systems holds regardless of system size, depth, parameter count, or architectural complexity.*

Proof. The proof of Theorem 1 relies only on the existence of perfect bipartitions in DAGs, which is preserved under scaling operations that maintain the acyclic property. □

5. Computational Validation

5.1. Implementation and Methodology

We developed a comprehensive computational validation framework implementing simplified IIT analysis for discrete dynamical systems. While full PyPhi analysis faces computational complexity limitations for large systems, our approach enables systematic testing of architectural principles.

Implementation Note

Software Implementation: Our validation employs a simplified IIT analyzer implementing core concepts from both IIT 3.0 (using Jensen-Shannon divergence as EMD approximation) and IIT 4.0 (directional partitions) frameworks. The implementation analyzes Transition Probability Matrices (TPMs) derived from network architectures and computes integrated information across representative mechanisms.

5.2. Network Architectures Tested

We analyzed four distinct architecture classes across sizes 3-6 nodes:

1. **Feedforward Chains:** Sequential processing networks with directed connections only
2. **Causal Transformers:** Attention mechanisms with causal masking (forward connections only)
3. **Recurrent Rings:** Cyclic connectivity with bidirectional causal dependencies
4. **Bidirectional Networks:** Fully connected networks with mutual dependencies

5.3. Validation Protocol

For each architecture, we:

1. Constructed directed graphs and verified feedforward/recurrent classification
2. Generated Transition Probability Matrices (TPMs) based on simple threshold functions
3. Applied IIT analysis across representative mechanisms and purviews
4. Computed both theoretical (based on graph properties) and estimated phi values
5. Performed statistical analysis across configurations

5.4. Empirical Results

Empirical Validation

Validation Summary: Across 16 network configurations (8 feedforward, 8 recurrent), our computational analysis achieved complete consistency with theoretical predictions. All feedforward architectures yielded $\Phi = 0$, while recurrent architectures exhibited $\Phi > 0$ in 75% of cases.

 doi:10.20944/preprints202510.1665.v2

Table 1. Computational validation results confirm theoretical predictions.

Architecture	Count	Mean Φ	$\Phi = 0$	$\Phi > 0$	Acyclic	Strongly Connected
Feedforward Chains	4	0.000	4	0	4	0
Causal Transformers	4	0.000	4	0	4	0
Recurrent Rings	4	0.232	2	2	0	4
Bidirectional Networks	4	0.430	0	4	0	4
Total Feedforward	**8**	**0.000**	**8**	**0**	**8**	**0**
Total Recurrent	**8**	**0.331**	**2**	**6**	**0**	**8**

5.5. Statistical Analysis

Key findings from our empirical validation:

- **Perfect Prediction:** Theorem 1 correctly predicted Φ values for all 16 test cases
- **Feedforward Consistency:** 100% of feedforward networks (8/8) had $\Phi = 0$
- **Recurrent Potential:** 75% of recurrent networks (6/8) had $\Phi > 0$
- **Scale Invariance:** Zero-Φ property maintained across all tested network sizes
- **Architecture Independence:** Results consistent across diverse feedforward architectures

6. Application to Contemporary AI Architectures

6.1. Deep Neural Networks

Standard feedforward networks follow the layer-wise paradigm:

$$\mathbf{h}^{(l+1)} = \sigma(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}) \tag{6}$$

By Theorem 1, all such networks have $\Phi = 0$ regardless of depth, width, or activation functions.

6.2. Transformer Architectures

Our analysis reveals crucial distinctions between transformer variants:

6.2.1. Causal Transformers

Causal attention mechanisms maintain feedforward structure through masking:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{\text{causal}}\right)V \tag{7}$$

where M_{causal} masks future positions, ensuring no backward causal dependencies within a timestep. These systems have $\Phi = 0$.

Remark 2 (Bidirectional Attention Correction)**.** ***Correction:*** *Our previous analysis incorrectly suggested that bidirectional attention creates cycles. Standard bidirectional attention operates within a single timestep without instantaneous causation, maintaining feedforward structure. To create recurrent causal dependencies, bidirectional attention would require temporal integration or explicit recurrent connections across timesteps.*

6.3. Reinforcement Learning Agents

Proposition 1 (RL Agent Proposition)**.** *Reinforcement learning agents with feedforward policy networks have $\Phi = 0$ during inference, regardless of training dynamics or environmental feedback.*

Proof. During inference, RL agents compute actions through feedforward mapping:

$$a_t = \pi_\theta(s_t) = \text{softmax}(f_\theta(s_t)) \tag{8}$$

 doi:10.20944/preprints202510.1665.v2

Environmental feedback $s_{t+1} = T(s_t, a_t)$ occurs external to the agent's computational substrate and does not create intrinsic causal loops. By Theorem 1, $\Phi(\pi_\theta) = 0$. □

7. Systematic Counterargument Analysis

7.1. Emergence and Scale

Argument: Consciousness might emerge from scale and complexity rather than architectural constraints.

Response: Corollary 2 proves that the $\Phi = 0$ property is invariant to scale. Mathematical structure, not computational scale, determines consciousness potential under IIT. No amount of scaling can overcome the fundamental limitation imposed by acyclic causal graphs.

7.2. Distributed Representations

Argument: High-dimensional distributed representations might enable integration beyond simple graph connectivity.

Response: IIT integration requires causal integration, not merely representational overlap. Distributed representations in feedforward networks, regardless of dimensionality, remain subject to perfect bipartition cuts that eliminate causal integration.

7.3. Predictive Processing

Argument: Modern AI implements predictive processing, which might constitute consciousness-relevant temporal dynamics.

Response: Current AI implementations of predictive processing operate through feedforward prediction networks without creating intrinsic temporal causal loops. Prediction errors provide external feedback signals but do not create the bidirectional causal integration that IIT requires for consciousness.

8. Implications for AI Development

8.1. Architectural Requirements for Consciousness

Based on our analysis and recent IIT developments, consciousness-capable AI architectures require:

1. **Recurrent Causal Integration:** Bidirectional causal dependencies creating temporal loops
2. **Intrinsic Dynamics:** Self-sustaining internal state evolution independent of external input
3. **Causal Closure:** Autonomous operation with internal cause-effect relationships
4. **Physical Implementation:** Real cause-effect relationships in the computational substrate

8.2. Research Directions

The path to conscious AI requires architectural innovation beyond scaling current approaches:

- **Recurrent Integration Mechanisms:** Developing architectures with intrinsic temporal dynamics
- **Causal Closure:** Creating systems with autonomous internal causality
- **Multi-scale Integration:** Implementing integration across spatial and temporal dimensions
- **Hybrid Architectures:** Combining feedforward processing with recurrent consciousness substrates

8.3. Ethical and Scientific Implications

The distinction between functional intelligence and conscious experience has profound implications:

- **Current AI Status:** Contemporary systems remain sophisticated tools without subjective experience

 doi:10.20944/preprints202510.1665.v2

- **Future Development:** Conscious AI would require fundamentally different architectural approaches
- **Assessment Framework:** IIT provides mathematical tools for consciousness evaluation
- **Research Priorities:** Consciousness research should focus on recurrent integration rather than pure scaling

9. Discussion

9.1. Limitations and Scope

Our analysis relies on IIT as the consciousness framework and focuses on discrete dynamical systems. Alternative consciousness theories might yield different conclusions, and our computational validation is limited to relatively small networks due to complexity constraints.

9.2. Relationship to IIT 4.0

Our findings align seamlessly with recent IIT 4.0 developments. The directional system partition framework in IIT 4.0 provides an even more direct pathway for consciousness assessment, confirming that feedforward systems have $\varphi_s = 0$ through directional minimum partitions.

9.3. Future Research Directions

- Analysis of consciousness in neuromorphic and quantum architectures
- Development of efficient consciousness measurement algorithms for large-scale systems
- Investigation of hybrid biological-artificial conscious systems
- Exploration of ethical frameworks for conscious AI development

10. Conclusion

We have confirmed through mathematical analysis and computational validation that feedforward AI architectures necessarily yield zero integrated information under both IIT 3.0 and 4.0 formalisms. This fundamental result applies regardless of scale, complexity, or architectural sophistication, including contemporary systems like transformers, CNNs, and reinforcement learning agents.

Our empirical validation across 16 diverse network configurations achieved complete consistency with theoretical predictions, with 100% of feedforward systems yielding $\Phi = 0$ and 75% of recurrent systems exhibiting $\Phi > 0$.

The path to conscious AI requires architectural innovation beyond scaling current feedforward approaches. Understanding these requirements is crucial for scientific progress, ethical consideration, and technological development as we navigate the complex landscape of artificial minds.

Whether humanity should pursue conscious AI remains an open question requiring careful consideration of benefits, risks, and ethical implications. Our work establishes a mathematical foundation for AI consciousness assessment that will become increasingly important as AI systems grow more sophisticated.

References

1. G. Tononi, "An information integration theory of consciousness," *BMC Neuroscience*, vol. 5, pp. 42, 2004.
2. M. Oizumi, L. Albantakis, and G. Tononi, "From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0," *PLOS Computational Biology*, vol. 10, no. 5, pp. e1003588, 2014.
3. L. Albantakis, M. Massimini, M. Rosanova, and G. Tononi, "Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms," *arXiv preprint arXiv:2212.14787*, 2022.
4. W. G. P. Mayner, W. Marshall, L. Albantakis, G. Findlay, R. Marchman, and G. Tononi, "PyPhi: A toolbox for integrated information theory," *PLOS Computational Biology*, vol. 14, no. 7, pp. e1006343, 2018.
5. P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. Fleming, et al., "Consciousness in artificial intelligence: Insights from the science of consciousness," *arXiv preprint arXiv:2308.08708*, 2023.
6. G. Marcus and E. Davis, "GPT-3, consciousness, and the hard problem of AI," *Communications of the ACM*, vol. 66, no. 7, pp. 54–63, 2023.

 doi:10.20944/preprints202510.1665.v2

7. M. Mitchell, "The debate over understanding in AI's large language models," *Proceedings of the National Academy of Sciences*, vol. 120, no. 13, pp. e2215907120, 2023.
8. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, et al., "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.
9. D. J. Gunkel, *Robot rights*, MIT Press, 2018.