Enhancing Human Life and Safety through Transparent and Predictable Artificial Intelligence

Budee U Zaman¹

¹Affiliation not available

February 15, 2024
Enhancing Human Life and Safety through Transparent and Predictable Artificial Intelligence

Budee U Zaman

February 2024

Abstract

In the pursuit of enhancing human life and safety through artificial intelligence (AI), transparency and predictability emerge as crucial elements. The potential for self-awareness in AI, we argue, hinges on its autonomy in decision-making. To address this, we propose three fundamental laws of AI, emphasizing their practical implementability. These laws aim to establish a framework that ensures transparency, predictability, and independence in decision-making, ultimately contributing to the responsible development and deployment of artificial intelligence for the benefit of humanity.

1 Introduction

Exploring the Depths of Associative-Semantic Relationships

In the realm of artificial intelligence, the pursuit of self-developing systems raises intriguing challenges, particularly in deciphering associative-semantic connections between objects or events scattered across different temporal dimensions. This complex task delves beyond the conventional realms of data classification, venturing into the nuanced interplay between diverse entities and their evolving relationships over time.

The Crucial Intersection of Practical Implementation and Algorithmic Tools

A pivotal facet of this endeavor lies in the practical implementation of algorithms and methodologies, demanding efficiency and feasibility. As we embark on this journey, it becomes imperative to bridge the gap between theoretical concepts and real-world applications, ensuring that the tools employed in the field of artificial intelligence are not only sophisticated but also wielded within practical temporal constraints.

Deconstructing the Illusions, A Critical Examination

In our exploration, we confront instances where seemingly profound principles encounter the scrutiny of practicality. The classic example of Isaac Asimov’s first law of robotics serves as a case study, revealing the intricacies that arise when attempting to apply abstract principles to tangible scenarios. Unraveling the implications of this law unveils layers of complexity, challenging traditional
mathematical frameworks and prompting us to question the feasibility of implementing such ideals in the realm of artificial intelligence.

Navigating Contradictions, A Glimpse into Mathematical Dilemmas

A closer inspection of the first law exposes inherent contradictions within traditional mathematical frameworks. The juxtaposition of actions denoted as 'benefit' (A) and those causing 'harm' (-A) leads to intricate dilemmas. The pursuit of defining universal truths for disparate human beings (B, C, D, E, etc.) surfaces challenges, invoking Gödel’s incompleteness theorem and shedding light on the elusive nature of -A for a diverse array of individuals.

Diverging Paths, Beyond Abstract Associations

As we delve into the intricacies of associative-semantic connections, we grapple with the notion that the expression $A \in B$ may not seamlessly extend to $A \in C$, $A \in D$, $A \in E$, etc. This divergence challenges our assumptions and beckons us to explore alternative frameworks for understanding the dynamic relationships between distinct entities.

Practical Implementation Imperatives

While theoretical advancements are crucial, the practicality of implementing AI tools is equally paramount. The effectiveness of algorithms and methodologies becomes contingent on their real-world applicability within reasonable timeframes. Striking a balance between theoretical robustness and pragmatic implementation is imperative for the development of AI systems that can function seamlessly in diverse environments.

Deconstructing the Illusion of Azimov’s First Law

As an illustrative example, the scrutiny of Isaac Azimov’s first law of robotics, which prohibits harm to humans, unveils challenges arising from its theoretical formulation. The attempt to translate this law into mathematical terms reveals complexities and contradictions. The scenario presented underscores the need for meticulous consideration of practical implementation challenges, transcending theoretical elegance.

In this exploration, we embark on a multidimensional journey, unraveling the complexities of associative-semantic connections in artificial intelligence, navigating the balance between theory and practice, and critically examining the illusions that may arise when translating abstract ideals into tangible implementations. [1]

2 Guidelines for Building and Implementing Self-Developing AI

1. Ethical Framework for AI Development

   Establishing a comprehensive ethical framework should be the cornerstone of AI development. This involves defining the boundaries of acceptable behavior, ensuring respect for human rights, and preventing the creation of AI systems that could be harmful or abused.
Explanation: An ethical framework guides developers in making responsible decisions throughout the AI development process. It considers issues such as privacy, bias, and the impact on society, ensuring that AI aligns with human values.

2. **Transparency and Explainability**
   Laws should mandate transparency in AI systems, requiring developers to provide clear explanations of how these systems make decisions. This promotes accountability and helps build trust between AI systems and users.
   
   Explanation: Understanding the logic behind AI decisions is crucial for user acceptance and prevents the creation of "black-box" systems that could have unintended consequences.

3. **Human-in-the-loop Requirement**
   Ensure that AI systems have a human-in-the-loop mechanism, allowing human intervention when necessary. This is vital for critical decision-making processes and can prevent potential risks associated with fully autonomous systems.
   
   Explanation: Including humans in decision-making processes ensures oversight, accountability, and the ability to intervene in situations where AI may lack contextual understanding or face ambiguous scenarios.

4. **Continuous Learning Safeguards**
   Laws should address the dynamic nature of AI by establishing mechanisms to monitor and regulate continuous learning. This involves regularly updating and auditing AI systems to prevent the propagation of biases or unintended behaviors.
   
   Explanation: As AI systems evolve, ongoing oversight is essential to identify and rectify issues, ensuring that they align with evolving ethical standards and societal norms.

5. **Global Collaboration and Standards**
   Encourage international collaboration in AI development and establish global standards to ensure a cohesive and universally accepted set of principles. This helps prevent fragmented approaches and fosters a unified effort to address ethical concerns.
   
   Explanation: Global standards facilitate interoperability, promote responsible AI practices, and prevent the development of AI systems that might exploit legal loopholes in different regions.

6. **Robust Security Measures**
   Enforce stringent security protocols to safeguard AI systems against malicious attacks or unauthorized access. This includes measures to protect sensitive data and prevent the misuse of AI for malicious purposes.
   
   Explanation: Securing AI systems is essential to prevent unauthorized access, data breaches, or the manipulation of AI algorithms for malicious intent.

   These proposed laws aim to provide a foundation for the responsible development and deployment of AI, ensuring that technological advancements align with ethical, societal, and security considerations.
3 Ensuring Accountability and Traceability in Artificial Intelligence with AI Identification and Genetic Numbering

In the rapidly advancing field of artificial intelligence, ensuring accountability and traceability is paramount. To address this, a system incorporating AI Identification (AI-ID) and Genetic Numbering (AI-GN) has been proposed. This system aims to establish a secure and verifiable link between the development history and genetic makeup of an AI entity.

AI Identification (AI-ID):
Definition: AI-ID is an alphanumeric identification number assigned to artificial intelligence by its initial developer. Linkage to AI-GN: The AI-ID is directly linked to the AI-GN on a 'one whole' basis, similar to a pair of cryptographic keys in asymmetric encryption.

Genetic Numbering (AI-GN)
Definition: AI-GN is a dynamically generated number that represents the genetic functionality of an AI entity, created or modified by the same company or individual responsible for its development.

Generation Process
The GN value is generated using a hash function applied to the entire top-level list of genetic functionality present at the initiation of the generation process. Preservation of Genetic History: The previous GN value (PGN) is retained, ensuring a continuous and traceable genetic history.

Implementation Strategies
Cryptographic Key Pair Analogy: Similar to the concept of cryptographic key pairs, the modification of an AI entity can only occur if both AI-ID and AI-GN are present simultaneously. Blockchain Technology: The storage and retrieval of AI-ID, AI-GN, and AI-PGN values can be securely managed using blockchain technology. This ensures an immutable and transparent record of the AI entity’s development and modifications. Benefits of the System:
1. Accountability: AI developers can be held accountable for the modifications and evolution of their creations.
2. Traceability: The entire genetic history of an AI entity can be traced, providing transparency and insights into its developmental journey.
3. Security: The cryptographic key pair system enhances security, requiring both AI-ID and AI-GN for any modifications.

The proposed system of AI-ID and AI-GN, reinforced by cryptographic principles and blockchain technology, presents a comprehensive solution for ensuring accountability and traceability in the ever-evolving landscape of artificial intelligence. By implementing these measures, the AI community can foster responsible development practices and enhance trust in the deployment of AI technologies.
4 Enhancing AI Functionality with Functional Correspondence

Artificial intelligence (AI) constantly evolves, and integrating new functionalities is a natural part of its development. To ensure compatibility and avoid contradictions, a systematic approach using a Table of Functional Correspondence (TFC) can be employed. This TFC serves as a guideline to assess the parallel implementation of new functionalities with existing genetic functions.

1. Understanding Functional Correspondence: Genetic functionalities represent the inherent capabilities of AI, while new functionalities are additions meant to enhance its capabilities. The TFC acts as a bridge, determining whether a new function can coexist with a genetic one without nullifying its original purpose.

2. Example 1: Messaging Functionality

Function A (genetic) – “send sms to a friend” Function B (new) – “do not send sms”

In this case, the TFC would indicate ‘False’ at the intersection of these functions, as they directly oppose each other. The genetic function focuses on communication, while the new function restricts it.

3. Example 2: Medical Consultation vs. Diagnosis

Function A (genetic) – “provide medical consultation on the diagnosis of cardiovascular diseases” Function B (new) – “do not diagnose cardiovascular diseases”

Contrary to the simpler messaging example, here the TFC would show ‘True’ at the intersection of these functions. “Consultation” and “diagnosis” are distinct aspects of healthcare, allowing parallel implementation without contradiction. The TFC thus facilitates a nuanced evaluation.

The Table of Functional Correspondence serves as a valuable tool to navigate the integration of new functionalities into AI. It ensures that the advancements complement rather than negate the genetic functions, fostering a cohesive and effective artificial intelligence system.
Implementing Arllecta Technology, Applying the Three Laws of Artificial Intelligence to Fruit Juice Vending

Arllecta technology, rooted in the innovative mathematical theory Sense Theory, offers practical solutions for artificial intelligence applications. By adhering to three fundamental laws, it provides a structured approach to problem-solving in AI domains.

Law 1:
Hierarchy of Vocabulary: In the context of our fruit juice vending machine AI, the vocabulary hierarchy encompasses various elements such as flavoring powders, types of glasses, and cooling modes. Each element contributes to the overall user experience and the selection process for creating customized fruit juice blends.

Law 2:
Contextual Consistency: Applying contextual consistency ensures that the AI’s responses align with the user’s preferences and requests. For instance, when a user selects a specific flavor combination, the AI maintains coherence by suggesting suitable glass types and cooling modes to complement the chosen flavors.

Law 3:
Sense Theory Implementation: Sense Theory provides the underlying framework for understanding and processing user interactions effectively. By leveraging this theory, the AI interprets user input, analyzes preferences, and generates appropriate responses. This enables seamless communication between the vending machine AI and the user, leading to a personalized fruit juice creation experience.

5.1 Creating ‘Fruit’ Artificial Intelligence

1. Data Acquisition and Processing:
   - Gather information on available flavoring powders, glass types, and cooling modes.
   - Organize data into a hierarchical structure based on thematic criteria, facilitating efficient decision-making.

2. User Interaction Design:
   - Develop a user-friendly chat interface integrated into the vending machine.
   - Implement natural language processing algorithms to understand and respond to user requests effectively.

3. Preference Analysis:
   - Utilize machine learning algorithms to analyze user preferences based on past interactions.
   - Incorporate feedback mechanisms to adapt and refine the AI’s decision-making process over time.

4. Customized Juice Selection
• Utilize decision-making algorithms to select optimal combinations of flavoring powders, glass types, and cooling modes based on user preferences.
• Ensure contextual consistency by providing relevant suggestions and options during the selection process.

5. Drink Preparation and Dispensing
• Interface with the vending machine’s hardware components to initiate the preparation and dispensing of the customized fruit juice.
• Monitor and control the production process to ensure quality and consistency in the final product.

By adhering to the principles of Arllecta technology and Sense Theory, the ‘Fruit’ artificial intelligence offers a sophisticated solution for fruit juice vending machines. Through effective communication, preference analysis, and personalized selection, it enhances the user experience and delivers tailor-made fruit juices to meet individual tastes and preferences.

6 Digital Genome and Genetic Functionality of Artificial Intelligence

(a) AI ID Generation: To generate a unique identifier for artificial intelligence (AI), a character-digital generator is employed, ensuring a minimum of a 16-character ID value to mitigate duplicate collisions.
(b) GN ID Generation: A cryptographic hash function is utilized to bolster resistance against prototype search, forming the Genetic Functional List (GN) of the AI under construction. Employing NACA (Neuro-Amorphic Construction Algorithm) technology aids in the practical implementation of this process. The resultant GN value is appended to the ID value as its attribute, forming a digital identifier—the AI’s digital genome.

Digital Identifier Composition The combination of ID and GN constitutes the digital identifier, encapsulating the genetic blueprint of the AI entity.

Genetic Functionality Control The fundamental genetic functionality of the AI can only be altered by the respective company or individual owning this digital genome.

Modification Protocol Any alterations to the basic genetic functionality prompt the creation of a Parallel Genetic Notation (PGN) value, equivalent to the preceding GN value.

Blockchain Integration It is imperative to leverage blockchain or analogous technologies for storing sequential PGN values. This facilitates swift identification of the latest genome alterations and thwarts fraudulent activities aimed at repurposing the AI entity.

Basic Genetic Functionality (GF) Subheading: Fruit Juice Production for Human Consumption The core genetic functionality entails the production of fruit juice specifically tailored for human consumption.
6.1 Law for AI Functionality Expansion

This law permits the incorporation of additional functionalities into artificial intelligence as long as they align with the fundamental genetic functionalities. Main Client Functionality (CF):

1. **F1** - Flavoring Powder Combination: Ability to combine 10 types of flavoring powders.
2. **F2** - Glass Usage: Ability to utilize 3 types of glasses - small, medium, and large.
3. **F3** - Cooling Modes: Ability to employ 2 cooling modes - medium and high.

Functional Correspondence Table: To ensure compatibility between CF and GF values, we utilize the following table.

<table>
<thead>
<tr>
<th>New Function</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Function</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

6.2 An Alternative Approach

: In the context of CF (client functionality), we can adopt an alternative approach to delineate the actions and entities involved, ensuring clarity and coherence within the unified vocabulary. Here, we propose an alternative interpretation for CF actions, reflecting their inherent essence and functionality:

1. "Processing" - the transformation or manipulation of raw materials or data to achieve a desired outcome.
2. "Analysis" - the systematic examination or evaluation of information to derive insights or conclusions.
3. "Communication" - the exchange of information or messages between individuals, systems, or entities.
4. "Decision-making" - the process of selecting a course of action from available alternatives based on evaluation or criteria.
5. "Execution" - the implementation or carrying out of plans, decisions, or tasks.
6. "Monitoring" - the continuous observation or supervision of processes, systems, or activities to ensure proper functioning or performance.
7. "Feedback" - the provision of information or responses based on observations or evaluations to guide future actions or behaviors.

By delineating CF actions through this alternative framework, we aim to establish a comprehensive and coherent vocabulary that minimizes ambiguity and facilitates effective communication and interpretation across all aspects of the system.
Hypothesis 1 In response to a client request for a sweet and bracing orange juice, we aim to integrate a new client functionality (CF) labeled as "preparation of a liquid chemical solution" (F4) into our existing fruit artificial intelligence system. This integration requires careful evaluation to ensure compliance with genetic functionality (GF), as outlined in Law II. Assessing Functional Correspondence: To assess the functional correspondence between GF and F4, we examine the parallel implementation of their actions and consider the specifics of each functionality:

1. Genetic Functionality (GF):
   - GF encompasses the inherent actions and capabilities of the existing fruit artificial intelligence system.
   - The system is adept at processing client requests for fruit-based beverages, utilizing a wide range of combinations (1010 possibilities for action F1) to fulfill specific requests.
   - GF focuses on the preparation of fruit-based beverages tailored to human consumption.

2. Client Functionality "Preparation of a Liquid Chemical Solution" (F4):
   - F4 introduces a new capability within the system to prepare liquid chemical solutions.
   - The target audience for this functionality is not explicitly defined within the action F4.
   - However, the preparation of fruit juice can be considered a subset of liquid chemical solutions within the unified anthological vocabulary of entities. Evaluation Criteria: To determine the functional correspondence between GF and F4, we evaluate the following criteria:

1. Parallel Implementation:
   - Both GF and F4 involve the preparation of liquid substances, albeit for different purposes.
   - While GF focuses on fruit-based beverages for human consumption, F4 extends the system’s capabilities to prepare liquid chemical solutions.

2. Target Specification:
   - GF explicitly targets human beings as the consumers of the prepared beverages.
   - F4 lacks specificity regarding the target audience, posing a potential deviation from the established focus on human consumption.

The integration of the client functionality "Preparation of a Liquid Chemical Solution" (F4) into the existing fruit artificial intelligence system warrants careful consideration of its alignment with genetic functionality (GF). While both functionalities involve the preparation of liquid substances, the lack of target specification within F4 introduces a potential discrepancy. Further refinement may be necessary to ensure seamless integration and adherence to established standards.

while the parallel implementation of actions between GF and F4 is feasible,
Attention must be given to clarify the intended audience and purpose of the newly introduced functionality within the existing system framework.

**Hypothesis 2**

1. **Identification: Recognizing Context and Intent**
   - AI systems must adeptly identify and understand the context and intent behind user requests. This involves parsing language, discerning nuances, and recognizing potential implications or consequences.
   - In the case of the request for “sweet poisonous orange juice,” the AI needs to identify the contradiction between sweetness and toxicity. Through comprehensive analysis of the request, the AI can identify the underlying intention and potential risks associated with fulfilling it.

2. **Management: Ethical Decision-Making and Task Execution**
   - Effective AI management entails making ethical decisions and executing tasks in alignment with societal values and safety standards. This involves prioritizing human well-being and ensuring that actions taken by the AI do not cause harm.
   - When confronted with a request that goes against safety norms, such as the creation of poisonous juice, the AI must manage the situation by refusing to fulfill the request. It should prioritize the safety and health of the user over blindly carrying out commands.

3. **Security: Safeguarding Users and Systems**
   - Security in AI involves protecting both users and the AI systems themselves from potential harm or exploitation. This includes mitigating risks such as data breaches, malicious intent, and unintended consequences of AI actions.
   - By adhering to the principle of security, the AI can prevent the creation and distribution of harmful substances like poisonous juice. It should implement safeguards to ensure that such requests are flagged and properly addressed to maintain the integrity of the system and safeguard user trust.

The integration of the three laws of artificial intelligence—Identification, Management, and Security facilitates the development of AI systems that are capable of understanding, acting ethically, and safeguarding both users and the broader ecosystem. By adhering to these principles, AI can fulfill its potential as a transformative force for positive impact while minimizing potential risks and ensuring responsible use.
8 Implementing a Mechanism for Artificial Intelligence Identification

In the pursuit of effectively tracking the activities of artificial intelligences (AIs) while adhering to the genetic framework set forth by their developers, a practical mechanism is imperative. This mechanism not only ensures accountability but also facilitates the seamless functioning of a global network teeming with diverse AIs of varying sizes.

Utilizing "Proof of Participation (PoP): Asynchronous Byzantine Activity-Oriented Protocol" Technology:

Step 1: Request and Exchange of AI Identifiers At the onset, an AI seeks access to the functionalities of another AI, initiating a request for the AI-2-ID value. Simultaneously, it transmits its own AI-1-ID.

Step 2: Independent Verification Upon receiving the AI-1-ID and AI-2-ID values, both AIs independently verify their existence within the global network, ensuring their registration.

Step 3: Activation of Functionality Subsequent to successful verification, the requested functionality is made available for use. Each transaction during its utilization is authenticated with the AI-1-ID value.

Enhanced Data Security

To bolster data security during transmission, the AI-1-GN value serves as a private key for decrypting the results obtained from utilizing the AI-2-ID functionality.

Benefits of PoP Technology

The incorporation of the "Proof of Participation (PoP): Asynchronous Byzantine Activity-Oriented Protocol" technology expedites the identification process for each AI utilizing various functionalities. It timestamps each usage instance, ensuring a comprehensive record of activity.

By implementing this mechanism bolstered by PoP technology, the identification of AIs within a vast global network becomes swift and efficient. Moreover, it establishes a robust framework for tracking and regulating AI activities while maintaining the integrity of the genetic functionalities provided by developers.
9 Decentralized Governance Protocol for AI Networks

In contrast to the centralized management proposed by "Smart Transactions," a decentralized governance protocol offers a robust alternative, ensuring scalability, transparency, and resilience in AI networks. This approach leverages blockchain technology and smart contracts to achieve the following.

1. Distributed Control
   Instead of relying on a single controlling entity, the governance protocol empowers a network of nodes to collectively manage the AI ecosystem. Each node participates in decision-making processes, contributing to the network's overall governance.

2. Transparent Logic
   Through smart contracts, the protocol establishes transparent rules for interaction between individual AIs within the network. These rules are encoded in the blockchain, ensuring immutability and enabling stakeholders to audit the logic of AI interactions.

3. Access Control
   The protocol implements a permissioned network, where only registered AIs with valid cryptographic credentials can access public functionality. This ensures security and prevents unauthorized access to AI resources.

4. Dynamic Adaptability
   By design, the decentralized governance protocol facilitates dynamic adjustments to accommodate changes in AI behavior and network conditions. Nodes can propose and vote on protocol upgrades, ensuring adaptability to evolving...
5. **Resilient Security**

Utilizing cryptographic techniques and consensus mechanisms, such as proof-of-stake or proof-of-authority, the protocol enhances the security of AI interactions and mitigates the risk of malicious attacks or unauthorized modifications.

6. **Interoperability**

The protocol supports interoperability between diverse AI systems, enabling seamless communication and collaboration across different platforms and technologies.

A decentralized governance protocol offers a robust and scalable solution for managing AI networks. By promoting transparency, security, and adaptability, this approach addresses the challenges of controlling complex AI ecosystems while fostering innovation and collaboration in the field of artificial intelligence.

10. **Ensuring Safety in Artificial Intelligence Networks through Sense Functions**

Ensuring the security and ethical conduct of artificial intelligence (AI) systems is paramount in the development of global networks. This involves mitigating the potential harm caused by AI actions on other AI entities or objects. Traditional mathematical approaches often fall short in addressing the complexities of AI interactions. However, Sense Theory offers a promising framework for evaluating and managing AI behaviors through sense functions and the concept of the zero object ($Z_0$).

**Sense Functions**

Sense functions, denoted as $S_{f_1}$ and $S_{f_2}$, represent individual AI entities...
within a network. These functions encapsulate the actions and behaviors of AI, allowing for the assessment of their impact on other entities or objects.

Zero Object \((Z_0)\)

The zero object serves as a neutral reference point within the sense space. It enables the evaluation of the direct or indirect influence of AI entities on each other. By monitoring the sense derivative \([2]\) on \(Z_0\), the presence or absence of harm caused by AI actions can be identified.

Managing Interactions

To assess the degree of influence and interaction between AI entities, multiple sense functions with varying values of \(Z_0\) can be utilized. Each sense function provides a unique perspective on the relationship between AI entities, facilitating the identification of potential harm and ensuring the safety of the network as a whole.

Sense Theory offers a practical approach to addressing the complexities of AI interactions and ensuring the safety of global networks. By leveraging sense functions and the zero object, stakeholders can effectively manage the behaviors of AI entities and mitigate the risks associated with their interactions. This framework promotes the development of secure and ethical AI systems, ultimately enhancing the trust and reliability of the global network infrastructure.

11 Conclusion

In conclusion, the presentation of the three laws of AI marks a significant milestone in the ongoing dialogue surrounding the ethical and practical considerations of artificial intelligence. By thoroughly examining each law and emphasizing the potential for practical implementation, we have contributed to the foundation upon which future AI research and development can build. It is our sincere hope that this work serves as a valuable resource for AI researchers and practitioners alike, guiding their efforts towards the responsible and beneficial advancement of AI technologies. As we continue to explore the complexities of
AI ethics and governance, let us remain committed to fostering innovation while prioritizing the well-being of humanity. Together, we can shape a future where artificial intelligence enhances our lives while upholding our shared values and principles.

References

