Ifiok: Journal of Interdisciplinary Studies Vol. 8, No. 2, Dec., 2024
ISSN 2408-6800
http://ifiokjournal.com

Beyond the Halting Problem: Undecidability in Formal Logic and its Implications for AI Research

UKANGA, Abasiofon Sunday
Department of Philosophy
Akwa Ibom State University, Nigeria
ukangaabasiofon@gmail.com

&

UDOFIA, Christopher Alexander, Ph.D
Department of Philosophy
Akwa Ibom State University, Nigeria
christopherudofia@aksu.edu.ng

Abstract

This paper, *Beyond the Halting Problem: Undecidability in Formal Logic and its Implication for AI Research* examines the Turing test and the associated halting problem or undecidability problem. The undecidability problem is a position which argues that given a logical system, there are arithmetic formulas which are true but cannot be proven true within the system. Alan Turing showed how that a computer program will either halt or run indefinitely given a certain input, proving that the halting problem is undecidable for all computer programs. This limitation in the traditional rule based systems in logic and artificial intelligence (AI) necessitates an exploration of alternative approaches that go beyond the rigid confines of predefined rules. The paper examines the shortcomings of rule based systems in handling complex, uncertain, and dynamic environments, highlighting their inability to adapt, learn, and generalize effectively. It then explores as alternatives; the probabilistic methods, connectionist models and reinforcement learning, analyzing their strengths in dealing with AI applications. The paper concludes by discussing the future of AI, ethical issues raised and the need for a robust and holistic approach which integrates diverse methods and promotes collaboration between different disciplines.

Keywords: Undecidability, Halting Problem, Artificial Intelligence, Embodied AI

Introduction

Alan Turing, a visionary mathematician and computer scientist, is widely recognized as one of the founding fathers of artificial intelligence (AI). His seminal work, "Computing Machinery and Intelligence," published in 1950, laid the groundwork for the field, proposing the now-famous Turing Test as a means of assessing a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. While the Turing Test has sparked significant debate and shaped much of AI research, a closer examination of Turing's broader vision reveals a deeper understanding of intelligence, particularly the crucial role of learning. In his paper, Turing made a profound statement: "a machine that could learn from its mistakes would be able to acquire much more knowledge than a machine that could not."

This statement, often overlooked in discussions of the Turing Test, foreshadowed a paradigm shift in AI research, moving beyond the limitations of rule-based systems and embracing the power of learning from experience. Rule-based systems, prevalent in early AI, relied on predefined sets of instructions or rules, limiting their ability to adapt to changing environments and unexpected situations. Turing's vision of learning machines, however, anticipated the development of AI systems that could adapt and improve their performance over time through interaction with the world.

Akman (2000) argues that “the Turing test, inspite of its intuitive appeal, is vulnerable to a number of justifiable criticisms. One of the most important of these criticisms is its bias toward purely symbolic problem solving tasks”. It does not test abilities requiring perceptual skill or manual dexterity even though these are important components of human intelligence. Formal systems, such as axiomatic set theory or propositional logic, provide a framework for rigorous and precise mathematical reasoning. These systems typically consist of axioms, or basic truths, and rules of inference, which allow us to derive new truths from existing ones. The goal is to create a system that is both powerful enough to encompass a wide range of mathematical knowledge and consistent enough to avoid contradictions.

The concept of undecidability raises profound questions about the nature of knowledge and the limits of our understanding. Undecidability suggests that our knowledge is inherently incomplete, as there will always be truths that are beyond our ability to prove or disprove definitively. According to Harnish, “Formal systems, while powerful tools for reasoning, cannot capture all truths”(Harnish, 2002 p.31). They provide a framework for rigorous deduction but are fundamentally limited in their scope. Undecidability highlights the importance of intuition, experience, and other non-formal methods in guiding our understanding of the world. These methods can often provide valuable insights that formal systems cannot capture.

The concept of undecidability, while seemingly negative, offers a valuable perspective on the limits of our knowledge and the power of computation. It compels us to embrace the inherent complexity of the world and to explore alternative approaches to reasoning and problem-solving.The discovery of undecidability problems within formal logic has profoundly impacted our understanding of knowledge, computation, and the limitations of reasoning. While undecidability presents challenges, it also opens up new avenues for exploration and understanding. By embracing the inherent incompleteness of knowledge and exploring alternative approaches, we can continue to advance our understanding of the world and the limits of our own reasoning.

The Halting Problem and Undecidabilty:The Power and Limits of Formal Systems

Interestingly, Turing machines, which have abstract properties and hinge upon how algorithms work, were proposed to deal with a very pressing problem in mathematics in the 1930's, the *Entscheidungsproblem* originally posed by the mathematician David Hilbert in 1928(Brodkorb 2019, pp 1), this problem consists in whether or not it is possible to find a general algorithmic procedure to resolve *in principle* all the mathematical problems. Kurt Godel's groundbreaking work in the 1930s,*On Formally Undecidable Propositions Of Principia Mathematica and Related Systems,* in his incompleteness theorems demonstrate that any sufficiently powerful formal system, capable of expressing basic arithmetic, will inevitably

contain statements that are true but cannot be proven within the system itself. This means that even with a perfectly consistent set of axioms and rules, there will always be truths that lie beyond the reach of formal proof.

Godel's first incompleteness theorem states that any consistent formal system containing basic arithmetic is incomplete, it cannot prove all true statements expressible within the system(Godel,1992, p.173). This theorem was a profound revelation, demonstrating that there are inherent limitations to the power of formal systems to capture all mathematical truths.

The second incompleteness theorem goes even further, proving that no consistent formal system containing basic arithmetic can prove its own consistency(Godel, 1992, p.175). This result has significant implications for the foundations of mathematics, suggesting that we cannot rely on formal systems alone to definitively establish the truth of all mathematical statements. Godel's theorems, while focused on arithmetic, have profound implications for other formal systems. The concept of undecidability extends beyond arithmetic and applies to a wide range of logical formalisms, including:

- **Propositional Logic:** While seemingly simpler than arithmetic, propositional logic, dealing with truth values and logical connectives, also exhibits undecidable problems. For example, the "satisfiability problem" for propositional logic, asking whether a given logical formula can be made true by assigning truth values to its variables, is known to be NP-complete,

implying its computational complexity and undecidability for large instances.

- **First-Order Logic**: First-order logic, which allows quantification over variables and introduces predicates, also suffers from undecidability. The "decision problem" for first-order logic, determining whether a given formula is universally true, has been proven undecidable. This means that there is no algorithm that can always determine the truth value of any first-order logic formula.
- **Modal Logics**: Modal logics, which introduce modal operators like "possibly" and "necessarily", are also subject to undecidability. Deontic and epistemic modal logics have been proven to be undecidable, meaning there is no general procedure for determining the truth value of all formulas in those logics.

Turing, then, postulated a revolutionary idea to deal with this issue, the terms ‘mechanical procedure’ and ‘machine’ had to be formalized. He then imagined a mechanical device that performs a finitely definable calculation procedure, that is, an idealized *abstract* machine that has a discrete set of different possible states, which are to be finite in number (even though it may be a very large number).

Despite this apparent limitation, the machine possesses no limit as to the possible calculations, as it has a set of instructions that act with independence of the size of the numbers. Another important point to bear in mind is that the input need not be restricted in size.

Actually the machine uses an unlimited *external* storage capacity, which is usually depicted as paper that serves for doing the calculations and producing the output. This in turn need not be limited in size either.

The machine, however, is not supposed to internalize the external data or the calculations. It deals with the data or calculations that are immediately involved in the operation carried out. The potentially unlimited size of the input versus the finitude of states and steps for the calculations suggests, again, that Turing machines are abstractions rather than machines one could actually construct. This point is remarked by Penrose thus:

> it is the unlimited nature of the input, calculation space, and output which tells us that we are considering only a mathematical idealization rather than something that could be actually constructed in practice…The marvels of modern computer technology have provided us with electronic storage devices which can, indeed, be treated as unlimited for most practical purposes (Penrose 1989 p.35)

These features, which resemble modern computers, as will be examined, and the fact that the TM is only an idealization explain why typical examples are usually pictured as an indefinitely long tape divided into squares (or cells). Additionally, a 'head' or scanner moves backwards and forwards along this tape, square by square, 'remembering' some of the scanned symbols or discrete states. At any given time *t*, the head, which is in an internal state (q_0, q_1, q_2,

... ,q_n), scans a particular square of the tape and its symbol (b_1, b_2, b_3, ... , b_n). On the basis of that particular symbol, the internal configuration of the head, and its microcode or program, the head prints or erases a symbol and then may proceed to the next square (provided that it has not yet reached the final stage of the computation). This process, which goes in accordance with the instructions of the head, is repeated until the machine has reached the state that represents the solution of a problem, which is printed on the tape.

In the case of a typical TM, with numbers 0 and 1 for the squares —for the sake of simplicity— the possible behaviour of the machine, which is *completely* determined by the head and its internal state at *t*, is the following:

1) The head reads the symbol that is in the square, which, combined with its current state, constitutes an input I_n. This consists of a pair <Current State, Read Symbol>

2) Given I_n and the microcode program of the head, the machine tries to determine an output triple, O_n, that is, <Write Symbol, Move Step (or stop), New State>. If the machine is unable to determine an Out, it *halts* (and so does it, when the computation has finished).

3) On the basis of O_n, the machine writes a symbol (in this case 1 or 0) in the square, moves the head to the left or right (or stays where it

is), and enters a new state. Eventually the machine returns to step 1, if it has not reached the final state.

Step 2 is fundamental to grasp how a TM works, for it turns inputs into outputs by a *program* that gives *a finite set of rules expressed in pairs* <In, Out>. In fact, the program, which generally comprises IF-THEN rules, can be expressed as quintuples of the form <Current State, Read Symbol, Write Symbol, Move Step (or stay), New State>. Given its internal configuration and the scanned symbols, the head only *follows* the microcode program rules until it finds the pair <In, Out>, and proceeds to print the output. If the machine does not find that pair, it gets 'stuck', so to speak, and halts. However, if the machine finds the pair, it produces an output, and keeps computing until it has found the solution.

Beyond Rule-Based Systems Exploring Alternative Approaches to Circumventing Undecidability in Artificial Intelligence

The idea of embodiment in intelligence, deeply rooted in phenomenological and pragmatic philosophies, challenges the Cartesian dualism that separates mind and body. It emphasizes the essential role of a physical body in shaping our perception of the world, our actions within it, and our overall understanding of intelligence. Merleau-Ponty, a French phenomenologist, argued that the body is not simply a passive instrument of the mind but rather an active and dynamic participant in our experience of the world (Merleau-Ponty 2012, 125). He emphasized the importance of the body's perceptual and motor capabilities in shaping our understanding

of the world. Perception, for Merleau-Ponty, is not a passive reception of sensory information but rather an active process of embodiment and engagement with the environment. Heidegger, argued that human existence is fundamentally characterized by "being-in-the-world," meaning that we are always already embedded in a world that is both physical and meaningful (Heidegger 1962, pp.22; Lawhead 2015, pp.536). This perspective suggests that intelligence is not merely a matter of abstract thought but rather arises from our practical engagement with the world, including our bodily interactions with objects and our understanding of the tools and technologies we use. John Dewey, emphasized the importance of experience in learning, arguing that knowledge is not simply acquired through passive reception of information but rather through active engagement with the world(godfrey-smith 2013, pp.286). This perspective, aligned with the principles of pragmatism, suggests that learning is inherently embodied, involving the use of our senses, motor skills, and our interaction with our environment.

Turing, influenced by the rise of computing and the formalization of logic, focused primarily on the computational aspects of intelligence. His work emphasized the idea of a universal Turing machine, capable of simulating any computation and potentially replicating human intelligence through a combination of algorithms and symbols. The Turing Test, proposed by Turing in his seminal 1950 paper "Computing Machinery and Intelligence," sought to determine whether a machine could exhibit intelligent behavior indistinguishable

from that of a human. The test involved a human evaluator interacting with both a human and a machine through a text-based interface, judging the machine's ability to generate responses that were indistinguishable from those of a human. This focus on linguistic competence, however, neglects the embodied nature of intelligence, potentially overlooking crucial aspects of human cognition. Critics argue that the Turing Test only assesses the machine's ability to mimic human behavior, rather than demonstrating a true understanding of the world. This points to the possibility of a machine passing the test without possessing genuine understanding or intelligence.

While Turing's primary focus was on the computational aspects of intelligence, he did acknowledge the importance of physical embodiment in certain scenarios. In his 1950 paper, he noted that "a machine that could learn from its mistakes would be able to acquire much more knowledge than a machine that could not."(Turing 1950, 435) This implies a recognition that learning, a crucial aspect of intelligence, is inherently linked to the ability to act in the world and receive feedback from those actions. Embodied intelligence argues that true understanding arises from the interaction of a physical body with the environment. The Turing Test, however, relies on a disembodied text-based interface, neglecting the crucial role of sensory experience, motor action, and the body's influence on cognition. Embodiment is crucial for learning. The Turing Test does not fully account for the complexities of embodied learning, which

involves the ability to interact with the world, receive feedback, and adapt behavior based on those experiences. We shall limit our research to the connectionist, probabilist and reinforcement learning models of artificial intelligence (AI).

The Connectionist models of AI: Embodied Intelligence, Moving Beyond Symbolic Representations

While rule-based systems, often referred to as expert systems, were successful in well-defined domains, they struggled to generalize and adapt to complex and dynamic environments. The connectionist approach, “inspired by the structure and function of the human brain emerged as a powerful alternative paradigm” (Nilsson 1998, p.169). Connectionist models (CM), based on artificial neural networks (ANN), offer a fundamentally different approach to AI. They draw inspiration from the structure and function of the human brain, employing a network of interconnected nodes, or neurons, that communicate with each other through weighted connections. This architecture allows for distributed representation, where knowledge is encoded across the network rather than in specific symbols or rules.

Connectionist models are not simply a technical advancement; they also carry significant philosophical implications that challenge our traditional understanding of intelligence and cognition. One of the most significant contributions of connectionist models is their emphasis on embodied intelligence. This perspective argues that

intelligence is not merely a matter of manipulating symbols but rather arises from the interaction of an organism with its environment. It challenges the Cartesian dualism that separates mind and body, suggesting that consciousness is an emergent property of embodied systems. Connectionist models are intrinsically embodied, meaning that they are designed to interact with and learn from the physical world. It views intelligence not as a disembodied process but rather emerges from the interplay between an organism's body, its environment, and its actions within that environment. Connectionist models move beyond the limitations of formal logic, embracing the complexities of real-world situations and the dynamic interplay between an organism and its environment.

Connectionist models employ distributed representation, where knowledge is encoded across the network, with each neuron contributing to the overall representation. Neural networks models various cognitive processes, such as perception, memory, and language processing. This contrasts with the symbolic approach, where knowledge is typically encoded in specific symbols or rules. Meaning in CM is context-dependent, emerging from the activation patterns of neurons within the network.

Connectionist models have significant implications for our philosophical understanding of AI, challenging our traditional views of consciousness, cognition, and the nature of intelligence itself. CM offer a new perspective on cognition, suggesting that it is not solely a matter of symbolic manipulation but rather involves complex

computations carried out by interconnected neural networks. CM provides a plausible cognitive architecture for understanding how the brain processes information, learns, and adapts. It further highlights the importance of the body in shaping cognition, emphasizing the role of sensory input, motor control, and the interaction with the environment.

The Probabilistic Model of AI: Embracing Uncertainty

Probabilistic models, grounded in probability theory and statistics, offer a fundamentally different approach to representing and reasoning about intelligence. They embrace uncertainty, representing information in terms of probabilities, enabling systems to reason under uncertainty and make decisions based on the likelihood of different outcomes. It aims to remedy the halting problem through, making decisions based on the available evidence and accounting for the possibility of uncertainty. Probabilistic models can learn from data and adjust their beliefs based on new evidence, making them more adaptable to changing environments. They are more robust and can generalize better to new situations, making them more flexible and capable of handling complex real-world problems.

Probabilistic Epistemology: a process epistemological framework

Probabilistic models embrace uncertainty, acknowledging that our knowledge of the world is often incomplete and imperfect. This philosophical stance “challenges the traditional view of knowledge as a collection of certain truths, recognizing that our understanding of

the world is always provisional and subject to revision"(Russell, Stuart and Peter Norvig, 2021,p.271). It rejects the notion of absolute certainty, arguing that all knowledge is ultimately probabilistic and based on degrees of belief. Probabilistic models emphasize the importance of evidence in forming beliefs, allowing for the updating of beliefs based on new information and experience.

Probabilistic models emphasize the role of subjective belief in reasoning and decision-making. They recognize that individuals may hold different beliefs based on their prior experience, background knowledge, and personal perspectives. This subjective element is reflected in the use of prior probabilities, which represent an individual's initial beliefs before any new evidence is considered. Probabilistic models provide a rigorous framework for making decisions in the face of uncertainty. They allow for the quantification of risk, enabling informed decision-making based on the likelihood of different outcomes.

Probabilist epistemology models a process perspective of knowledge claims, suggesting that knowledge is not a collection of certain truths but rather a set of beliefs that are constantly evolving based on evidence and experience. This perspective challenges the traditional view of knowledge as a static and objective entity which is in itself unchanging, emphasizing the dynamic and subjective nature of belief. Knowledge is seen as a dynamic process, constantly evolving in response to new information and experiences.

Probabilist Logic: Bayesian Inference

Probabilistic models, particularly Bayesian inference, provide a framework for understanding and formalizing the concept of rationality in the face of uncertainty. They suggest that “rational decision-making involves updating beliefs based on new evidence, taking into account prior beliefs and the likelihood of different outcomes”(Ghahramani, 2012,p.1). Bayesian inference is a powerful framework for updating our beliefs about the world based on new evidence. It's a cornerstone of probabilistic reasoning, offering a systematic and mathematically rigorous approach to understanding how evidence can influence our knowledge. This approach, named after the 18th-century English mathematician Thomas Bayes, provides a logical structure for combining prior knowledge with new data to arrive at updated beliefs. The origins of Bayesian inference can be traced back to Thomas Bayes posthumously published paper, "*An Essay towards solving a Problem in the Doctrine of Chances*"(1763). This groundbreaking work introduced a theorem, now known as Bayes' Theorem, which provides a mathematical formula for updating beliefs based on new evidence.

Bayes's Theorem: This fundamental theorem provides a way to calculate the probability of a hypothesis (H) given some observed evidence (E), denoted as P(H|E). It states:

$$P(H|E) = [P(E|H) * P(H)] / P(E)$$

Where:

P(H|E): The posterior probability of the hypothesis given the evidence.

P(E|H): The likelihood of observing the evidence given the hypothesis is true.

P(H): The prior probability of the hypothesis (our initial belief).

P(E): The probability of observing the evidence often called the marginal likelihood(Triola, 2016,p.12)

Bayesian inference provides a coherent logical structure for updating beliefs based on evidence. It involves three key components:

1. **Prior Belief:** The prior probability reflects our initial belief about the hypothesis before observing any evidence. This belief can be based on previous experience, background knowledge, or any other relevant information. The prior belief is inherently subjective and can vary across individuals, reflecting their individual experiences and biases.
2. **Likelihood Function:**The likelihood function quantifies the probability of observing the evidence given that the hypothesis is true. It represents how well the observed evidence supports the hypothesis. The likelihood function plays a crucial role in Bayesian inference, as it provides a way to assess the strength of the evidence in favor of the hypothesis.
3. **Posterior Belief:**The posterior probability represents our updated belief about the hypothesis after considering the evidence. It reflects the combined influence of our prior belief and the new evidence. The posterior belief is a balance between the prior belief and the likelihood of the evidence, with the relative weights determined by the strength of the evidence.

Probabilistic models offer a new perspective on intelligence, suggesting that it is not merely a matter of “manipulating symbols

according to predefined rules but rather involves the ability to reason under uncertainty, learn from experience, and adapt to changing environments"(Griffiths 2011, p.3). This perspective challenges the traditional view of intelligence as a deterministic and rule-based process, highlighting the importance of probabilistic reasoning, adaptive learning, and the ability to handle incomplete information. Probabilistic models hold significant promise for the future of AI, offering a powerful framework for creating intelligent systems that can reason under uncertainty, learn from experience, and make informed decisions in complex and dynamic environments.

The Reinforcement Learning Paradigm: A World of Action, Reward, and Adaptation

Reinforcement learning (RL), drawing inspiration from behavioral psychology and cognitive science, offers a fundamentally different approach to AI. It focuses on agents that learn through interactions with their environment, receiving feedback in the form of rewards or punishments. It is a "machine learning method in which the agent learns local behavior by doing actions and seeing the result of actions. For every good deed the agent receives a positive feedback and for every bad deed the agent receives a negative feedback" (Makkar 2024 p. 120). This feedback guides the agent's actions, allowing it to learn optimal behaviors over time.

> where the agent's actions influence the state of the environment, and the environment provides feedback in the form of rewards. It emphasizes learning through trial and error, with the agent exploring different actions and adapting its behavior based on the received rewards. RL systems are capable of learning

> and adapting to changing environments, continually improving their performance based on feedback received from their interactions. Optimal behaviors emerge from the agent's interactions with the environment, rather than being explicitly programmed, reflecting the dynamic and adaptive nature of intelligence. (Sutton and Barto 2018, p.16)

RL emphasizes the active role of the agent in shaping its own intelligence. Unlike rule-based systems that passively follow predefined instructions, RL agents are active participants in their environment, taking actions and learning from their consequences. This perspective resonates with the philosophical notion of embodied cognition, which argues that intelligence is not merely a matter of manipulating symbols but rather arises from the interaction of an organism with its environment. RL aligns with the idea that intelligence is grounded in the physical body and its interactions with the environment. It emphasizes the active nature of intelligence, with agents taking actions and shaping their own experiences.

Reward plays a crucial role in RL, serving as the primary feedback mechanism that guides the agent's learning. This raises philosophical questions about the nature of value and how rewards are assigned. Rewards reflect the subjective value judgments of the designer or the environment, shaping the agent's learning and goal orientation. One of the central challenges in RL is aligning the agent's values with the values of its creators or the wider society, ensuring that the agent's actions are beneficial and aligned with human interests. The design of reward functions is crucial for guiding the

agent's learning in a desired direction, requiring careful consideration of the desired outcomes and potential unintended consequences.

RL highlights the dynamic interplay between agent and environment, with the agent's actions influencing the state of the environment, and the environment providing feedback in the form of rewards. This feedback loop drives the learning process, shaping the agent's behavior over time. The agent's learning is situated in its environment, with the agent's actions and experiences shaping its understanding of the world. The interplay between agent and environment can be viewed as a dynamic system, constantly adapting and evolving based on their interactions. The agent's optimal behavior emerges from its interactions with the environment, rather than being explicitly programmed.

RL emphasizes the role of experience in learning, with agents acquiring knowledge through repeated interactions with their environment. This perspective contrasts with the rule-based approach, where knowledge is explicitly encoded by human experts. RL has the potential to create more intelligent and adaptable AI systems that can handle complex and dynamic environments. It offers a valuable tool for understanding how humans learn and adapt, providing insights into the cognitive processes underlying intelligent behavior. It is already being used in a wide range of applications, including game playing, robotics, autonomous driving, and healthcare.

Ethical Considerations for AI Systems in the Quest for General AI

The pursuit of Artificial General Intelligence (AGI) represents one of the most ambitious goals in the field of artificial intelligence. “While the potential benefits of AGI are immense ranging from solving complex global issues to enhancing human capabilities, the ethical considerations surrounding its development and deployment are equally critical”(Bishop 2006,p.5). the notion of artificial general intelligence is grounded in the idea of AI systems developing to the extent of intelligence combines the cognitive skills of humans, performing tasks with better efficiency and accuracy than humans; independently and without strict supervision. However, the ethical risks associated with AGI must be addressed or mitigated to ensure balance and remain within the core objectives of AI systems; being primarily the technological advancement of the world through simplification of tasks.

Safety and Control

The safety of AGI systems is paramount. Unlike narrow AI, which is designed for specific tasks, AGI must navigate complex, multifaceted environments. This requires robust design to minimize unintended consequences. Researchers must “implement rigorous testing protocols and adopt strategies for fail-safe mechanisms ensuring that AGI operates within established ethical boundaries”(Braband and Schäbe 2020, pp.15). The alignment problem ensuring that AGI goals and behaviors align with human values is a significant challenge. As AGI systems grow more capable, the risk of misalignment increases. Approaches such as value

learning, inverse reinforcement learning, and cooperative inverse reinforcement learning are being explored to tackle this issue. Continuous human oversight and intervention mechanisms may also be necessary to ensure AGI operates as intended.

Responsibility and Accountability

As AGI systems become more autonomous, the question of responsibility emerges. Who is accountable for the actions taken by an AGI? This complicates legal and ethical frameworks, necessitating the development of new regulatory standards. It raises “critical questions regarding liability in the event of harm caused by AGI, emphasizing the need for clear guidelines on the responsibilities of developers, operators, and users” (Good fellow et al. 2016, p.188).Developers must adhere to ethical design principles at every stage of AGI development. This includes transparency in algorithms, fairness in data usage, and consideration of the societal impacts of AGI deployment. Ethical oversight committees can assist in evaluating design processes to ensure adherence to these principles.

Bias and Fairness

AGI inherits biases present in training data, which can lead to discriminatory outcomes. As AGI systems make decisions affecting individuals and communities, it is essential to implement bias detection and mitigation strategies. Techniques such as fairness-aware machine learning and diverse dataset curation should be

employed to promote equitable outcomes. To enhance fairness, stakeholder engagement must be inclusive, considering diverse perspectives and experiences in the development of AGI systems. Engaging marginalized communities in the design process can yield insights into potential biases and inform more equitable AGI outcomes.

Privacy and Surveillance

AGI systems often rely on vast amounts of data, raising privacy concerns. Safeguarding personal information is critical to maintaining trust between users and AGI systems. Data anonymization, secure storage practices, and strict adherence to data protection regulations are necessary to mitigate privacy risks. The deployment of AGI in surveillance systems poses ethical dilemmas. While it can enhance security, unchecked surveillance can infringe on civil liberties. An ethical framework must be established to regulate the use of AGI in surveillance, balancing security needs with individual rights.

Societal Impacts and Global Considerations

AGI has the potential to displace jobs, leading to economic inequality and social unrest. Policymakers must proactively address these implications by developing safety nets and re-skilling programs for affected workers. Ensuring a just transition will be vital in maintaining social stability. A denial of posterity in the framework of AGI could be contradictory to her objectives of development(Otto 2020). AGI development transcends national boundaries, requiring international cooperation on regulations and ethical standards.

Establishing a global framework for AGI governance can help align the interests of various stakeholders, reducing the risk of a race to the bottom in ethical standards.

Also, as AGI systems become more advanced, they may pose existential risks to humanity. The possibility of AGI systems making decisions independently raises concerns about their alignment with human survival. Researchers must prioritize the identification and mitigation of these risks through safety protocols and strategic planning. A fundamental question in the quest for AGI is how these systems will coexist with humanity. The rights of the individual in the society must be protected at all cost and should be at the fore of AGI considerations (Obioha 2021; Obioha 2014). Rather than viewing AGI as a replacement for human abilities, the focus should be on co-evolution and collaboration. Envisioning a future where AGI enhances human capabilities can foster a more positive outlook toward its development. Adhering to ethical principles throughout the development process can create a framework wherein AGI not only advances technological progress but also aligns with the broader goals of humanity.

Conclusion

The halting problem is a fundamental limit on the power of computation. It demonstrates that even with the most powerful computers, there will always be problems that are inherently impossible to solve algorithmically. This understanding compels us to rethink our approach to artificial intelligence (AI) and to explore

alternative frameworks for creating truly intelligent machines. The discovery of undecidability does not diminish the power of algorithms, but it compels us to acknowledge their inherent limitations. Its renaissance encourages us to explore new computational models, to develop novel techniques for dealing with uncertainty, and to seek inspiration from the multifaceted nature of human intelligence. Consequently embracing and developing ethical standards for the control of artificial intelligent systems is crucial for the growth and sustainability of AI. It is pertinent that humanity continues to be the centre piece of every development drive, hence AI policies ought to be developed with these values in focus.

References

Akman, Varol and Patrick Blackburn.(2000) “Editorial: Alan Turing and Artificial Intelligence” *Journal of Logic, Language, and Information.* 9(1). pp.391–395

Bishop, Christopher M. (2006) *Pattern Recognition and Machine Learning.* Springer.

Braband, Jens & Schäbe, Hendrik. (2020). “On Safety Assessment of Artificial Intelligence” *Dependability* 20(4).

Brodkorb,Laurel. (2019) *The Entscheidungsproblem and Alan Turing*. Georgia College And State University.

Ghahramani, Zoubin.(2012)*Probabilistic Modeling, Machine Learning And The Information Revolution.* MIT CSAIL.

Godel, Kurt.(1992)*On Formally Undecidable Propositions Of Principia Mathematica And Related systems.* TRANS Meltzer, B. Dover Publications.

Godfrey-Smith, Peter,(2014) "John Dewey's Experience And Nature" *TOPIO* 33(1), pp. 285-291

Goodfellow, Ian.(2016). *Deep Learning*. MIT Press.

Griffiths, Thomas. (2011). *Connecting Human And Machine Learning Via Probilistic Models Of Cognition.* University Of Califonia Press.

Harnish, Robert M.(2002). *Minds, Brains, Computers*. Blackwell.

Heidegger, Martin.(1962) *Being and Time*. Translated by John Macquarrie and Edward Robinson, Harper & Row.

Lawhead, William.(2015) *The Voyage Of Discovery: A Historical Introduction To Philosophy*. Cengage Learning.

Makkar, Priyanka. (2024) "Reinforcement Learning: A Comprehensive Overview". *International Journal Of Innovative Research In Computer Science And Technology(IJIRCST)* 12(2), pp. 119-125.

Merleau-Ponty, Maurice. (2012). *Phenomenology of Perception*. Translated by Colin Smith, Routledge,

Nilsson, Nils J.(1998) Artificial Intelligence: A New Synthesis. Morgan Kaufmann.

Obioha, Precious Uwaezuoke (2014). "A Communitarian Understanding of the Human Person as a Philosophical Basis for Human Development" *The Journal of Pan African Studies*, 6(8), pp 247-267

Obioha, Precious Uwaezuoke. (2021) "An Afro-Communal Ethic for Good Governance" *Acta Universitatis Danubius Administratio* 13(1), pp. 20-38

Otto, Dennis (2020). "Ethics of Posterity for Environmental Development of the Niger Delta" *GNOSI: An Interdisciplinary Journal of Human Theory and Praxis.* 3(3), pp 83-96

Russell, Stuart, and Peter Norvig.(2021)*Artificial Intelligence: A Modern Approach*. Pearson Education.

Sutton, Richard S., and Andrew G. Barto.(2018)*Reinforcement Learning: An Introduction*. MIT Press.

Triola, Mario.(2016)*Bayes' Theorem*. University of Washington Seattle Press.

Turing, Alan M. (1950)."Computing Machinery and Intelligence." *Mind,* 59(236), pp. 433-460.