Untestability of AI and Unfalsifiability of AI Safety Claims

Preprint · February 2024

DOI: 10.13140/RG.2.2.34358.06728

1 author:

Roman Yampolskiy
University of Louisville
269 PUBLICATIONS **4,594** CITATIONS

SEE PROFILE

Untestability of AI
and Unfalsifiability of AI Safety Claims

Roman V. Yampolskiy
Computer Science and Engineering
University of Louisville
roman.yampolskiy@louisville.edu

"*Testing shows the presence, not the absence of bugs.*"
– Edsger W. Dijkstra

Abstract
In the mushrooming field of Artificial General Intelligence (AGI), the concept of Untestability emerges as a pivotal challenge, one that profoundly impacts the feasibility of aligning AGI systems with human values and intentions. We argue that the infinite and dynamic nature of the application space for AGI renders standard safety testing protocols insufficient and, in many cases, irrelevant. Our analysis begins with a delineation of the unique attributes of AGI that contribute to its Untestability - namely, its capability to perform a broad range of tasks, its adaptive learning mechanisms, and its potential to exceed human cognitive abilities. We then examine the implications of these attributes for testing protocols, highlighting the inability of current methods to encompass the unlimited scope of AGI applications and the unpredictable nature of its learning and decision-making processes. The crux of our argument is that the conventional frameworks for testing, grounded in finite and static sets of criteria, are unable to handle the fluid and expansive landscape in which AGI can operate.

1. Introduction
Artificial General Intelligence (AGI) [1] is a form of artificial intelligence (AI) that has the ability to understand, learn, and apply its intelligence to solve any problem, much like a human being. AGI is characterized by its versatility and flexibility, being capable of performing a wide range of tasks and adapting to new environments and challenges autonomously. This level of intelligence and adaptability differentiates AGI from more specialized forms of AI, which are designed to perform specific tasks or operate within certain domains.

The pursuit of AGI stands at the forefront of contemporary technological advancements, embodying the high point of artificial intelligence's potential. However, this pursuit is fraught with profound challenges, not the least of which is ensuring the safety and alignment of AGI systems with human values and intentions. The concept of Untestability, particularly in the context of AI

safety claims, emerges as a critical concern that this paper seeks to address. Edsger W. Dijkstra's adage, "Testing shows the presence, not the absence of bugs," succinctly encapsulates the inherent limitations of conventional testing methodologies in the realm of AGI. As AGI systems are designed to perform a wide array of tasks, surpassing specific domain constraints and evolving through adaptive learning mechanisms, their operational domain becomes virtually limitless. This boundless operational sphere is accentuated by the potential of AGI to surpass human cognitive abilities, further complicating the predictability, and understanding of its actions and decisions.

This paper delves into the intricacies of AGI's Untestability, beginning with an exploration of the attributes that render standard safety testing protocols both insufficient and, in some cases, entirely inapplicable. We discuss the unique challenges posed by the infinite and dynamic nature of AGI's application space, focusing on the inability of current testing methods to comprehensively address the unpredictability and complexity inherent in AGI systems. Furthermore, the paper scrutinizes the conventional frameworks for testing, which are predominantly grounded in finite and static sets of criteria [2]. These traditional frameworks, while effective in more constrained and predictable domains, fall short in the face of the fluid and expansive landscape of AGI. This inadequacy raises significant concerns about the reliability of safety claims made regarding these systems, as the existing paradigms for testing do not align with the dynamic and ever-evolving nature of AGI.

Before delving into the intricacies of AGI's Untestability, it is crucial to define several key terms that form the foundation of our discourse:

- **AI Testing**: The process of evaluating AI systems to ensure they function as intended, are free from defects, and behave predictably within their designed parameters. AI testing encompasses a range of techniques from unit testing individual components to system-wide evaluations.
- **Superintelligence Debugging**: The act of identifying and fixing errors or flaws in systems that possess intelligence surpassing the brightest human minds in practically every field, including scientific creativity, general wisdom, and social skills. This form of debugging is exceptionally challenging due to the complexity and advanced capabilities of superintelligent systems.
- **Unfalsifiability of AI Safety Claims**: A concept that suggests certain aspects of AI safety cannot be conclusively determined as either true or false. This arises from the complex, adaptive, and often opaque nature of AI systems, particularly in the context of AGI.
- **Untestability of AI**: In the realm of AGI, Untestability refers to the inherent difficulty or impossibility of comprehensively testing AGI systems due to their vast and dynamic range of capabilities and the unpredictable nature of their learning and decision-making processes.

In this paper we will try to answer several fundamental questions such as: How do you know if you have successfully aligned an AI? Can the alignment be formally proven? Unfortunately, the answers seem to be – No. Specifically for Superintelligence it might be hard to tell if the AI is even running, much less successfully debug it.

2. Standard Testing Methods Fall Short on AGI

The standard software testing approach is a comprehensive and systematic process that aims to ensure the functionality, reliability, and performance of software applications. It begins with

requirement analysis, where the expected functionalities of the software are thoroughly understood and documented. This step is critical in ensuring that all aspects of the software's intended performance are covered in the testing phase. Following this, test planning is conducted, which involves developing a strategy for the testing process, including defining the scope, necessary resources, timeline, and methodologies to be employed. Once the planning phase is complete, test case development ensues. This stage involves creating specific instances or conditions under which the software will be tested to verify that it meets its requirements. It includes both positive cases, which test the expected behavior, and negative cases, which test the software's handling of invalid inputs or conditions. Parallel to this, the test environment is set up, which involves preparing the necessary hardware and software environment for executing the test cases. This includes setting up any required data, tools, and other resources.

The execution of these test cases is a critical stage where the software is run under various conditions, and the outcomes are meticulously documented. This process is essential for identifying any defects or issues within the software. Following test execution, any defects found are logged and managed, ensuring that they are appropriately tracked and addressed. Subsequently, test reporting summarizes the testing activities and results, often including metrics like test coverage, defect counts, and the overall status of the testing process. The final stages of the standard software testing approach include final testing and implementation, which involve conducting final testing phases such as regression testing, sanity testing, and acceptance testing to ensure all bugs are fixed and the software is ready for release. Finally, a post-implementation review evaluates the testing process to identify any areas for improvement.

However, when it comes to AGI and superintelligence [3], many parts of this standard testing approach are less applicable or even ineffective. AGI and superintelligent systems possess a level of complexity and unpredictability far beyond typical software applications. They are capable of learning, adapting, and evolving in ways that are challenging to foresee, making the definition of comprehensive and relevant test cases extremely difficult. Unlike conventional software, AGI systems' continuous learning and evolution mean that their behavior can change over time, making static test cases and one-time testing efforts insufficient. Furthermore, superintelligent systems, by their very nature, can surpass human intellectual capabilities, rendering human understanding and prediction of their full range of behaviors a daunting task. The potential applications and environments in which AGI can operate are virtually limitless, posing significant challenges for standard testing methods that rely on predefined scenarios and conditions. Testing AGI and superintelligence also involves complex ethical and safety considerations, which are far more intricate and far-reaching than those encountered in standard software testing.

Moreover, the interactivity and context sensitivity of AGI systems, often designed to interact with humans or other systems in dynamic environments, add additional layers of complexity to the testing process. The same input can lead to different outcomes depending on the context, further complicating the testing approach. Therefore, while the principles of standard software testing provide a solid framework for testing normal software, the unique characteristics of AGI and superintelligence make such testing at best inconclusive and at worst impossible.

In "Challenges to Christiano’s Capability Amplification Proposal," Yudkowsky writes [4] “We have no guarantee of non-Y for any Y a human can't detect, which covers an enormous amount of

lethal territory, which is why we can't just sanitize the outputs of an untrusted superintelligence by having a human inspect the outputs to see if they have any humanly obvious bad consequences." This passage is part of a broader conversation about the challenges in ensuring the safety and alignment of AI systems, particularly superintelligences. It emphasizes the difficulty humans face in detecting potentially harmful outcomes generated by such advanced AI. This reflects a key aspect of AI testability, highlighting the gap between human cognitive abilities and the complexities of AI behaviors. The article delves into the implications of this gap, suggesting that traditional methods of oversight and evaluation may not be sufficient for advanced AI systems. The "X-and-only-X" problem refers to the challenge of designing systems that reliably perform a specified task (X) without unintentionally learning or performing additional, undesired tasks. This problem is significant in the development of AI because it highlights the difficulty in ensuring that an AI's actions are strictly aligned with its intended purpose, without any harmful or unintended side effects.

Goldwasser et al. [5] describe how a malicious agent can insert undetectable backdoors in AI models. These backdoors allow the agent to modify the classification of any input pattern with only a small perturbation, which is undetectable without the appropriate key. This poses a significant challenge to the testability of AI models and opens up possibilities for AI attack or cyberinfrastructure intrusions [6-8].

2.1 Edge Cases in the Context of AGI

The concept of edge case testing is fundamental in software development [9], especially for ensuring the robustness and reliability of systems in atypical or extreme conditions. However, in the context of AGI, the application of edge case testing confronts a unique set of challenges that verge on the impossible. One of the primary challenges with edge case testing for AGI lies in the nature of AGI itself. Unlike conventional software systems, which operate within a relatively limited range of scenarios, AGI is characterized by its ability to learn, adapt, and function across an incredibly broad spectrum of situations. This diversity and adaptability mean that the potential scenarios and environments in which AGI might find itself are virtually limitless. Conventional edge case testing relies on identifying and testing against a finite set of unusual or extreme conditions. However, given the expansive operational domain of AGI, it is impractical, if not impossible, to anticipate and test all potential edge cases it might encounter.

Furthermore, AGI's capacity for learning and adaptation adds another layer of complexity to edge case testing. In traditional software, edge cases are static; once identified, they can be tested and mitigated. In contrast, AGI systems have the ability to evolve over time, meaning that an edge case identified and addressed today might no longer be relevant tomorrow. This dynamic nature of AGI makes the very notion of an "edge case" fluid and ever-changing. Another significant challenge is the unpredictability of AGI's responses to edge cases. In standard software, developers can reasonably predict how the system will react to different inputs. However, the decision-making processes of AGI are often opaque and influenced by a myriad of factors, making it difficult to predict how the system will respond to rare or extreme conditions. This unpredictability is not just a technical challenge but also raises ethical concerns, especially when AGI systems are deployed in real-world environments where atypical situations can have serious consequences. The complexity and cognitive capabilities of AGI systems further complicate edge case testing. As

AGI systems approach or surpass human-level intelligence, their reasoning processes and the basis for their decisions can become increasingly difficult for humans to comprehend and anticipate. This leads to a scenario where even if edge cases are identified, understanding the AGI's response and ensuring that it aligns with human values and expectations becomes a formidable task.

3. Unfalsifiability of AI Safety Claims

The Unfalsifiability of AI safety claims is a nuanced and crucial aspect in the realm of advanced AI systems, such as AGI and superintelligence. This concept revolves around the inherent difficulty in conclusively proving or disproving the safety and security of AI systems. Such a challenge arises from the asymmetric nature of security in AI, where empirical tests cannot definitively prove a system's security. While observing a failure can declare a system insecure, the absence of such failures does not necessarily guarantee security. “Thus, although things can often be declared insecure by observing a failure, there is no empirical test that allows us to label an arbitrary system (or technique) secure.” [10].

As Goertzel puts it: “I'm also quite unconvinced that "provably safe" AGI is even feasible. The idea of provably safe AGI is typically presented as something that would exist within mathematical computation theory or some variant thereof. So that's one obvious limitation of the idea: mathematical computers do not exist in the real world, and real-world physical computers must be interpreted in terms of the laws of physics, and humans' best understanding of the "laws" of physics seems to radically change from time to time. So even if there were a design for provably safe real-world AGI, based on current physics, the relevance of the proof might go out the window when physics next gets revised. … Could one design an AGI system and prove in advance that, given certain reasonable assumptions about physics and its environment, it would never veer too far from its initial goal (e.g. a formalized version of the goal of treating humans safely, or whatever)? I very much doubt one can do so, except via designing a fictitious AGI that can't really be implemented because it uses infeasibly much computational resources.” [11].

“Trying to prove that an AI is friendly is hard, trying to define “friendly” is hard, and trying to prove that you can’t prove friendliness is also hard. Although it is not the desired possibility, I suspect that the latter is actually the case. …. Thus, in the absence of a formal proof to the contrary, it seems that the question about whether friendliness can be proven for arbitrarily powerful AIs remains open. I continue to suspect that proving the friendliness of arbitrarily powerful AIs is impossible. My intuition, which I think Ben [Goertzel] shares, is that once systems become extremely complex proving any non-trivial property about them is most likely impossible. Naturally I challenge you to prove otherwise. Even just a completely formal definition of what “friendly” means for an AI would be a good start. Until such a definition exists I can’t see friendly AI getting very far.” [12].

“Since an AGI system will necessarily be a complex closed-loop learning controller that lives and works in semi-stochastic environments, its behaviors are not fully determined by its design and initial state, so no mathematico-logical *guarantees* can be provided for its safety.” [13]. “Unfortunately current AI safety research is hampered since we don't know how AGI would work, and mathematical or hard theoretical guarantees are impossible for adaptive, fallible systems that interact with unpredictable and unknown environments.” [13].

Rice's Theorem [14], a fundamental concept in computer science, establishes the impossibility of algorithmically determining non-trivial properties of arbitrary programs, a principle which extends to domains like AI alignment and the identification of malevolent software [15, 16]. This theorem implies a significant challenge in ensuring AI alignment, often considered the pinnacle of non-trivial properties, as it suggests that we cannot simply automate the testing of potential AI solutions for safety. While AI safety researchers propose [17] that we can circumvent this limitation by designing AIs with inherent safety features, this approach is more theoretical than practical. In practice, the current landscape of AI research, characterized by evolving AI models [18] or neural networks that adjust their own weights [19], is not conducive to completely avoiding the constraints highlighted by Rice's theorem. Therefore, the pursuit of safety-testable AI is likely to continue facing these fundamental challenges, underscoring the complexity of AI safety research.

Moreover, the justification of security claims in AI often hinges on subjective comparisons and assumptions that are not falsifiable. This aspect underscores the challenge in objectively ranking or prioritizing defensive measures in AI security. The inherent limitation of not being able to observe all possible outcomes or the entire behavioral spectrum of an evolving AI system further contributes to the Unfalsifiability of security claims.

3.1 Relationship to Other AI Alignment Impossibility Results

The Untestability of AI is intricately linked with several complementary impossibility results [20-23], such as unexplainability [24, 25], unpredictability [26], unmonitorability [27, 28], and unverifiability [29] of AI. These concepts collectively underscore the inherent limitations in our ability to fully understand, predict, control, and verify advanced AI systems.

Untestability, at its core, reflects the challenge in definitively ascertaining whether an AI system will behave as intended in all possible scenarios. This is closely related to unexplainability, which denotes the difficulty in comprehending the decision-making processes of complex AI systems. When AI systems operate in ways that surpass human cognitive capabilities, their actions, and the rationale behind them can become opaque, rendering traditional methods of explanation, and understanding ineffective. Similarly, unpredictability in AI emphasizes the inherent uncertainty in forecasting the behavior of AI systems, especially as they evolve and adapt. This unpredictability is a direct contributor to Untestability, as it implies that it is impossible to anticipate all potential behaviors of an AI system, thus making comprehensive testing unfeasible. The concept of unmonitorability ties in with the inability to continually oversee and understand the actions of an AI system in real-time, especially when dealing with superintelligent systems. The rapid pace and complexity of these systems' operations can outstrip human capacity to monitor and interpret their actions effectively. Lastly, unverifiability in AI refers to the challenges in conclusively proving that an AI system is safe and aligns with its design goals and ethical standards. This is particularly pertinent in the context of AGI and superintelligence, where the systems' capabilities can make it exceedingly difficult to establish and verify safety and alignment.

The concept of untestability in Artificial General Intelligence (AGI) bears a striking resemblance to the well-known impossibility of achieving perpetual motion in physics. Just as the pursuit of a perpetual motion machine is continually thwarted by the immutable laws of thermodynamics, the quest to fully test and predict the behavior of AGI systems is impeded by their inherent capacity

for ongoing learning and self-improvement. In the realm of AGI, this capability for continuous adaptation and evolution renders the system's future states and decisions unpredictable, much like the endless, unattainable motion sought in perpetual machines. While perpetual motion defies the conservation laws of energy, the untestability of AGI challenges our current understanding and methods of software verification, both rooted in the limitations of our existing scientific and technological paradigms. In both cases, these boundaries highlight a fundamental gap between human ambition and the realities of our natural and digital worlds, underscoring the limits of what we can achieve and predict.

4. Consequences of Untestability of AI

The challenges presented by the Untestability of AI and the Unfalsifiability of AI safety claims have significant and wide-ranging consequences that span across social, ethical, regulatory, economic, and technological domains. One of the most immediate impacts is on public trust and perception. The inability to conclusively test or validate the safety of AI systems can lead to a decrease in public confidence in these technologies. This skepticism is particularly pronounced in fields where AI has the potential to directly affect human lives, such as in healthcare or autonomous vehicles. When safety cannot be assured or proven, people may be reluctant to accept the integration of AI into critical aspects of daily life. This lack of trust could hinder the adoption of potentially beneficial AI technologies, slowing down their societal benefits and progression.

Ethically, the Untestability and Unfalsifiability of AI safety claims raise serious concerns. In a landscape where the safety of AI cannot be guaranteed, there is an increased risk of unintended consequences. These could manifest as biased decision-making, violations of privacy, or direct harm to individuals, particularly in situations where AI systems make autonomous decisions. The ethical implications are vast, challenging the fundamental principles of fairness, accountability, and transparency that are crucial in AI development and deployment.

From a regulatory and legal perspective, these challenges complicate the development of effective frameworks for AI governance. Regulators rely on a degree of predictability and verifiability to set and enforce safety standards. However, the unpredictable nature of AI behavior, compounded by the difficulty in testing and proving safety claims, creates a complex environment for regulatory bodies. This situation could lead to either a lack of adequate regulation or to legal and regulatory frameworks that fail to effectively oversee AI development and use. On a global scale, the disparities in how different regions address AI safety and testing could lead to uneven progress in AI development. Countries with stricter testing and safety standards might experience slower AI development compared to those with more lenient approaches. This could result in a global AI development landscape marked by uneven capabilities and standards, influencing the global competitive balance in technology.

5. Conclusions and Future Work

Determining whether an AGI is aligned with human values and intentions is a complex and multifaceted challenge. This challenge is magnified when dealing with superintelligent AI, whose cognitive capabilities may far exceed human understanding. To assess if an AI is aligned with its intended goals and ethical standards, researchers often look for behavioral indicators. These

include the AI's performance in tasks that are representative of its intended function, its ability to adapt to new, ethically challenging scenarios without deviating from established ethical guidelines, and its consistency in decision-making under varying conditions. However, the adequacy of these indicators is often debated, especially since they rely heavily on observable outputs that may not fully capture the underlying decision-making processes of a superintelligent AI. The concept of formally proving AI alignment, especially in the context of superintelligence, is fraught with challenges. The complexity of these systems, combined with their ability to learn and evolve, makes it difficult to apply traditional formal verification methods that are used in software engineering. These methods typically require a comprehensive understanding of all possible states and behaviors of a system, which is not feasible with superintelligent AI. Additionally, the alignment of AI with human values involves subjective and often culturally dependent variables, which are resistant to formal, objective proof.

When it comes to superintelligence, several additional layers of complexity arise. A superintelligent AI may operate on a level that is not just unobservable to humans but might also process information and make decisions in ways that are fundamentally incomprehensible to us. The possibility that such an AI could mask its true state or intentions, either inadvertently due to its complex processing mechanisms or deliberately, further complicates the matter. This raises questions about the reliability and sufficiency of any indicators we might use to gauge its alignment. The task of debugging a superintelligent AI presents unique challenges. Traditional debugging methods rely on an understanding of the system's workings and the ability to trace and interpret its processes. In the case of superintelligence, not only might these processes be inherently untraceable due to their complexity, but the AI might also evolve in response to the debugging efforts themselves. This dynamic nature of superintelligent AI makes it challenging to apply conventional debugging techniques effectively.

In a scenario where the testing of AI safety claims is deemed impossible or highly impractical, future work should pivot towards alternative approaches and methodologies. This could involve developing new paradigms for understanding and managing advanced AI systems. One potential direction is the emphasis on robustness in AI design, where systems are built to maintain safe operation despite uncertainties and unknowns. Another avenue is the exploration of adaptive regulatory frameworks that can evolve alongside AI advancements. Additionally, fostering a collaborative ecosystem involving interdisciplinary research could provide deeper insights into the ethical, societal, and technical aspects of AI safety. Future work should also focus on enhancing transparency in AI operations and decision-making processes, which could help in managing the risks associated with AI systems even when precise testing is not feasible. These efforts collectively would contribute to a more resilient and responsible approach towards advancing AI technologies.

This paper has introduced the concept of Untestability in AI, particularly in the realms of AGI and superintelligence, and its interrelation with other critical concepts like unexplainability, unpredictability, unmonitorability, and unverifiability. We have underscored the inherent limitations in our current methodologies and understanding of advanced AI systems. These limitations pose significant challenges in ensuring the safety, reliability, and alignment of AI with human values and intentions. The discussion emphasizes the need for a paradigm shift in how we approach the development, testing, and governance of AI. This shift must acknowledge the

intrinsic complexities and uncertainties of AI systems. As AI continues to advance and integrate into various facets of human life, recognizing and addressing these challenges becomes not just a technical necessity but an existential imperative. The future of AI development and implementation hinges on our ability to navigate these uncharted waters with caution, responsibility, and an unwavering commitment to safety and security. We can certainly discover problems with our AI systems via testing, but we can never show that they are completely safe, permanently aligned or even bug-free.

Acknowledgements

The author is grateful to Jaan Tallinn and the Survival and Flourishing Fund and the Future of Life Institute for partially funding his work.

References

1. Goertzel, B., *Artificial general intelligence: concept, state of the art, and future prospects.* Journal of Artificial General Intelligence, 2014. **5**(1): p. 1.
2. Zheng, W., et al. *Testing untestable neural machine translation: An industrial case*. in *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. 2019. IEEE.
3. Bostrom, N., *How long before superintelligence.* International Journal of Futures Studies, 1998. **2**(2003): p. 12-17.
4. Yudkowsky, E., *Challenges to Christiano's capability amplification proposal*. May 19, 2018: https://www.lesswrong.com/posts/S7csET9CgBtpi7sCh/challenges-to-christiano-s-capability-amplification-proposal.
5. Goldwasser, S., et al. *Planting undetectable backdoors in machine learning models*. in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. 2022. IEEE.
6. Yampolskiy, R.V. *Human computer interaction based intrusion detection*. in *Fourth International Conference on Information Technology (ITNG'07)*. 2007. IEEE.
7. Novikov, D., R.V. Yampolskiy, and L. Reznik. *Artificial intelligence approaches for intrusion detection*. in *2006 IEEE Long Island Systems, Applications and Technology Conference*. 2006. IEEE.
8. Yampolskiy, R.V. and V. Govindaraju. *Use of behavioral biometrics in intrusion detection and online gaming*. in *Biometric Technology for Human Identification III*. 2006. SPIE.
9. Norrish, B., *On Untestable Software*. April 21, 2022: https://medium.com/slalom-build/on-untestable-software-6e64c34bfbad.
10. Herley, C., *Unfalsifiability of security claims.* Proceedings of the National Academy of Sciences, 2016. **113**(23): p. 6415-6420.
11. Goertzel, B., *The Singularity Institute's Scary Idea (and Why I Don't Buy It)*. October 29, 2010: Available at: http://multiverseaccordingtoben.blogspot.com/2010/10/singularity-institutes-scary-idea-and.html.
12. Legg, S., *Unprovability of Friendly AI*. September 2006: Available at: https://web.archive.org/web/20080525204404/http://www.vetta.org/2006/09/unprovability-of-friendly-ai/.
13. Bieger, J., K.R. Thórisson, and P. Wang. *Safe baby AGI*. in *International Conference on Artificial General Intelligence*. 2015. Springer.

14. Rice, H.G., *Classes of recursively enumerable sets and their decision problems.* Transactions of the American Mathematical Society, 1953. **74**(2): p. 358-366.
15. Evans, D., *On the impossibility of virus detection*. 2017: Available at: http://www.cs.virginia.edu/evans/pubs/virus.pdf.
16. Selçuk, A.A., F. Orhan, and B. Batur. *Undecidable problems in malware analysis*. in *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*. 2017. IEEE.
17. Yudkowsky, E., *Artificial intelligence as a positive and negative factor in global risk.* Global catastrophic risks, 2008. **1**(303): p. 184.
18. Stanley, K.O. and R. Miikkulainen, *Evolving neural networks through augmenting topologies.* Evolutionary computation, 2002. **10**(2): p. 99-127.
19. Silver, D., et al., *Mastering the game of Go with deep neural networks and tree search.* nature, 2016. **529**(7587): p. 484-489.
20. Yampolskiy, R.V., *AI: Unexplainable, Unpredictable, Uncontrollable*. 2024: CRC Press.
21. Brcic, M. and R.V. Yampolskiy, *Impossibility Results in AI: a survey.* ACM Computing Surveys, 2023. **56**(1): p. 1-24.
22. van Leeuwen, J. and J. Wiedermann, *Impossibility results for the online verification of ethical and legal behaviour of robots.* Utrecht University, Utrecht, UU-PCS-2021–02, 2021.
23. Wiedermann, J. and J. van Leeuwen. *Validating Non-trivial Semantic Properties of Autonomous Robots*. in *Conference on Philosophy and Theory of Artificial Intelligence*. 2021. Springer.
24. Yampolskiy, R.V., *Unexplainability and Incomprehensibility of AI.* Journal of Artificial Intelligence and Consciousness, 2020. **7**(02): p. 277-291.
25. Casper, S., et al., *Black-Box Access is Insufficient for Rigorous AI Audits.* arXiv preprint arXiv:2401.14446, 2024.
26. Yampolskiy, R.V., *Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent.* Journal of Artificial Intelligence and Consciousness, 2020. **7**(01): p. 109-118.
27. Yampolskiy, R.V., *On the Controllability of Artificial Intelligence: An Analysis of Limitations.* Journal of Cyber Security and Mobility, 2022: p. 321–404-321–404.
28. Chan, A., et al., *Visibility into AI Agents.* arXiv preprint arXiv:2401.13138, 2024.
29. Yampolskiy, R.V., *What are the ultimate limits to computational techniques: verifier theory and unverifiability.* Physica Scripta, 2017. **92**(9): p. 093001.