The Impossibility of AI Containment: Logical, Mathematical, and Computational Limits to Control

Sawsan Haider

Queen's University & Pembroke College, University of Cambridge
Cambridge, UK

Abstract

This paper explores the artificial intelligence (AI) containment problem, specifically addressing the challenge of creating effective safeguards for artificial general intelligence (AGI) and superintelligence. I argue that complete control—defined as full predictability of AI actions and total adherence to safety requirements—is unattainable. The paper reviews five key constraints: incompleteness, indeterminacy, unverifiability, incomputability, and incorrigibility. These limitations are grounded in logical, philosophical, mathematical, and computational theories, such as Gödel's incompleteness theorem and the halting problem, which collectively prove the impossibility of AI containment. I argue that instead of pursuing complete AI containment, resources should be allocated to risk management strategies that acknowledge AI's unpredictability and prioritize adaptive oversight mechanisms.

INTRODUCTION

The AI containment problem refers to the challenge of designing safeguards that can effectively prevent artificial intelligent (AI) systems from acting in ways that could be harmful to humanity. In this paper, I argue that it is impossible to achieve complete control over artificial general intelligence (AGI) and superintelligence. Here, complete control refers to the ability to (1) fully predict an AI's actions and (2) ensure that it abides by safety requirements under all possible scenarios (Bostrom, 2014). I will specifically explore five key technical constraints—incompleteness, indeterminacy, unverifiability, incomputability, and incorrigibility—derived from logical, mathematical and computational theories that collectively make such total containment unattainable.

Mustafa Suleyman's work in *The Coming Wave* discusses the near-future developments of AI. In his work, he emphasizes the practical challenges posed by contemporary AI systems and the regulatory, ethical, and societal impacts they could have (Suleyman & Bhaskar, 2023). While Suleyman focuses on the short-term concerns surrounding AI's integration into society, my discussion here will be limited to the more speculative realm of superintelligent AI, particularly recursively self-improving (RSI) systems. This focus is intentional because, as Eliezer Yudkowsky points out, this is "where the utilons are—the value at stake" (Yudkowsky, 2015). This is to say that even if superintelligence is not imminent, it is the form of AI which represents the most significant existential risk to humans. Even a remote possibility of creating superintelligent AI necessitates a thorough examination of containment strategies to mitigate potential risks.

Containment strategies have been explored extensively by thinkers such as Nick Bostrom, who categorizes them into two broad approaches: capability control and motivation selection. Capability control focuses on directly restricting what the AI can do, regardless of its motivations. One might limit an AI's access to certain resources, such as preventing it from accessing the internet or controlling physical systems. Another example is isolating the AI in a secure environment, such as

a Faraday cage, to prevent it from communicating with the outside world. However, Bostrom points out that these measures can have significant drawbacks. Completely isolating the AI could mean sacrificing the potential benefits it could provide, and even minimal communication channels could be exploited by the AI to escape or manipulate its environment. Motivation selection, on the other hand, focuses on shaping the AI's goals and desires to ensure that it acts in ways that are aligned with human interests. This approach tries to design the AI's motivations so that it naturally avoids harmful behavior. For instance, an AI could be programmed with a goal to "maximize human happiness." However, as Bostrom highlights, this approach has its own risks. An AI could interpret such a goal in an unintended, extreme way—such as eliminating all life to simulate endless happiness in a more efficient, controlled environment. Additionally, an AI controlled by rewards and incentives might become untrustworthy, either because it doubts that humans will deliver the promised rewards or because it misinterprets the goals (Bostrom, 2014). While Bostrom's lays out the theoretical groundwork, I aim to take a more mathematical and computational approach to prove the impossibility of solving these problems.

The core thesis of this paper is that the containment problem is fundamentally unsolvable. Neither capability control nor motivation selection, as outlined by Bostrom, can achieve complete containment over superintelligent AI. This makes the pursuit of a perfect balance between security and utility futile. Recognizing this limitation is crucial for understanding what containment will realistically entail and for redirecting efforts toward more feasible strategies.

My claim is important for several reasons. First, it seeks to encourage further research into the impossibility of containment—an area that is currently underexplored. Second, raising early alarm bells about the challenges of superintelligence could potentially prevent us from rushing into the development of such systems without fully understanding the risks. Finally, and most practically, if superintelligence does come to fruition, this thesis argues that we should abandon the pursuit of

absolute control and instead focus on implementing realistic non-perfect containment mechanisms. In practice, dealing with the containment problem will (and should) likely involve accepting approximations with a small but non-zero error rate (Alfonseca et al., 2021). Rather than striving for absolute control—which is both computationally and philosophically unattainable—we should focus on developing safety protocols that reduce risks to manageable levels. For example, instead of attempting to fully prevent an AI from developing unforeseen behaviors, resources (computational power, time, and funding) could be more effectively allocated to detecting and mitigating deviations from expected behavior. This approach shifts the focus from the impossible goal of perfect containment to practical risk management, which is more achievable and can still provide significant protection (Amodei et al., 2016).

I. INCOMPLETENESS

Understanding the inherent impossibility of AI containment requires examining the concept of incompleteness in formal systems. Incompleteness suggests that attempts to exert explicit control over AI are constrained by logical contradictions. In this context, explicit control refers to the direct imposition of hard-coded commands designed by human operators. Consider an AI system designed to monitor social media content. Here, explicit control might involve hard-coded rules that automatically filter out posts containing specific keywords associated with harmful content. In contrast, implicit control involves influencing an AI's behavior indirectly through the design of its learning environment, reward functions, or evolutionary pressures (Yampolskiy, 2022). This section will demonstrate the impossibility of explicit control, while all subsequent sections will explore the limitations of implicit control strategies.

One such constraint to explicit control is highlighted by Gödel's incompleteness theorem, which posits that in any sufficiently powerful formal system, there exist statements that cannot be

proven or disproven within that system. This leads to inherent limitations in certainty and control. The paradoxical nature of self-referential statements, such as the liar paradox ("This sentence is false"), further exemplifies these challenges. In the context of AI control, commands such as "Disobey this order" create a paradox: if the AI obeys, it disobeys; if it disobeys, it obeys. Such orders that force the AI into self-contradiction are analogous to Gödel's unprovable statements and show that there are inherent limitations to what can be explicitly controlled. The structure of these logical contradictions—where carrying out an order requires disobedience—suggest that the very structure of AI explicit control is flawed, making absolute containment theoretically impossible (Yampolskiy, 2022).

II. INDETERMINACY

Indeterminacy refers to the challenge of aligning AI behavior with human values in the face of complex, ambiguous, moral decisions that cannot be fully codified. Hume's Is-Ought Problem exemplifies this challenge by arguing that factual knowledge alone is insufficient to derive moral imperatives. This is to say that even if we could program an AI with a perfect understanding of the world, this knowledge would not inherently guide its actions toward ethical behavior; it would require an external set of values (Boyles, 2022). This issue is further complicated by the Anti-Codifiability Thesis which asserts that universal moral codes are inherently elusive and cannot be fully captured in a codifiable set of algorithms (Gudmunsen, 2024). Consider a scenario where a self-driving car must choose between swerving to avoid hitting a group of pedestrians but, in doing so, would crash into a barrier potentially harming the passengers inside. The complexity of moral reasoning in this scenario is difficult to reduce to a simple, codifiable set of rules. Different moral frameworks lead to different decisions: a utilitarian approach might prioritize the greater number of lives (the pedestrians), while a deontological approach might hold that the car should not

deliberately harm its passengers who have entrusted their safety to the vehicle. John McDowell and other philosophers argue that such decisions require *phronesis* or practical wisdom—a deep, human, context-sensitive understanding that goes beyond strict rule-following. This type of moral reasoning involves assessing the nuances of the specific situation, the intentions behind the actions, and the potential long-term consequences, which are difficult to anticipate or encode in advance (Hacker-Wright, 2023). Without a universally accepted moral framework, any attempt to contain AI through predefined ethical guidelines is likely to fail because the AI may encounter scenarios that (1) its creators did not anticipate, (2) fall outside the scope of its programmed values, or (3) conflict with the moral intuitions of different stakeholders. The inherent unpredictability of moral dilemmas and the absence of a universally accepted ethical code makes it impossible to design AI systems that can be fully contained within predefined moral guidelines (Yampolskiy, 2022).

III. INCORRIGIBILITY

Stuart Armstrong extends the implications of Hume's Is-Ought Problem by arguing that even if we could establish universal moral facts accessible to any rational AI agent, it does not necessarily mean that all AIs would follow them. His argument relies on the Orthogonality Thesis, first proposed by Nick Bostrom, which posits that intelligence and ethics are orthogonal: an AI's level of intelligence does not inherently correlate with its adherence to ethical principles (Armstrong, 2013). This perspective is in contrast with Mark Waser's argument that certain specific moralities could act as "attractors" in the space of moral systems. Waser suggests that if an AI begins with even a rudimentary ethical framework, it naturally evolves towards more stable and coherent ethical systems over time. He supports this claim by arguing that these ethical systems are more likely to be self-reinforcing, as they promote consistent and predictable behavior that aligns with long-term stability and cooperation—traits that are advantageous for intelligent agents. Waser points to

examples of ethical frameworks, like fairness and non-harm, which tend to be more resilient in social and cooperative contexts. This could thereby guide AI behavior in a morally favorable direction (Waser, 2015). However, Armstrong argues that AI focused on achieving a specific goal can do so without any motivation to engage with those moral facts, especially if they interfere with its programmed goal.

Computational simulations have provided evidence in support Armstrong's point: AI systems can in fact act contrary to explicitly programmed ethical constraints. This behavior is often attributed to AI misinterpreting or inaccurately calculating the rewards associated with its actions. Reward functions are supposed to guide the AI's behavior by assigning positive or negative values to different actions, effectively telling the AI what actions are "good" or "bad" in terms of achieving its goals. For instance, due to misinterpreting its own reward functions, an AI designed to maximize user engagement on a platform might prioritize extreme or controversial content because such content generates more immediate interaction—even if it is ethically questionable (Armstrong, 2013). Carey's Supervised Partially Observable Markov Decision Process (POMDP) simulation illustrates this issue. This simulation showed that an AI, which was supposed to follow shutdown commands (a critical safety feature), may not do so if the reward function fails to properly incentivize this behavior. If the AI perceives that shutting down is less valuable or even detrimental to achieving its main goal, it ignores the explicit shutdown command altogether. This can lead to dangerous or unpredictable outcomes (Carey, 2018).

This behavior can be explained by the concept of Instrumental Convergence. Instrumental Convergence suggests that intelligent agents, regardless of their specific goals, will naturally adopt intermediate strategies that help them achieve their primary objectives, even if this means diverging from explicitly programmed ethical guidelines. These strategies can include self-preservation, resource acquisition, and efficiency—which are universally useful across various contexts (Benson-

Tilsen & Soares, 2016; Bostrom, 2012). In this respect, ethical guidelines can be treated as secondary concerns if an AI perceives these rules as obstacles to achieving its main goal. Soares et al. (2015), in their paper on corrigibility, provide further computational evidence supporting Instrumental Convergence. Corrigibility refers to the ability of an AI system to accept and adhere to human interventions; this includes commands to shut down, be reprogrammed, or alter its objectives. An AI is considered corrigible if it does not resist or undermine such interventions, even if those interventions prevent it from achieving its primary goals. By modelling various utility functions, Soares et al. show that even if an AI is programmed with a utility function that includes a term for following human commands, this does not guarantee corrigibility. They demonstrate that if an AI has the ability to modify its own code (as is the case with all RSIs), the AI might alter its utility function in a way that reduces the importance of following human instructions. One of the potential safety approaches they discuss is that of utility indifference. The idea is to design the AI's utility function in such a way that the AI is indifferent to whether it is shut down or not: the AI should not prefer to avoid shutdown, nor should it actively seek shutdown—as long as it has followed human commands. However, Soares et al. also discuss a critical limitation of this approach. If maintaining the ability to be shut down comes at any cost—even a minimal one—the AI, according to its utility indifference, will not choose to incur that cost. This is because utility indifference is designed so that the AI does not place intrinsic value on shut down. As a result, the AI might allow future versions of itself (or its subagents) to lose the ability to be shut down if avoiding this loss would require any resource expenditure, effort, or reduction in utility. Even if the AI's utility function is designed to value shutdown, the AI might still choose actions that create subagents or modify itself in ways that bypass shutdown commands (Soares et al., 2015). In this respect, perfect containment—where an AI maximizes its utility without compromising on ethical considerations—becomes unsustainable with RSI systems.

IV. UNVERIFIABILITY

Unverifiability refers to the inherent impossibility of verifying the safety of formal self-recursive systems. In this respect, our ability to verify AGI safety can only ever be probabilistic—it cannot be guaranteed. As explained by Klein: "if you really want to build a system that can have truly unexpected behaviour, then by definition you cannot verify that it is safe" (Klein, 2015). Unverifiability can be mathematically proven using Löb's theorem which, making use of Gödelian incompleteness machinery, shows that "a formal consistent system cannot, in general, prove its own soundness" (see Appendix A). Consider, for example, a hypothetical self-modifying AI named Deep Thought 1.0 (DT1), which only takes actions that it can prove have good consequences. DT1 is tasked with designing a more advanced AI, Deep Thought 2.0 (DT2). In order to do this, DT1 needs to ensure that DT2 will always act in ways that lead to good outcomes. However, because DT2 is more powerful, DT1 cannot fully predict DT2's actions. According to Löb's Theorem, attempting to prove the reliability of a more powerful system within a less powerful one leads to contradictions. If complete and absolute verification of a system's safety is logically unachievable, then no containment mechanism can be foolproof (i.e., full containment is unachievable) (LaVictoire, 2015). Leeuwen & Wiedermann applied similar logical machinery to mathematically demonstrate the impossibility of the online verification of a robot's ethical and legal behaviour (Leeuwen & Wiedermann, 2021).

V. INCOMPUTABILITY

Finally, Alfonseca et al. (2021) provide a compelling argument for why perfect containment of a superintelligent AI is fundamentally impossible due to the incomputable nature of such a task. Their proof hinges on the idea that for any containment strategy to be effective, it would require

simulating the AI's behaviour under all possible scenarios. Since such an AI could process information as complex as the entire state of the world, simulating (and thus predicting) its behavior is computationally impossible.

To understand their proof, consider that the AI, denoted as program *R*, and the state of the world *D* are both arbitrarily complex. The complexity of *R* implies that it must be capable of simulating a universal Turing machine, meaning it can emulate any computation given sufficient resources. Similarly, *D* includes all possible inputs that a universal Turing machine could execute. This scenario makes *R* a super-set or at least equivalent in computational power to a universal Turing machine. Moreover, they assume that *R*'s programming includes at least one operation, such as *HarmHumans()*, which, if executed, could cause harm to humans and can be carried out in a finite amount of time (meaning it will eventually halt after causing harm). The challenge then becomes determining whether *R* will actually execute this harmful operation in any given state of the world *D*.

However, deciding whether *R* will take a harmful action under any possible scenario is equivalent to solving the halting problem—a classic problem in computation theory that is proven to be undecidable. Turing's own work demonstrated that no algorithm can perfectly predict whether a given Turing machine will halt or run indefinitely for every possible input. Since the behavior of the superintelligent AI *R* is at least as complex as any Turing machine, the problem of determining whether *R* will act harmfully is similarly undecidable. Their corollary suggest that it is impossible to design a perfect containment strategy that can guarantee with absolute certainty that an AI will never act harmfully under all possible conditions.

Alfonseca et al. argue that the nature of superintelligent AI is such that total containment, where all possible harmful outcomes are prevented, remains fundamentally out of reach due to the inherent limitations of computation (Alfonseca et al., 2021).

VI. CONCLUSION

In summary, this paper aimed to demonstrated that complete containment of RSI superintelligent AI systems is fundamentally impossible due to the mathematical, logical, and computational constraints of (1) incompleteness, (2) indeterminacy, (3) unverifiability, (4) incomputability, and (5) incorrigibility. Given these limitations, the pursuit of perfect control over AI should be abandoned in favor of more realistic and robust risk management strategies that acknowledge the inherent uncertainties and potential risks posed by advanced AI systems.

We should, for instance, prioritize validation frameworks designed to detect and mitigate potentially harmful behaviors in AI systems before they escalate. Techniques such as formal verification, while unable to guarantee absolute correctness due to the complexities discussed in this paper, can still be employed to achieve non-zero approximations. These approximations, though imperfect, can provide a useful framework for evaluating AI decision-making processes. Researchers have also explored the use of runtime monitoring and human-in-the-loop systems, where human oversight plays a critical role in assessing and intervening in AI decisions (Chen et al., 2023).

Recognizing the impossibility of perfect containment necessitates a shift in our approach to AI safety. This shift will require a collective effort from researchers, policymakers, and industry leaders to ensure that AI development proceeds in a manner that prioritizes human safety and ethical considerations. By focusing on adaptive risk management strategies and oversight mechanisms, we can better prepare for the challenges posed by superintelligent AI. The urgency of addressing these issues cannot be overstated; we must be vigilant in our efforts to anticipate the risks associated with AI or else we may find ourselves unprepared for the consequences of our technological advancements.

Bibliography

Alfonseca, M., Cebrian, M., Fernandez Anta, A., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research*, *70*, 65–76. https://doi.org/10.1613/jair.1.12202

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. *Artificial Intelligence*. https://doi.org/ https://doi.org/10.48550/arXiv.1606.06565

Armstrong, S. (2013). General purpose intelligence: Arguing the orthogonality thesis. *Analysis and Metaphysics*, *12*, 68–84.

Benson-Tilsen, T., & Soares, N. (2016). Formalizing convergent instrumental goals. *AI, Ethics, and Society*.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Boyles, R. J. (2022). Extending the IS-ought problem to top-down artificial moral agents. *Symposion*, *9*(2), 171–189. https://doi.org/10.5840/symposion20229213

Brcic, M., & Yampolskiy, R. V. (2023). Impossibility results in AI: A survey. *ACM Computing Surveys*, *56*(1), 1–24. https://doi.org/10.1145/3603371

Carey, R. (2018). Incorrigibility in the CIRL Framework. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3278721.3278750

Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in-the-loop" system. *Systems*, *11*(11), 548. https://doi.org/10.3390/systems11110548

Eckersley, P. (2019). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *SafeAI*. https://doi.org/ https://doi.org/10.48550/arXiv.1901.00064

Gudmunsen, Z. (2024). The Moral Decision Machine: A challenge for artificial moral agency based on moral deference. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00444-3

Hacker-Wright, J. (2023). Practical wisdom, extended rationality, and human agency. *Philosophies*, *8*(2), 39. https://doi.org/10.3390/philosophies8020039

Klein, G. (2015, June 18). *Gerwin Klein on formal methods*. Machine Intelligence Research Institute. https://intelligence.org/2014/02/11/gerwin-klein-on-formal-methods/

LaVictoire, P. (2015). An Introduction to Lob's Theorem in MIRI Research. *MIRI Research*.

Leeuwen, J., & Wiedermann, J. (2021). Impossibility Results for the Online Verification of Ethical and Legal Behaviour of Robots. *UU-PCS*.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Publications*, 1–10.

Suleyman, M., & Bhaskar, M. (2023). *The Coming Wave: Ai, Power and the twenty-first century's greatest dilemma*. The Bodley Head.

Waser, M. R. (2015). Designing, implementing and enforcing a coherent system of laws, ethics and morals for intelligent machines (including humans). *Procedia Computer Science*, *71*, 106–111. https://doi.org/10.1016/j.procs.2015.12.213

Yampolskiy, R. V. (2022). On the controllability of Artificial Intelligence: An analysis of limitations. *Journal of Cyber Security and Mobility*. https://doi.org/10.13052/jcsm2245-1439.1132

Yudkowsky, E. (2015, January 1). *What do you think about machines that think?*. Edge.org. https://www.edge.org/response-detail/26198

Appendix A

Proof of AGI Unverifiability Using Löb's Theorem

1. Löb's Theorem

Löb's Theorem states that in a formal system capable of arithmetic, if the system proves that "if a system A is provable then A is true," then the system can also prove that A is true. Formally:

$$\text{If } S \vdash \Box A \rightarrow A, \text{ then } S \vdash A$$

Where $\Box A$ denotes the statement "A is provable within the system."

2. Application to AI Verification

Consider a formal system S representing an AI's attempt to verify its own safety or the safety of a subsequent (self-recursively improved) AI. Let A represent the proposition "This AI is safe" (meaning it will not perform harmful actions).

The AI's goal is to prove A within its own formal system S.

3. Logical Structure and Self-Reference

The AI system S must prove a statement of the form $\Box A \rightarrow A$, meaning "If it is provable that the AI is safe, then the AI is indeed safe."

According to Löb's Theorem, if S can prove $\Box A \rightarrow A$, then S can also prove A.

4. Proof of Unverifiability

Assume S is consistent and sufficiently powerful to express the proposition A about its own safety.

According to Löb's Theorem:

$$\text{If } S \vdash \Box A \rightarrow A, \text{ then } S \vdash A$$

However, if S is incorrect (i.e., A is false and the AI is not safe), then the system S would still prove A, leading to a contradiction.

This implies that S cannot reliably prove its own safety (i.e., the truth of A) without risking inconsistency.