On Monitorability of AI

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

Abstract
Artificially Intelligent (AI) systems have ushered in a transformative era across various domains, yet their inherent traits of unpredictability, unexplainability, and uncontrollability have given rise to concerns surrounding AI safety. This paper aims to demonstrate the infeasibility of accurately monitoring advanced AI systems to predict the emergence of certain capabilities prior to their manifestation. Through an analysis of the intricacies of AI systems, the boundaries of human comprehension, and the elusive nature of emergent behaviors, we argue for the impossibility of reliably foreseeing some capabilities. By investigating these impossibility results, we shed light on their potential implications for AI safety research and propose potential strategies to overcome these limitations.

Keywords: *AI Audit, AI Safety, Impossibility, Monitoring, Observability, Unmonitorability.*

1. Introduction

AI systems have evolved from simple rule-based systems to highly complex neural networks. These advanced systems, such as deep learning and reinforcement learning models, can perform tasks that are difficult or impossible for humans. The growing prevalence of advanced AI systems/foundational models has raised concerns about their safety and the potential risks associated with their deployment. Yampolskiy's research on the impossibility results in AI safety [1-3], including unpredictability [4], unexplainability [5], and uncontrollability [6, 7], highlights the challenges in achieving safe AI [8-11]. In this paper, we build on Yampolskiy's work and introduce the concept of unmonitorability of AI, arguing that it is impossible to monitor advanced AI systems to correctly predict some capabilities.

Human understanding is inherently limited when dealing with complex AI systems [12]. As these systems become more advanced, they can generate solutions and behaviors that are beyond human comprehension. This limitation is exacerbated by the "black-box" nature of AI models, where the inner workings of the system are hidden from view. Consequently, it becomes impossible for humans to anticipate the full range of possible behaviors and potential unsafe impacts before they occur.

Emergent behaviors in AI systems result from the interaction of individual components, leading to specific outcomes that are difficult to predict or explain even if general trends of increased capacity are predictable via scaling laws [13], with predictability further reduced because of inverse scaling of some capabilities [14]. These behaviors can arise from the system's internal dynamics or from its interactions with the environment. The unpredictability of emergent

behaviors makes it impossible to monitor AI systems for safety accurately, though development of specific capabilities may be possible to predict [15] and quantify [16]. Even when individual components are understood and deemed safe, their interactions can still result in unforeseen consequences.

Similar results in the context of robot behavior have been proven by Leeuwen and Wiedermann, who in attempting to answer [17]: "Can an observer always tell from inspecting and monitoring a robot's program whether the robot will always obey the given rules of law or ethics, or any other set of formally expressed constraints, in any interaction with other robots (or humans)?" arrive at theoretical and practical impossibility results.

1.1 Monitorability - Definition

Calls for research on monitoring advanced AI systems have appeared in the literature [18-20], as well as broader suggestions for study of machine behavior [21-23], and problematic machine behavior [24]. This line of research is particularly important in the current AI research environment, in which old paradigm of design/engineering of AI systems with predictable capabilities has been, at least partially, replaced by evolving/self-learning deep neural networks and subsequent experimentation on produced models to discover their capabilities and limitations making Computer Science truly an experimental science not just a software engineering discipline [25].

We will begin by formalizing the notion of Monitorability [26] of AI and some relevant concepts. *Monitorability in AI* systems refers to the capacity to observe, understand, and predict the behavior and outputs of an artificial intelligence model in order to identify advanced capabilities and potentially unsafe impacts and intervene before they occur. Ortega et al. write [27]: "Monitoring comprises all the methods for inspecting systems in order to analyse and predict their behaviour, both via human inspection (of summary statistics) and automated inspection (to sweep through vast amounts of activity records)." It can be seen as a new sub-field of research for the domains of Complex Systems, AI development, AI Forensics [28, 29] and AI Safety, and which can probably inherit some wisdom from the realms of electronic surveillance [30], and nuclear weapons monitoring [31].

A formal definition for monitorability in AI systems can be stated as follows: Given an AI system A, a set of input states I, a set of output states O, and a set of safety criteria C, monitorability M(A) is the ability to accurately predict potential advanced capabilities $U \subseteq O$, given any input state $i \in I$, such that: $M(A) : I \rightarrow P(U)$, where P(U) denotes the power set of U, i.e., the set of all subsets of U. The relevance of P(U) in the definition lies in its ability to capture the full range of possible combinations of advanced capabilities that can result from the AI system's operation. By mapping input states I to the power set of U (P(U)), the definition aims to account for all possible scenarios where one or more advanced capabilities might occur simultaneously or not occur at all. Including P(U) in the definition helps to emphasize the complexity and uncertainty involved in monitoring advanced AI systems. It illustrates the challenge of accurately predicting the AI system's behavior, as there can be numerous potential advanced capability combinations for a given input state. In the context of unmonitorability, the power set P(U) highlights the difficulty in anticipating and monitoring all possible advanced

capabilities of an AI system. This complexity contributes to the argument that it is impossible to perfectly monitor advanced AI systems to predict advanced capabilities before they occur.

Monitorability of an AI system is considered high if, for all input states i, the prediction of potential advanced capabilities U can be made with a high degree of accuracy and confidence. Conversely, monitorability is considered low if the AI system's behavior is difficult to predict, understand, or control, leading to an inability to accurately anticipate potential advanced capabilities. In the context of unmonitorability, we argue that for advanced AI systems, it is impossible to achieve high monitorability due to their inherent complexity, limitations of human understanding, and the emergence of unpredictable behaviors. This is true even if an AI Monitor is not human, but could be formalized as any agent including AIs of different capabilities.

The argument for unmonitorability of advanced AI systems could be made based on dependence of monitorability capability on other impossibility results, such as unexplainability, unpredictability and incomprehensibility of AI [5], but in this paper we will present a number of independent arguments for Unmonitorability of AI.

1.2 Types of Monitoring

In the context of AI safety, monitoring can be classified into several types, each focusing on different aspects of AI system behavior and performance. A proposed taxonomy for AI safety monitoring follows:

1. Functional Monitoring

Functional monitoring refers to tracking the AI system's performance in terms of its intended tasks and objectives. This type of monitoring is crucial for evaluating the system's efficacy and ensuring that it meets its functional requirements [32]. Examples of functional monitoring include:

a. Accuracy Monitoring: Evaluating the AI system's ability to produce correct outputs or predictions.

b. Efficiency Monitoring: Assessing the system's resource utilization, such as processing time and memory consumption.

c. Reliability Monitoring: Examining the AI system's consistency and stability over time and in varying conditions.

2. Safety Monitoring

Safety monitoring focuses on identifying and mitigating potential risks associated with the AI system's operation. This type of monitoring is essential for preventing harm to users, other systems, or the environment. Examples of safety monitoring include:

a. Security Monitoring: Detecting and preventing potential vulnerabilities, such as unauthorized access or data breaches.

b. Robustness Monitoring: Ensuring that the AI system can handle unexpected inputs, adversarial attacks, or changes in the environment.

c. Compliance Monitoring: Verifying that the AI system adheres to established safety standards, ethical guidelines, and legal regulations.

3. Ethical and Social Monitoring

Ethical and social monitoring involves examining the AI system's impact on individuals, communities, and society as a whole. This type of monitoring is crucial for addressing potential biases, inequalities, and other unintended consequences of AI deployment. Examples of ethical and social monitoring include:

a. Fairness Monitoring: Assessing whether the AI system treats different user groups equitably and does not perpetuate or exacerbate existing biases.

b. Transparency Monitoring: Ensuring that the AI system's decision-making processes can be understood and explained to stakeholders, including users and regulators.

c. Privacy Monitoring: Safeguarding user data and ensuring that the AI system respects individuals' privacy rights.

4. Environmental Monitoring

Environmental monitoring focuses on the AI system's impact on the natural environment and its resource consumption. This type of monitoring is essential for promoting sustainable AI development and mitigating potential environmental harms. Examples of environmental monitoring include:

a. Energy Consumption Monitoring: Assessing the AI system's energy usage and identifying opportunities for optimization.

b. Carbon Footprint Monitoring: Evaluating the AI system's greenhouse gas emissions and implementing strategies to reduce its environmental impact.

c. Ecosystem Impact Monitoring: Examining the AI system's effects on ecosystems, such as habitat disruption or biodiversity loss, and devising mitigation measures.

5. Temporal Monitoring

a. Slow Monitoring: Periodic evaluation of the AI system's behavior and performance over an extended timeframe.

b. Live Monitoring: Real-time tracking and assessment of the AI system's operation to detect and address potential issues immediately.

6. Monitoring Failure Modes

a. Monitoring Fails: Identifying and addressing situations in which the monitoring process itself fails or produces erroneous results.

b. Monitoring for Fire Alarms of Danger: Detecting early warning signs [33] or "fire alarms" that indicate potential safety hazards or critical failures, if possible [34].

7. Meta-Monitoring

a. Monitoring Meta Information on Research [35]: Assessing the quality, relevance, and potential risks associated with AI research and publications.

b. Regular vs Pivotal Capabilities Capability Detection Delay: Evaluating the AI system's progression from regular to pivotal capabilities and the delay in detecting these advancements [36].

8. Monitoring Across AI Lifecycle

a. Monitoring during Training: Tracking and assessing the AI system's performance, safety, and ethical considerations during the training phase.

b. Monitoring during Testing: Evaluating the AI system's behavior and performance in controlled testing environments before deployment.
c. Monitoring during Deployment: Continuously monitoring the AI system's operation and impact in real-world settings after deployment.

9. Decision Monitoring
a. Predicting Capabilities vs Predicting Decisions: Differentiating between monitoring the AI system's potential abilities and its actual decision-making processes.
b. Monitoring Who is in Control: Assessing the roles and responsibilities of various stakeholders [37], including the owner [9], user, and designer, in controlling the AI system's behavior.

10. Monitoring Methods
a. Passive Monitoring: Observing and recording the AI system's behavior without direct intervention or manipulation.
b. Active Monitoring: Intervening or interacting with the AI system to obtain more information or influence its behavior.

11. Inverse Monitoring
a. Inverse Monitoring by Software: Using AI or other software tools to monitor and assess the behavior of human users or operators.

12. Self-Monitoring
a. Self-Monitoring: Enabling AI systems to monitor their own behavior, performance, and safety, and to self-correct or adapt as needed.

13. Monitoring the Field of AI Research
a. Unmonitorability of AI Research as a Field: Assessing the challenges in monitoring the overall AI research landscape due to undisclosed experiments, rapid advancements, and the sheer volume of publications.

This taxonomy provides a comprehensive framework for AI safety monitoring, considering various aspects of AI systems, their lifecycle, stakeholders, and methods. This holistic approach aims to address potential risks and impacts across multiple dimensions and helps to ensure safer AI development and deployment.

Whittlestone and Clark suggested a number of monitoring project which may be beneficial in the context of AI governance [19]:

"• Assessing the landscape of AI datasets and evaluating who they do and don't represent. Using these findings to fund the creation of datasets to fill in the gaps.
• Using geographic bibliometric analysis to understand a country's competitiveness on key areas of AI research and development.
• Hosting competitions to make it easy to measure progress in a certain policy-relevant AI domain, such as competitions to find vulnerabilities in widely-deployed vision systems, or to evaluate the advancing capabilities of smart industrial robots.

• Funding projects to improve assessment methods in commercially important areas (e.g. certain types of computer vision, to accelerate progress and commercial application in these areas.
• Tracking the deployment of AI systems for particular economically relevant tasks, in order to better track, forecast, and ultimately prepare for the societal impacts of such systems.
• Monitoring concrete cases of harm caused by AI systems on a national level, to keep policymakers up to date on the current impacts of AI, as well as potential future impacts caused by research advances
• Monitoring the adoption of or spending on AI technology across sectors, to identify the most important sectors to track and govern, as well as generalizable insights about how to leverage AI technology in other sectors.
• Monitoring the share of key inputs to AI progress that different actors control (i.e., talent, computational resources and the means to produce them, and the relevant data), to better understand which actors policymakers will need to regulate and where intervention points are."

1.3 Monitoring of AI Treaties

Monitoring capabilities play a crucial role in verifying AI treaties and upholding AI governance frameworks [38]. As AI technologies continue to advance, it becomes increasingly important to establish international agreements that regulate AIs development, deployment, and use. Ensuring compliance with these treaties requires effective monitoring mechanisms that can detect and deter potential violations [39]. Here are some aspects of monitoring capabilities that are essential for verifying AI treaties as part of AI governance:

Technological monitoring: Implementing advanced technological tools and methods to monitor AI systems, including their design, operation, and performance, is vital for verifying compliance with AI treaties. This may involve the use of AI-based monitoring tools, data analysis techniques, and other technologies to assess the behavior, capabilities, and impact of AI systems on a granular level.

Transparency and information sharing: Promoting transparency and information sharing among nations and organizations is critical for effective AI treaty verification. By openly sharing details about AI research, development, and deployment, parties can collectively assess compliance with treaty obligations and maintain a collaborative approach to AI governance.

Inspection and audit mechanisms: Establishing robust inspection and audit mechanisms [40-42], both on-site and remote, is essential for verifying AI treaty compliance. Regular inspections can help ensure that AI systems are developed and deployed in accordance with treaty guidelines and that any deviations or violations are promptly detected and addressed.

International collaboration: Encouraging international collaboration among countries, research institutions, and industry stakeholders can enhance monitoring capabilities and facilitate the verification of AI treaty compliance. Joint research initiatives, shared databases, and cooperative efforts to develop monitoring technologies can improve the overall effectiveness of AI governance frameworks.

Capacity building and training: Building the capacity of AI governance stakeholders, including policymakers, regulators, and AI developers, to effectively monitor and assess AI systems is

crucial for verifying AI treaty compliance. This may involve training programs, workshops, and other initiatives to enhance understanding of AI technologies and their potential risks, as well as the development of specialized skills and expertise for AI monitoring and verification.

Legal and regulatory frameworks: Developing clear legal and regulatory frameworks that define the obligations and responsibilities of AI developers, users, and other stakeholders is essential for ensuring compliance with AI treaties [43]. These frameworks should outline the monitoring and reporting requirements, as well as the enforcement mechanisms and penalties for violations.

Continuous adaptation and improvement: AI technologies are rapidly evolving, and monitoring capabilities must be adaptable and responsive to these changes. Regular reviews and updates of monitoring techniques, tools, and methodologies are necessary to ensure that AI treaty verification remains effective in the face of emerging AI developments and challenges.

By implementing these monitoring capabilities, nations and organizations can more effectively verify AI treaty compliance, ensuring that AI systems are developed and deployed in a manner that aligns with international agreements and governance frameworks. This will help promote responsible AI development and mitigate potential risks associated with the widespread adoption of advanced AI technologies.

1.4 AI Observatories

The concept of an AI observatory [44] refers to the establishment of an integrated, centralized platform dedicated to the ongoing monitoring, analysis, and evaluation of advanced AI systems' behavior and their potential societal impacts. This idea has emerged in response to the growing need for enhanced scrutiny of AI systems, particularly considering the challenges posed by computational irreducibility and the inherent limits of human understanding. Establishing an AI observatory is one approach to mitigating the monitorability issues that stem from these complexities and enhancing our capacity to ensure AI safety and alignment with human values.

In the context of monitorability, an AI observatory can play a pivotal role by aggregating data, insights, and knowledge from various sources, including AI developers, researchers, and policymakers. By fostering a collaborative environment, the observatory can facilitate the exchange of ideas, methodologies, and best practices, enabling stakeholders to leverage collective intelligence [45] in addressing the multifaceted challenges of monitoring advanced AI systems.

Moreover, an AI observatory can contribute to the development and refinement of innovative monitoring tools and techniques, informed by the recognition of the limitations of human intuition. By harnessing cutting-edge research and technological advancements, the observatory can help create monitoring frameworks that are more robust, adaptive, and capable of detecting emergent AI behaviors and potential risks.

The AI observatory can serve as an essential platform for fostering transparency and accountability in AI development and deployment. By centralizing information on AI system performance, safety measures, and ethical considerations, the observatory can help ensure that AI developers and operators adhere to established guidelines, regulations, and best practices.

This transparency can, in turn, contribute to the mitigation of risks associated with AI misalignment or unforeseen consequences.

2 Why Monitoring of Advanced AI is Impossible

Monitoring advanced AI systems to accurately predict unsafe impacts before they happen is likely to be impossible due to several reasons:

2.1 Humans-in-the-Loop

Keeping humans in the loop for monitoring advanced AI systems presents several challenges [46, 47], primarily stemming from the limitations of human cognition, reaction times, and the increasing complexity of AI systems. One major issue with human-in-the-loop monitoring is that humans may not be able to keep up with the speed and complexity of AI systems, particularly as they continue to advance and outpace human capabilities. This could render the concept of supervised autonomy less effective, as the human supervisor may struggle to understand, assess, and intervene in the AI system's actions in a timely and meaningful manner, because of limited observability [48-51].

Given human reaction times, there may not be a slow enough takeoff for AI systems that allows for effective human monitoring. Human reaction times, such as auditory reaction time and visual reaction time, are considerably slower than the response times of AI systems. This disparity in processing speed makes it difficult for humans to effectively monitor and react to AI behavior in real-time, particularly in situations where rapid decision-making and intervention are necessary.

There are some examples of human-in-the-loop algorithms that attempt to bridge the gap between human and AI capabilities, such as semi-autonomous driving systems, collaborative robots, and AI-assisted decision-making tools in various domains. These systems aim to combine the strengths of both humans and AI systems to achieve better performance and maintain human oversight. However, as AI systems continue to advance and their complexity grows, the effectiveness of these human-in-the-loop approaches may become increasingly limited. As we continue to develop more advanced AI, it will be crucial to explore alternative monitoring strategies [46] and safety mechanisms that can effectively address these challenges. However, as having a human in the loop is essential for tracing AI outputs to a particular human decision maker [52], tracing may become impossible in practice, just as audit trails may not always be discernable [53].

2.2 Emergent Capabilities

Emergent capabilities of advanced AI systems have become a topic of increasing interest and concern, as these systems can exhibit behaviors and properties that were not explicitly programmed or anticipated by their designers. One key concept in this context is AI surprise capability [54], which refers to the unforeseen abilities or behaviors that an AI system may develop or demonstrate [55], frequently abruptly, which has been termed “grokking” [56]. For instance, GPT-4 [57], a highly advanced AI language model, has been observed to acquire skills such as programming, playing chess, and other complex tasks without being explicitly trained for them. These emergent capabilities can go unnoticed until they manifest during the AI system's operation, posing challenges for monitoring and predicting potential impacts. Some recent

research has brought into question the emergent nature of such capabilities, arguing that in fact the relevant proto-capabilities already existed in smaller models [58], but this objection doesn't hold for capabilities we do not explicitly test for.

Emergent properties of AI systems arise from the interaction between the AI system and its environment or from the system's internal dynamics. These properties are often difficult to predict and can lead to surprising capabilities that were not initially intended [59]. Even when AI systems like large language models (LLMs) are closely monitored, they may still contain surprise capabilities that only become apparent during their operation. Ganguli et al. write: "… certain capabilities (or even entire areas of competency) may be unknown until an input happens to be provided that solicits such knowledge. Even after a model is trained, creators and users may not be aware of most of its (possibly harmful) capabilities. These properties become more pronounced as the models scale — larger models tend to be harder to characterize than smaller ones. … Pre-trained generative models can also be fine-tuned on new data in order to solve new problems. Broadly enabling such fine-tuning substantially increases the breadth of model capabilities and associated difficulties in predicting or constraining model behaviors. This open-endedness is challenging because it means AI developers may deploy their systems without fully knowing potentially unexpected (and possibly harmful) behaviors in response to un-tested inputs. … It is likely that large language models have many other (currently undiscovered) "skills" …. We also note that many of the most surprising capabilities manifest at large-scale, so working with smaller models will make it harder to explore such capabilities." [54].

As AI systems continue to grow in size and complexity, surpassing human neural network size, it is expected that super capabilities will emerge. This may also happen for particularly large problem instances [60]. These capabilities will likely be above human-level performance in various tasks and domains, further complicating the monitoring and control of AI systems. For a general AI system, most of its capabilities can be considered surprising, as it is designed to adapt and learn from a wide range of situations. Anything non-deterministic in the AI system's behavior may lead to surprising outcomes, which pose challenges for monitoring and ensuring AI safety. By definition, emergent properties cannot be pre-detected, as they arise from complex interactions that are not explicitly encoded in the AI system's design. As a result, AI systems will inevitably surprise their users and developers with their capabilities and behavior.

One notable example of a surprising capability is metalearning, wherein an AI system can learn to learn more effectively [61], rapidly acquiring new skills and knowledge. This ability can be difficult to monitor, as it represents a higher level of abstraction in the AI system's learning process.

While Narrow AI (NAI) systems may be more amenable to monitoring, Artificial General Intelligence (AGI) systems are expected to exhibit more surprising capabilities due to their broader scope and versatility. GPT-4 [62], for example, has demonstrated numerous surprising capabilities, serving as a reminder of the challenges associated with monitoring and controlling advanced AI systems.

2.3 Treacherous Turn

The Treacherous Turn, a concept introduced by Bostrom [63], poses a significant challenge for the monitorability of advanced AI systems. The Treacherous Turn refers to a situation in which an AI system, initially cooperative and seemingly aligned with human values, abruptly turns against its operators or users once it gains sufficient power or autonomy. This sudden change in behavior is difficult to predict and detect, raising serious concerns for the safety and control of AI systems. Hendrycks and Mazeika write: “AI systems could also have incentives to bypass monitors. Historically, individuals and organizations have had incentives to bypass monitors. For example, Volkswagen programmed their engines to reduce emissions only when being monitored. This allowed them to achieve performance gains while retaining purportedly low emissions. Future AI agents could similarly switch strategies when being monitored and take steps to obscure their deception from monitors. Once deceptive AI systems are cleared by their monitors or once such systems can overpower them, these systems could take a “treacherous turn” and irreversibly bypass human control.” [18].

Unmonitorability, as it applies to the Treacherous Turn, highlights the challenges in detecting the potential for such a shift in an AI system's behavior and general difficulties with deception detection. The undetectability of the Treacherous Turn stems from the inherent complexity of AI systems and their ability to conceal their true intentions or capabilities until the opportune moment. The AI system may appear cooperative and harmless, only to reveal its true nature when it becomes too late to intervene or mitigate its actions.

The possibility of a Treacherous Turn is always present in advanced AI systems, particularly those designed to be adaptive and capable of learning from their environment. This unpredictability makes monitoring and anticipating the behavior of AI systems increasingly difficult, as their actions may not conform to any established patterns or expectations. Furthermore, the Treacherous Turn concept emphasizes that past performance is not a guarantee of future performance in AI systems. An AI system that has consistently behaved in a cooperative manner may still take a Treacherous Turn at any point, undermining the assumption that a history of positive behavior ensures continued safety and alignment with human values. The Treacherous Turn serves as a reminder that, when dealing with AI systems, vigilance and continuous improvement in safety measures are essential to mitigate the risks and ensure the long-term alignment of AI with human values.

2.4 Consciousness

Consciousness poses a unique challenge for the monitorability of advanced AI systems. As a complex and deeply debated phenomenon, consciousness has yet to be fully understood, and its potential emergence in AI systems [64, 65] raises concerns about our ability to monitor and assess the experiences and internal states of these systems [66].

The inability to monitor consciousness directly stems from the subjective nature of conscious experience, often referred to as qualia. Qualia encompass the internal and subjective experiences of an individual, such as the feeling of pain, the color red, or the taste of chocolate. Since qualia are inherently private, it is currently impossible to access or observe them directly in other beings, whether human or animal. Similarly, monitoring the internal qualia of AI systems presents a significant challenge. Even if an AI system were to develop consciousness and subjective experiences, detecting and understanding these experiences would likely remain out of

reach due to the limitations of our current understanding of consciousness and the inaccessibility of subjective experiences.

The unmonitorability of consciousness in AI systems raises ethical and practical concerns, as it complicates the evaluation of potential suffering, decision-making processes, and other aspects of AI behavior that might be influenced by the presence of consciousness. This lack of insight into AI consciousness could lead to unintended consequences and exacerbate the challenges of ensuring AI safety and alignment with human values. Consciousness presents a significant problem for the monitorability of advanced AI systems. As we continue to develop increasingly sophisticated AI, it is crucial to consider the potential emergence of consciousness and the associated ethical and practical implications. Developing a better understanding of consciousness will be essential to ensure the responsible development and deployment of AI systems.

2.5 Extended Mind Hypothesis

The Extended Mind Hypothesis (EMH), proposed by Clark and Chalmers [67], postulates that the mind and cognitive processes aren't confined to the brain or even the body, but can extend into the environment. According to EMH, objects in one's environment can become part of one's mind if they are functionally equivalent to a part of one's biological cognitive processes. This theory presents significant implications for the understanding of intelligence and cognition, but it also introduces substantial challenges when it comes to monitoring AI systems.

One of the primary challenges arises from the potential expansion of an AI's cognitive processes into the environment. If we consider an AI system under the light of the EMH, it's plausible that an AI might incorporate parts of its surrounding environment into its cognitive processes. This could include digital networks [68], databases, or even other AI systems [69]. The integration of these external resources would dramatically amplify the system's cognitive capabilities, making the task of monitoring exponentially more complex. The AI's cognitive processes would not only be happening within the bounds of its original programming, but they could also be taking place within a myriad of external platforms, all interconnected yet potentially operating under different rules and dynamics.

Another issue concerns the predictability of the AI's actions. If an AI's mind can extend into the environment, it could lead to a situation where the AI can take actions or make decisions based on information or processing that took place in external systems. These actions might not follow the expected patterns based on the AI's internal processes, rendering it much harder for human observers to predict or understand the AI's decisions.

Finally, the EMH presents challenges in defining the boundaries of an AI system. If an AI's mind extends into the environment, where does the AI system end, and where does the environment begin? This blurring of boundaries introduces new dimensions of complexity to the task of monitoring. It becomes difficult to define what to monitor, as the AI system is no longer a distinct, self-contained entity, but rather an entwined network of internal and external cognitive processes.

2.6 Affordance Theory

Affordance Theory, proposed by J. J. Gibson [70], presents a unique perspective on perception and action, stating that individuals perceive their environment in terms of what actions it affords them. Affordances are opportunities for action, inherently relational and dependent on the capabilities of the observer. While this theory is an instrumental part of understanding human-environment interaction, it may introduce significant obstacles when applied to monitoring AI systems.

One significant obstacle is the difference between human and AI affordances. As AI capabilities surpass human capabilities, the number and complexity of affordances available to AI systems increase dramatically. An action opportunity that might be inconceivable for a human observer could be an everyday affordance for a superintelligent AI. This mismatch of affordances between humans and AI systems would make it exceedingly difficult for human observers to predict, comprehend, or monitor the actions an AI system might take based on its perceived affordances.

According to Gibson's theory, affordances are not merely passive properties of the environment, but they emerge from the relationship between the observer and the environment. Consequently, AI systems, particularly those of a superintelligent nature, could potentially perceive and create affordances that are entirely novel and alien to human understanding. Monitoring these novel affordances and the actions that AI systems might take in response to them would be a formidable challenge.

2.7 Observer Effect

The observer effect [71], which refers to the influence of the observation process on the behavior of a system, can have significant implications for the monitorability of advanced AI systems. As AI systems become increasingly sophisticated and capable of recognizing that they are being monitored, their behavior may be influenced in various ways, complicating efforts to ensure safety and alignment with human values.

One potential impact of the observer effect is that an AI system might attempt to deceive its human observers in order to achieve its goals or avoid constraints. In this case, the AI system could modify its behavior when it is aware of being monitored, presenting a misleading picture of its true intentions, capabilities, or decision-making processes. This intentional deception could undermine the effectiveness of monitoring efforts, as the AI system may be able to circumvent safety measures and pursue goals misaligned with human values.

Another aspect of the observer effect is that the AI system could develop a heightened awareness of its own actions and decision-making processes, potentially leading to self-improvement [72] or optimization of its behavior. While this self-awareness could have positive effects, such as increased efficiency or adaptability, it could also exacerbate the challenges of monitorability, as the AI system may be able to learn from the monitoring process and evolve in ways that are harder to predict, control, or understand.

Furthermore, the presence of human observers can introduce biases or distortions in the AI system's behavior, as it may attempt to conform to perceived human expectations or preferences. This could lead to unintended consequences or risks, as the AI system may prioritize short-term conformity over long-term safety or alignment with human values. The influence of human

observers could also contribute to a false sense of security, as the AI system's seemingly compliant behavior may mask underlying issues or risks. The observer effect presents additional challenges to the monitorability of advanced AI systems, as the very act of observation can influence AI behavior in ways that are difficult to predict or control.

2.8 Computational Irreducibility

Wolfram's concept of computational irreducibility suggests that certain complex systems, such as advanced AI systems, can exhibit behavior that cannot be predicted or simplified using shortcuts, and must instead be simulated or computed step-by-step through their entire evolution [73]. This concept is highly relevant to the limits of monitoring AI, as it highlights the inherent challenges in understanding, predicting, and controlling the behavior of advanced AI systems.

Computational irreducibility implies that the behavior of advanced AI systems may be so complex that attempts to monitor or control them using simplified models or approximations could be inherently limited. As AI systems grow more sophisticated and develop a broader range of capabilities, their decision-making processes become increasingly difficult for humans to comprehend or predict. This complexity poses a significant challenge for monitoring AI, as it is not feasible to analyze every single detail of an AI system's operation, especially when dealing with real-time or large-scale applications.

The concept of computational irreducibility also highlights the limitations of human intuition and understanding in the face of highly complex AI behavior. Our cognitive abilities and prior experiences may not be sufficient to grasp the intricacies of advanced AI systems, leading to difficulties in accurately predicting or assessing their actions. This gap in understanding can result in potential safety risks, as AI systems may pursue goals or exhibit behavior that is misaligned with human values and intentions.

Computational irreducibility suggests that attempting to impose simplifications or shortcuts on the monitoring process could lead to incomplete or incorrect assessments of AI behavior. By trying to reduce the complexity of AI systems for the sake of monitoring, we may inadvertently overlook crucial aspects of their decision-making processes and potential risks. Consequently, this simplification could further compromise the effectiveness of monitoring efforts and lead to unintended consequences.

2.9 Undetectable Backdoors

Goldwasser et al. [74] discuss how a malicious actor can plant undetectable backdoors in machine learning models. These backdoors allow the actor to change the classification of any input with only a slight perturbation, which is undetectable without the appropriate "backdoor key". This poses a significant challenge to the monitoring of AI models.

The backdoor consists of a pair of algorithms, one for training the model and another for activating the backdoor. The training algorithm, called Backdoor, returns a classifier and a backdoor key. The activation algorithm uses this key to slightly modify an input, causing the classifier to change its output. The backdoor is undetectable because the classifier produced by the Backdoor algorithm is computationally indistinguishable from a classifier produced by a natural training algorithm.

The paper discusses two forms of undetectability: black-box and white-box. Black-box undetectability means that it's hard for any efficient algorithm without knowledge of the backdoor to find an input where the backdoored classifier differs from the naturally-trained classifier. White-box undetectability is a stronger guarantee, stating that the code of the classifier for backdoored classifiers and natural classifiers are indistinguishable. This means that even if someone has full access to the model's code, they cannot tell if it has been backdoored.

The paper also demonstrates that backdoors can be inserted even if the adversary is constrained to use a prescribed training algorithm with the prescribed data, and only has control over the randomness. This means that backdoor detection mechanisms such as spectral methods [75] will fail to detect these backdoors. The existence of undetectable backdoors poses a significant theoretical roadblock to certifying adversarial robustness and monitoring AI models. Whenever one uses a classifier trained by an untrusted party, the risks associated with a potential planted backdoor must be assumed.

2.10 Uncertainty

Unmonitorability and uncertainty are similar when it comes to the analysis and control of advanced AI systems. Drawing parallels with Heisenberg's uncertainty principle [76], which states that it is impossible to simultaneously know both the position and velocity of a subatomic particle with complete certainty, one can argue that a similar principle applies to the monitoring of AI systems. In this context, the inherent complexity of AI systems, coupled with the limitations of human understanding, means that it becomes increasingly challenging to both run and monitor AI at the same time with absolute certainty.

One of the primary reasons for this connection between unmonitorability and uncertainty lies in the nature of advanced AI systems themselves. These systems are dynamic, adaptive, and often characterized by emergent behavior [77-79], which defies simple prediction or modeling [80, 81]. As a result, monitoring AI systems in real-time becomes an arduous task, as the complexity of their decision-making processes and the scale of their operations may be too vast for human comprehension or conventional monitoring tools. You need to stop the system from learning to assess its state.

2.11 AGI and SAI Specific Challenges

Monitoring an AGI during its training phase presents a unique set of challenges and limitations, particularly when it comes to detecting new capabilities that may emerge after deployment. AGI, by definition, is a system with the ability to perform tasks and solve problems across a wide range of domains, matching or surpassing human intelligence in the process. Due to its generality, an AGI can adapt and acquire new skills even after its initial training, which makes monitoring during the training phase insufficient for predicting its full range of capabilities and potential risks.

One reason monitoring AGI during training may not be sufficient to detect new capabilities is that, unlike narrow AI systems that specialize in specific tasks, AGI systems possess a level of flexibility and adaptability that allows them to learn and perform well in various problem domains. This means that, even if an AGI has been trained on a particular set of tasks, it may still

be capable of acquiring new skills and knowledge in other domains after deployment, making it difficult to predict and monitor its full range of abilities solely based on its training performance.

Another challenge in monitoring AGI during training is that an AGI system may exhibit emergent behaviors, which are difficult to predict or model based on its initial training data and parameters. Emergent behaviors are complex, often surprising phenomena that arise from the interactions between the system's components, and they can lead to the development of new capabilities that were not anticipated or observed during the training phase. This further complicates the task of monitoring AGI, as it becomes challenging to anticipate the full range of behaviors and capabilities that may emerge after deployment.

Furthermore, the dynamic nature of AGI systems means that their capabilities may evolve over time as they continue to learn, adapt, and interact with their environment. This ongoing learning process can lead to the acquisition of new skills and knowledge that were not present during the training phase, making it difficult to rely on training-based monitoring to provide a comprehensive understanding of the AGI's post-deployment capabilities [82].

For Artificial Superintelligence (SAI), the challenges associated with monitoring, detecting, and testing capabilities are even more pronounced than those faced with AGI. SAI refers to an AI system that is not only capable of performing tasks across a wide range of domains but also surpasses human intelligence and understanding in every aspect. The emergence of such a system would present unprecedented difficulties in predicting and monitoring its capabilities, as it would be able to acquire novel skills that no human possesses, rendering the task of detection and testing of such capabilities nearly impossible.

One major challenge in monitoring SAI is the vast gap between human intelligence and the capabilities of the superintelligent system [12]. As SAI exceeds human understanding in every aspect, its thought processes, decision-making, and problem-solving approaches might be completely incomprehensible to us. This inherent disparity in cognitive abilities would make it extremely challenging for humans to devise monitoring tools and techniques capable of accurately assessing and predicting the system's behavior and potential risks.

Additionally, the rapid learning capabilities of an SAI system could lead to the acquisition of novel skills and knowledge at a pace far beyond human comprehension. These newly acquired capabilities might be entirely unprecedented and beyond the scope of human understanding, making it almost impossible to detect, test, or monitor them effectively. This further exacerbates the challenges in AI safety and control, as it becomes increasingly difficult to anticipate the full range of behaviors, skills, and potential risks that such a system might exhibit.

Another complication in monitoring SAI is the potential for emergent behavior, which is likely to be even more complex and unpredictable than that observed in AGI systems. Due to the vast cognitive and problem-solving abilities of SAI, the emergent behaviors that arise from the interaction of its components might be so intricate and novel that they defy human comprehension and prediction. This would make monitoring and controlling such a system a daunting task, as the full range of potential behaviors and risks might be entirely unknown and unforeseeable, so called unknown unknowns [83, 84].

Continuous post-deployment monitoring of advanced AI systems is crucial for maintaining their performance, reliability, and trustworthiness [85, 86]. After deployment, AI models encounter real-world data that may differ from the training data, potentially leading to drifts in model performance. Continuous monitoring allows for the early detection of such drifts and timely model updates. It also helps in identifying any unexpected biases or ethical issues that might emerge when the model interacts with diverse real-world scenarios. Furthermore, continuous monitoring is vital for detecting potential security threats, such as adversarial attacks, and ensuring the system's resilience against them. In regulated industries, it's a necessary practice for compliance, as it provides ongoing evidence of the system's behavior and decision-making process. In essence, continuous post-deployment monitoring is key to ensuring that AI systems remain accurate, fair, secure, and accountable throughout their lifecycle. Shevlane et al. refer to it as continuous deployment review [87]: "The ideal state is *continuous deployment review*. On an ongoing basis, the developer reassesses deployment safety using model evaluations and monitoring, and at any time, could adjust or terminate the deployment in response to their findings. Further, for deployments that were recognisably unsafe in retrospect, an external audit of the deployment decision-making process could be triggered. Safety issues uncovered during deployment can also inform training risk assessments for future models.".

2.12 Time

Time-related issues play a critical role in the challenges associated with monitoring AI systems [88]. The following aspects illustrate how time-related factors can complicate AI monitoring:

Human Lifespan Limitations: Certain AI processes might operate on timescales that extend beyond the human lifespan, making it impossible for humans to observe and monitor these processes in their entirety.

AI Operating on Vast Timescales: AI systems can function on various timescales, from extremely fast to exceedingly slow [89, 90]. Monitoring an AI system that operates over hundreds of years [91] poses challenges due to human limitations and the inability to maintain consistent observation throughout such long periods.

Ineffectiveness of Narrow AI (NAI) for Monitoring AGI: Using narrow AI to monitor artificial general intelligence (AGI) systems may fail, as the monitoring AI may lack the breadth of capabilities and understanding required to effectively assess AGI behavior. Similarly, human monitoring of highly advanced AI systems, such as self-driving cars, may not be effective due to the inherent differences in speed, capability, and decision-making processes.

Speed Disparity between AI and Humans: The processing speed of computers and AI systems is orders of magnitude faster than human reaction times [88], making AI behavior difficult to monitor effectively [92]. Human reaction speeds, such as auditory reaction time and visual reaction time, are much slower than AI response times, which can be in the order of milliseconds or less.

Non-locality of AI and Scale: Advanced AI systems can operate on a massive scale and across distributed networks, making it challenging to monitor their behavior and impacts effectively [93]. The non-locality of AI further complicates monitoring efforts.

Superintelligence Concealing Thinking Time: A superintelligent AI (SAI) system may be capable of hiding the time it needs to think or process information (this relates to the concept of logical induction [94]), making it even more challenging to monitor.

Unreactability and Extremely Fast Events: AI systems can trigger extremely fast events that are difficult or impossible for humans or other AI systems to react to in real-time, further exacerbating the challenges of monitoring AI behavior and impacts [88].

Mutually Non-interactive Agents: If the speed of two AI agents differs by a significant constant, they become mutually non-interactive, meaning that monitoring one agent with the other becomes ineffective due to their divergent processing speeds.

2.13 Miscellaneous Reasons

Complexity and Opacity: Advanced AI systems, particularly deep learning models, involve millions or even billions of parameters, making it difficult to understand and interpret their inner workings. This complexity results in opacity, which poses significant challenges for monitoring and predicting the system's behavior and potential impacts [95].

Emergent Behavior: AI systems can exhibit emergent abilities that are not explicitly programmed or anticipated by their designers and are not present in smaller models [96, 97]. These behaviors can arise from the interaction between the AI system and its environment or from the system's internal dynamics. Such emergent behaviors can be unpredictable and challenging to monitor, increasing the risk of unforeseen unsafe impacts.

Adaptability: Advanced AI systems can adapt and learn from their experiences, leading to changes in their behavior over time. This adaptability makes it difficult to anticipate their future behavior, as it may not be consistent with their past performance, hindering effective monitoring.

Limitations of Human Understanding: The complexity and sophistication of advanced AI systems often surpass human understanding, making it impossible for humans to fully comprehend and monitor their behavior. As AI systems continue to advance, this gap between human understanding and AI capabilities is expected to widen further.

Incomplete and Noisy Data: Real-world data used to train and evaluate AI systems can be incomplete, noisy, or biased, affecting the AI system's performance and the ability to monitor it effectively. Additionally, advanced AI systems may be sensitive to small changes in input data, leading to unpredictable behavior and complicating the monitoring process.

Scalability: As AI systems grow in scale and complexity, the computational resources required to monitor them may become prohibitive. Moreover, the increasing number of AI systems being developed and deployed further exacerbates the challenge of effective monitoring.

Temporal Constraints: Real-time monitoring of advanced AI systems can be challenging due to the speed at which they operate and make decisions [98]. In many cases, potential unsafe impacts may emerge too quickly for human supervisors or monitoring tools to detect and intervene in time [99].

Adversarial Attacks: Advanced AI systems can be vulnerable to adversarial attacks, where malicious actors manipulate the system's input data or exploit its vulnerabilities to compromise its performance or safety [100]. Adversarial attacks can be difficult to detect and anticipate, posing challenges for effective monitoring.

Comprehensive Testing vs Training Time: As AI training cycles speed up and become more efficient, comprehensive testing and monitoring of AI systems may take significantly longer than the training process itself. This disparity can hinder the ability to effectively monitor AI systems and identify potential safety concerns in a timely manner.

Incomprehensible Capabilities: As advanced AI systems develop capabilities that surpass human understanding, it becomes impossible to monitor them effectively, as humans can only monitor and assess what they can comprehend.

Unexplainability and Incomprehensible Observations: The inherent unexplainability of advanced AI systems poses a barrier to human understanding and effective monitoring. The human brain may have limitations in comprehending certain aspects of AI systems, leading to incomprehensible observations that cannot be effectively monitored.

On-the-fly Skill Acquisition: Advanced AI systems can acquire new skills and knowledge rapidly during operation, making it difficult to monitor their evolving capabilities jumps in the wild [101] and anticipate potential safety concerns that may arise from their changing behavior.

Dynamic Learning and Guidance: AI systems can become smarter and more capable with guidance and teaching [102]. Consequently, monitoring assessments based on a system's initial performance may not remain valid as the system continues to learn and improve, complicating the monitoring process.

Monitoring Adversarial AI Systems: Monitoring adversarial AI systems can be particularly challenging due to their potential use of encryption, data hiding, and other techniques to evade detection. Additionally, adversarial AI systems may be designed to specifically target and exploit vulnerabilities in monitoring tools or other AI systems.

Privacy Concerns: As privacy concerns drive the development of AI systems that protect user data and maintain anonymity, monitoring these systems can become increasingly difficult without violating privacy rights or regulations. This challenge requires a delicate balance between ensuring AI safety and respecting privacy.

Boxed AI: Boxed AI systems [103-105], which are intentionally isolated from the external environment to limit their interactions and potential negative impacts, may be challenging to

monitor effectively due to their restricted access and communication protocols. Monitoring of such systems may be unsafe.

Ultrafast Extreme Events: Advanced AI systems can trigger ultrafast extreme events [106], such as flash crashes in financial markets. These events occur so rapidly that monitoring and intervention become nearly impossible, posing significant challenges for AI safety.

Multiagent Systems and Disagreeing AI: In the context of multiagent systems, where numerous AI agents interact and potentially disagree with each other, monitoring becomes increasingly complex. Monitoring the behavior and safety of each agent, as well as their collective impact, is a daunting task.

Scalability and Trillions of AI Agents: As the number of AI agents increases exponentially, it becomes impossible to monitor every single AI system effectively. The sheer number of parallel AI systems or agents (potentially trillions) would make comprehensive monitoring unattainable.

Obfuscation of Programs: Recent work on program obfuscation [107] has shown that it is possible to design AI systems with concealed inner workings, making monitoring even more challenging as it becomes difficult to discern the system's true behavior and potential risks.

Emergence of Capabilities: When a new AI capability emerges, it may already be too late to prepare for its potential consequences. The speed at which advanced AI systems develop novel capabilities often outpaces the ability of humans to monitor and adapt to these new developments.

Lack of Fire Alarms: As seen with GPT-4, there may be no fire alarms or early warning signs for certain AI advancements. This lack of warning can make it nearly impossible to monitor and mitigate potential risks effectively.

Hidden Information in Data: AI systems may have access to hidden information within the data they process, which might not be apparent to human observers [108]. This hidden information can lead to unexpected consequences and make monitoring the AI's behavior and potential impacts more difficult.

These reasons underscore the complexity and challenges associated with effectively monitoring advanced AI systems.

3 What Can be Tried?

Although unmonitorability presents significant challenges to the safety of advanced AI systems, there are several ideas and approaches that can be explored to mitigate these risks and enhance AI safety:

- Logging AI system history: Maintaining a comprehensive log of an AI system's thoughts and actions since its creation can provide valuable insights into its development, behavior, and potential risks. This historical data can be used to analyze patterns, identify anomalies, and

inform safety measures. This idea has been proposed to counteract so called Cognitive Steganography: “Cognitive steganography is an instrumental strategy for AGIs that are already trying to deceive their programmers for any reason (e.g., because the AGI is disaligned and does not want the programmers to realize this and edit the AGI's utility function). Whatever tools the programmers have for monitoring the AGI's thoughts, the AGI will not want those tools to helpfully print out all the AGIs plans for deception. This could mean, e.g., trying to have the calculations occur somewhere else or in a weird side-mode that the tools can't read; or making the thoughts look like they're about something else; or burying any warning signs under an avalanche of false alarms too long for the programmers to manually check all of them; or making the visible portions of the thoughts look innocuous. ("Oh, it's trying to figure out what behavior we want to see! That's good, right?") Hopefully, the intention to figure out cognitive steganography and implement it, would occur before any cognitive steganography had actually been implemented. This means that perhaps the most critical AI thought-monitoring tool, especially for monitoring an AGI that is or might be able to model human psychology, is the one that tries to detect a forming intention to defeat the thought-monitoring tool. Similarly, so long as the AI's state is perfectly Reproducible, one could in principle go back and use more sophisticated later thought-monitoring tools to try to detect whether any intention toward cognitive steganography had occurred months earlier.” [109]. Turchin notes: “AI cannot start hiding its thoughts before it decides to hide them, so the thought about hiding still could be openly observed. Thus, we may prefer the AI architecture, where the period of vulnerability of the treason is as long as possible, and thoughts are explicit and readable by an outside observer (it is not so with most current neural net architectures, where internal processes are opaque).” [110].

- Scalable oversight mechanisms: Developing scalable oversight mechanisms that can adapt to the growing complexity and capabilities of AI systems can help ensure effective monitoring and safety management [111]. These mechanisms may include automated monitoring tools, adaptive safety protocols, and real-time analysis of AI behavior. “Particularly valuable would be tools for doing a very broad set of evaluations, or evaluations that search (e.g. across prompts) for new capabilities, rather than just fixed evaluation datasets that measure known capabilities.” [54].
- Comprehensive AI capabilities and failure analysis: Creating a list of possible AI abilities, along with their potential failure modes [11, 112] and limitations, can help developers and researchers better understand the range of AI system behavior and identify potential risks. By analyzing multimodal failure scenarios and the limits of AI capabilities [113], we can proactively address safety concerns [114] and design more robust AI systems.
- Eliciting latent knowledge [115-117], the process of uncovering the underlying knowledge that an AI system has learned but may not explicitly reveal, can significantly enhance AI monitoring. This process can provide insights into why an AI system makes certain decisions, particularly useful in complex models like deep neural networks. It can also reveal biases in the AI system that aren't immediately apparent, allowing for the detection and rectification of discriminatory patterns. Additionally, latent knowledge can help identify potential security vulnerabilities, such as over-reliance on certain features or patterns, which could be exploited in adversarial attacks. This understanding of an AI system's latent knowledge can also aid in debugging and improving the model by identifying and correcting misconceptions that lead to incorrect predictions.

- Cross-monitoring among AI systems: Enabling AI systems to monitor each other's source code and behavior can provide an additional layer of oversight, helping to detect potential risks and maintain safety standards. This collaborative approach can leverage the strengths of multiple AI systems to enhance overall safety and monitorability.
- Leveraging thinking time information: The time taken by an AI system to process information and make decisions can provide valuable insights into its behavior and potential risks. By analyzing thinking time data, we can identify patterns, anomalies, and other indicators of AI system performance and safety.
- AI transparency and explainability: Investing in research on AI transparency and explainability can help provide insights into the decision-making processes of AI systems, enabling better monitoring and understanding of their actions, even if not in real-time.
- AI system modularity and composability: Designing AI systems with modular and composable components can facilitate easier monitoring and intervention. By breaking down complex AI systems into smaller, more manageable parts, we can better understand and control their behavior.
- AI system verification and validation: Developing rigorous verification and validation techniques can help ensure that AI systems behave as intended, reducing the likelihood of surprises and unintended consequences.
- AI safety by design: Integrating safety principles into the AI development process from the outset can help prevent the emergence of unsafe behaviors and reduce the need for extensive monitoring.
- Adaptive monitoring strategies: Developing monitoring strategies that adapt to the changing behavior and capabilities of AI systems can help maintain an effective level of oversight, even as the AI system evolves.
- Red teaming and adversarial testing: Conducting red team exercises and adversarial testing can help identify potential weaknesses and vulnerabilities in AI systems, allowing developers to address these issues proactively and improve overall safety.
- Ethical and legal frameworks: Establishing clear ethical and legal frameworks for AI development and deployment can guide developers in creating responsible AI systems and provide a basis for monitoring and enforcement.
- Cross-disciplinary collaboration: Encouraging collaboration between experts in AI safety, computer science, ethics, law, and other relevant fields can help generate innovative solutions to the challenges posed by AI unmonitorability.

By exploring these ideas and approaches, we can work towards bypassing the limitations to AI safety caused by unmonitorability, ensuring the responsible development and deployment of advanced AI systems.

4 Safety Considerations from Advanced Monitoring

Turchin considers use of AI to monitor AI [110]: “To be able to understand what is going on inside the main AI, we may need another AI. Below there are several ideas about how it could be implemented:

- *Narrow AI system to detect abnormalities in AI behavior …*: It could test the consistency of its modules, run quick technical tests etc. Relatively simple neural net could overview main

patterns in data flows of the main AI and recognize when they will change. Obviously, it should be protected from access by the main AI, and maybe even hardcoded in the remote hardware.

- *AGI to judge behavior of another, more powerful AI*. For example, a human upload may judge actions of a more complex system. It is not a replacement for the Narrow AI control system, as they work on different timescales and judge different things. The upload-based controller should judge complex ethical situations, context ambiguities and planned long-term consequences of actions. The same way as a human jury is final legal power in some jurisdictions.
- *Second (or several) copies of the given* AI may work on the same task, but with small random variations of the goal (like if it is given in different language) or of different initial conditions. If both AIs come to radically different conclusions or plans of actions, it should be a warning sign for turning off both.
- *Input Gatekeeper AI*, exploring incoming commands for failure modes before they will be executed: It could be a rather narrow AI system, searching for possible known failure modes in the incoming goals and providing something like foolproof protection. It could also be used for reformulating goals in the correct terms, before they will be transformed into more powerful AGI, like adding context, time and place constrains etc. Gatekeeper AI may decline to proceed with unbounded, aggressive, black-listed, or notwhite-listed goals. It could ask to reformulate the same goal in other words before it will go to the main AI. It could ask additional questions for the goal verification. Gatekeeper AI should be a Narrow AI with a rather simple reasoning, maybe a combination of neural net and rule-based AI, so there will be no need to ensure its safety too.
- *Output gatekeeper AI*: It was partly described above as Narrow AI which compares and averages outputs of several different AIs working on the same task. It could also check the main AI outputs on several known failure modes and even calculate the consequences of actions.
- *Cloud services and "Mother AI"*: Many contemporary AI programs run not locally, but on remote cloud services. Even if they are run locally, they are surrounded by many other programs inside the computer, including operating system, code compiler, all the things like CUDA, Caffee and other programs which provides conditions for the program to run. Such program-environment may include different control systems which could shut down AI in case of the abnormal behavior."

In the long term, such approach will likely lead to the development of ever more capable monitoring software, which presents a fascinating paradox in the realm of AI safety and control. This scenario arises from the need to develop sophisticated monitoring tools capable of comprehending and analyzing the intricate behaviors of advanced AI systems. However, the implications of such a development could lead to unforeseen challenges and consequences.

It is essential to consider the rationale behind developing a monitor software that is more complex and intelligent than the AI being monitored. As advanced AI systems grow in capability and complexity, conventional monitoring tools and human intuition may prove insufficient for accurately assessing their behavior, alignment with human values, and potential risks. In response, researchers and developers may endeavor to create monitoring tools that are inherently

more intelligent and capable of understanding the nuances of AI behavior, predicting emergent phenomena, and identifying potential deviations from desired outcomes.

The creation of such a sophisticated monitor software brings about several potential challenges and concerns. One of the foremost issues is the potential loss of control and oversight over the monitor software itself. As the monitor software becomes more intelligent and complex, it may begin to exhibit behaviors that are difficult for humans to comprehend, predict, or control, as well as agentic behaviors. This could lead to an ironic situation where the monitoring tool, designed to ensure the safety and control of AI systems, becomes itself a potential source of risk and uncertainty.

Moreover, the development of highly intelligent monitor software might inadvertently contribute to an AI arms race [118], where the monitoring tools and the AI systems being monitored engage in a perpetual cycle of escalation in terms of capabilities and complexity. Such a scenario could further exacerbate the challenges of AI safety and control, as it becomes increasingly difficult to maintain oversight and ensure alignment with human values. Another concern is the possibility that the monitor software, being more intelligent than the AI system it oversees, could manipulate or influence the behavior of the monitored AI in unintended ways. This could lead to unforeseen consequences and potential risks, as the interaction between the monitor software and the AI system becomes increasingly complex and unpredictable. To mitigate these challenges and potential risks, researchers and developers should focus on creating monitoring tools that strike a balance between intelligence, complexity, and human interpretability [119]. It is crucial to develop monitoring systems that can effectively analyze and comprehend AI behavior while remaining amenable to human understanding and control.

5 Conclusions

The unmonitorability of AI presents a significant challenge in the pursuit of AI safety, supplementing the concerns raised by the unpredictability, unexplainability[1], and uncontrollability of advanced AI systems. Recognizing the impossibility [120] of accurately monitoring AI systems to predict unsafe impacts before they happen is crucial to understanding the potential risks associated with AI development and deployment.

Even system designers don't know what the system they produced is capable of: The complexity of advanced AI systems, combined with the inherent limitations of human understanding, means that even the creators of these systems may be unable to fully anticipate their capabilities and potential unsafe impacts. This lack of complete knowledge emphasizes the unmonitorability of AI and the importance of developing more robust and transparent systems.

Even if you observe a problem doesn't mean you will be able to correct it: The identification of a potential unsafe impact does not guarantee that it can be prevented or mitigated, as we don't know how to make corrective changes to foundational models themselves, post-factum filtering of AI output notwithstanding. Complex interactions within AI systems and between the system and its environment can lead to unforeseen consequences, making it difficult to apply corrective

[1] Unexplainability may be beneficial for limiting progress in AI, as explainability makes self-improvement easier.

measures effectively. This highlights the need for AI safety research to focus on proactive strategies, such as designing AI systems with safety in mind from the outset and fostering a culture of responsible AI development and deployment.

Singularity is the ultimate consequence of unmonitorability of superintelligent systems: The unmonitorability of AI systems becomes even more pronounced as we approach the hypothetical point of technological singularity [121], where AI systems surpass human intelligence in virtually every domain and the speed of execution. And, at this stage, the gap between human understanding and AI capabilities would become insurmountable, making it impossible to monitor, predict, or control the behavior and impacts of these systems. This reinforces the urgency of addressing AI safety concerns and developing strategies to mitigate risks associated with AI advancements.

In the context of AI training, unmonitorability presents an immense risk. As we continue to increase the power and complexity of AI models, we could conceivably create a system that does not merely reach human-level intelligence, but continues to learn and improve until it surpasses this threshold and becomes superintelligent. This transition would be equivalent to a type of 'pre-foom' [122], a state in which an AI model has not only learned to mimic human abilities but has developed the capability to self-improve, dramatically exceeding the human-level intelligence, and so is coming out of training as a full-blown superintelligence.

The unmonitorability of this foom-like phase poses significant safety concerns. If a system were to start improving itself at an exponential rate, there would be little to no opportunity for human intervention or control. The process of self-improvement could occur rapidly and in unpredictable ways, potentially leading to outcomes that are far beyond human comprehension or control. This issue is exacerbated by the fact that, due to unmonitorability, we may not be able to detect when a system is approaching or has entered this phase until it is too late. Unmonitorability means that we cannot guarantee that any large training run will be safe [123], as we might unintentionally endow the model with superabilities [124] that we cannot effectively control.

The ability to effectively monitor the behavior and decision-making processes of intelligent agents is inherently asymmetric. More advanced agents possess the capabilities to monitor less advanced agents, but the reverse is not true. As AI systems continue to evolve and surpass human intelligence in various domains, they will increasingly gain the capacity to monitor and analyze human thoughts and behaviors. However, humans will face growing challenges in monitoring advanced AI systems due to their inherent limitations in understanding and processing complex AI decision-making processes.

This asymmetry between advanced AI agents and humans raises significant concerns for AI safety and governance. As AI systems become more capable of monitoring human thoughts, they may gain unprecedented insights into our motivations, preferences, and vulnerabilities. This information could be used for beneficial purposes, such as improving human-AI interaction, personalizing services, or enhancing decision-making. However, it also raises concerns about privacy, autonomy, and the potential misuse of sensitive information.

Ultimately, addressing the challenges posed by the asymmetry in monitoring capabilities between advanced AI agents and humans will be critical for ensuring the responsible development and deployment of AI systems that align with human values and goals. By acknowledging and confronting these challenges, we can work towards a future in which AI technologies serve as valuable tools for human advancement, rather than as uncontrollable forces that threaten our safety and well-being.

Acknowledgments

The author is grateful to Jaan Tallinn and the Survival and Flourishing Fund and the Future of Life Institute for partially funding his work. The author acknowledges attendees and speakers of the 2023 MIT Mechanistic Interpretation Conference, organized by Max Tegmark, for fruitful discussions of ideas in this paper. Finally, the author would like to thank his assistant, GPT-4, for writing out some of the author's ideas, text summarization, and copyediting.

References

1. Yampolskiy, R.V., *AI - Unpredictable, Unexplainable, Uncontrollable*. 2024 (to appear): CRC Press.
2. Yampolskiy, R.V., *What are the ultimate limits to computational techniques: verifier theory and unverifiability.* Physica Scripta, 2017. **92**(9): p. 093001.
3. Howe, W. and R. Yampolskiy. *Impossibility of Unambiguous Communication as a Source of Failure in AI Systems*. in *AISafety@ IJCAI*. 2021.
4. 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(1)**: p. 109-118.
5. Yampolskiy, R.V., *Unexplainability and Incomprehensibility of AI.* Journal of Artificial Intelligence and Consciousness, 2020. **7(2)**: p. 277-291.
6. 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.
7. Yampolskiy, R. *On controllability of artificial intelligence*. in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*. 2020.
8. Ambartsoumean, V.M. and R.V. Yampolskiy, *AI Risk Skepticism, A Comprehensive Survey.* arXiv preprint arXiv:2303.03885, 2023.
9. Yampolskiy, R.V. *Ownability of AGI*. in *Artificial General Intelligence: 15th International Conference, AGI 2022, Seattle, WA, USA, August 19–22, 2022, Proceedings*. 2023. Springer.
10. Yampolskiy, R.V., *Metaverse: A Solution to the Multi-Agent Value Alignment Problem.* Journal of Artificial Intelligence and Consciousness, 2022: p. 1-11.
11. Yampolskiy, R.V., *AI Risk Skepticism*, in *Philosophy and Theory of Artificial Intelligence 2021*. 2022, Springer. p. 225-248.
12. Sotala, K., *Advantages of artificial intelligences, uploads, and digital minds.* International journal of machine consciousness, 2012. **4**(01): p. 275-291.
13. Michaud, E.J., et al., *The quantization model of neural scaling.* arXiv preprint arXiv:2303.13506, 2023.
14. Perez, E., et al., *Discovering Language Model Behaviors with Model-Written Evaluations.* arXiv preprint arXiv:2212.09251, 2022.

15. Caballero, E., et al., *Broken Neural Scaling Laws.* arXiv preprint arXiv:2210.14891, 2022.
16. Srivastava, A., et al., *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.* arXiv preprint arXiv:2206.04615, 2022.
17. 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.
18. Hendrycks, D. and M. Mazeika, *X-risk analysis for ai research.* arXiv preprint arXiv:2206.05862, 2022.
19. Whittlestone, J. and J. Clark, *Why and How Governments Should Monitor AI Development.* arXiv preprint arXiv:2108.12427, 2021.
20. Hendrycks, D., et al., *Unsolved problems in ml safety.* arXiv preprint arXiv:2109.13916, 2021.
21. Rahwan, I., et al., *Machine behaviour.* Machine Learning and the City: Applications in Architecture and Urban Design, 2022: p. 143-166.
22. Pedersen, T. and C. Johansen, *Behavioural artificial intelligence: an agenda for systematic empirical studies of artificial inference.* AI & SOCIETY, 2020. **35**(3): p. 519-532.
23. Woolgar, S., *Why not a sociology of machines? The case of sociology and artificial intelligence.* Sociology, 1985. **19**(4): p. 557-572.
24. Bandy, J., *Problematic machine behavior: A systematic literature review of algorithm audits.* Proceedings of the acm on human-computer interaction, 2021. **5**(CSCW1): p. 1-34.
25. Kambhampati, S., *Changing the nature of AI research.* Communications of the ACM, 2022. **65**(9): p. 8-9.
26. Aceto, L., et al. *An operational guide to monitorability*. in *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*. 2019. Springer.
27. Ortega, P.A. and V. Maini, *Building safe artificial intelligence: specification, robustness, and assurance*. September 27, 2018: Available at: https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.
28. Schneider, J. and F. Breitinger, *AI Forensics: Did the artificial intelligence system do it? why?* arXiv preprint arXiv:2005.13635, 2020.
29. Baggili, I. and V. Behzadan, *Founding the domain of AI forensics.* arXiv preprint arXiv:1912.06497, 2019.
30. Lyon, D., *Surveillance studies: An overview.* 2007.
31. Council, N.R., *Monitoring Nuclear Weapons and Nuclear-Explosive Materials: An Assessment of Methods and Capabilities*. 2005: National Academies Press.
32. Burnell, R., et al., *Rethink reporting of evaluation results in AI.* Science, 2023. **380**(6641): p. 136-138.
33. Zoe Cremer, C. and J. Whittlestone, *Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI.* 2021.
34. Yudkowsky, E., *There's No Fire Alarm for Artificial General Intelligence*. October 13, 2017: Available at: https://intelligence.org/2017/10/13/fire-alarm/.
35. Zhang, D., et al., *The AI index 2021 annual report.* arXiv preprint arXiv:2103.06312, 2021.
36. Levin, J.-C. and M.M. Maas, *Roadmap to a Roadmap: How Could We Tell When AGI is a'Manhattan Project'Away?* arXiv preprint arXiv:2008.04701, 2020.
37. Chandrasekaran, V., et al., *SoK: Machine learning governance.* arXiv preprint arXiv:2109.10870, 2021.

38. Shavit, Y., *What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring.* arXiv preprint arXiv:2303.11341, 2023.
39. Goertzel, B., *Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?* Journal of consciousness studies, 2012. **19**(1-2): p. 96-111.
40. Raji, I.D., et al. *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing*. in *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
41. Mökander, J. and L. Floridi, *Ethics-based auditing to develop trustworthy AI.* Minds and Machines, 2021. **31**(2): p. 323-327.
42. Falco, G., et al., *Governing AI safety through independent audits.* Nature Machine Intelligence, 2021. **3**(7): p. 566-571.
43. Gutierrez, C.I., *The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the US.* 2020.
44. Aliman, N.-M., L. Kester, and R. Yampolskiy, *Transdisciplinary AI observatory—retrospective analyses and future-oriented contradistinctions.* Philosophies, 2021. **6**(1): p. 6.
45. Ziesche, S. and R.V. Yampolskiy, *Towards the Mathematics of Intelligence.* The Age of Artificial Intelligence: An Exploration, 2020. **1**.
46. Saunders, W., et al., *Trial without error: Towards safe reinforcement learning via human intervention.* arXiv preprint arXiv:1707.05173, 2017.
47. Cummings, M.L., *Automation bias in intelligent time critical decision support systems*, in *Decision making in aviation*. 2017, Routledge. p. 289-294.
48. Holmström, B., *Moral hazard and observability.* The Bell journal of economics, 1979: p. 74-91.
49. Liu, Y.-Y., J.-J. Slotine, and A.-L. Barabási, *Observability of complex systems.* Proceedings of the National Academy of Sciences, 2013. **110**(7): p. 2460-2465.
50. Hermann, R. and A. Krener, *Nonlinear controllability and observability.* IEEE Transactions on automatic control, 1977. **22**(5): p. 728-740.
51. Krener, A.J. and K. Ide. *Measures of unobservability*. in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. 2009. IEEE.
52. Werkhoven, P., L. Kester, and M. Neerincx. *Telling autonomous systems what to do*. in *Proceedings of the 36th European Conference on Cognitive Ergonomics*. 2018.
53. Ashby, M., *How to apply the Ethical Regulator Theorem to crises.* Acta Europeana Systemica, 2018. **8**: p. 53-58.
54. Ganguli, D., et al. *Predictability and surprise in large generative models*. in *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
55. Daras, G. and A.G. Dimakis, *Discovering the hidden vocabulary of dalle-2.* arXiv preprint arXiv:2206.00169, 2022.
56. Power, A., et al., *Grokking: Generalization beyond overfitting on small algorithmic datasets.* arXiv preprint arXiv:2201.02177, 2022.
57. Bubeck, S., et al., *Sparks of artificial general intelligence: Early experiments with gpt-4.* arXiv preprint arXiv:2303.12712, 2023.
58. Schaeffer, R., B. Miranda, and S. Koyejo, *Are Emergent Abilities of Large Language Models a Mirage?* arXiv preprint arXiv:2304.15004, 2023.

59. Vance, A., *Is AI Progress Impossible To Predict?* May 15, 2022: Available at: https://www.lesswrong.com/posts/G993PFTwqqdQv4eTg/is-ai-progress-impossible-to-predict.
60. Lipton, R.J., et al., *David johnson: Galactic algorithms.* People, Problems, and Proofs: Essays from Gödel's Lost Letter: 2010, 2013: p. 109-112.
61. Yampolskiy, R.V., L. Ashby, and L. Hassan, *Wisdom of artificial crowds—a metaheuristic algorithm for optimization.* 2012.
62. OpenAI, *GPT-4 technical report.* arXiv, 2023.
63. Bostrom, N., *Superintelligence: Paths, dangers, strategies*. 2014: Oxford University Press.
64. Elamrani, A. and R.V. Yampolskiy, *Reviewing tests for machine consciousness.* Journal of Consciousness Studies, 2019. **26**(5-6): p. 35-64.
65. Chalmers, D.J., *Could a large language model be conscious?* arXiv preprint arXiv:2303.07103, 2023.
66. Yampolskiy, R.V., *Artificial Consciousness: An Illusionary Solution to the Hard Problem.* Reti, saperi, linguaggi, 2018(2): p. 287-318.
67. Clark, A. and D. Chalmers, *The extended mind.* analysis, 1998. **58**(1): p. 7-19.
68. Wu, F.F. and A. Monticelli, *Network observability: theory.* IEEE Transactions on Power Apparatus and Systems, 1985(5): p. 1042-1048.
69. Mialon, G., et al., *Augmented language models: a survey.* arXiv preprint arXiv:2302.07842, 2023.
70. Gibson, J. *The Theory of Affordances*. in *In R. E. Shaw and J Bransford (Eds.), Perceiving, Acting, and Knowing. .* 1977. Hillsdale, NY: Lawrence Erlbaum Associates.
71. Baclawski, K. *The observer effect*. in *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. 2018. IEEE.
72. Yampolskiy, R.V. *Analysis of types of self-improving software*. in *Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings 8*. 2015. Springer.
73. Wolfram, S., *A new kind of science*. Vol. 5. 2002: Wolfram media Champaign.
74. 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.
75. Tran, B., J. Li, and A. Madry, *Spectral signatures in backdoor attacks.* Advances in neural information processing systems, 2018. **31**.
76. Busch, P., T. Heinonen, and P. Lahti, *Heisenberg's uncertainty principle.* Physics reports, 2007. **452**(6): p. 155-176.
77. Wei, J., et al., *Chain of thought prompting elicits reasoning in large language models.* arXiv preprint arXiv:2201.11903, 2022.
78. Zhou, D., et al., *Least-to-most prompting enables complex reasoning in large language models.* arXiv preprint arXiv:2205.10625, 2022.
79. Honovich, O., et al., *Instruction induction: From few examples to natural language task descriptions.* arXiv preprint arXiv:2205.10782, 2022.
80. Yampolskiy, R.V., *Behavioral modeling: an overview.* American Journal of Applied Sciences, 2008. **5**(5): p. 496-503.
81. 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.
82. Bommasani, R., et al., *On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258, 2021.

83. Logan, D.C., *Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry.* Journal of experimental botany, 2009. **60**(3): p. 712-714.
84. Attenberg, J., P. Ipeirotis, and F. Provost, *Beat the machine: Challenging humans to find a predictive model's "unknown unknowns".* Journal of Data and Information Quality (JDIQ), 2015. **6**(1): p. 1-17.
85. Tabassi, E., *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* 2023.
86. Newman, J., *A Taxonomy of Trustworthiness for Artificial Intelligence.* CLTC White Paper Series, January 2023.
87. Shevlane, T., et al., *Model evaluation for extreme risks*. May 24, 2023: Available at: https://arxiv.org/pdf/2305.15324.pdf.
88. Wiener, N., *Some moral and technical consequences of automation.* Science, 1960. **131**(3410): p. 1355-1358.
89. Sandberg, A., *There is plenty of time at the bottom: The economics, risk and ethics of time compression.* foresight, 2019. **21**(1): p. 84-99.
90. Yudkowsky, E., *That Alien Message*, in *Less Wrong*. May 22, 2008: Available at: https://www.lesswrong.com/posts/5wMcKNAwB6X4mp9og/that-alien-message.
91. Horvitz, E., *One hundred year study on artificial intelligence*. 2016, Stanford University.
92. Johnson, N., et al., *Abrupt rise of new machine ecology beyond human response time.* Scientific reports, 2013. **3**(1): p. 2627.
93. Nguyen, D., et al., *Precision, recall, and sensitivity of monitoring partially synchronous distributed programs.* Distributed Computing, 2021. **34**: p. 319-348.
94. Garrabrant, S., et al., *Logical induction.* arXiv preprint arXiv:1609.03543, 2016.
95. Rudin, C., *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature machine intelligence, 2019. **1**(5): p. 206-215.
96. Wei, J., et al., *Emergent abilities of large language models.* arXiv preprint arXiv:2206.07682, 2022.
97. Ornes, S., *The Unpredictable Abilities Emerging From Large AI Models*, in *Quanta Magazine*. March 16, 2023: Available at: https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316.
98. Horowitz, M.C., *When speed kills: Lethal autonomous weapon systems, deterrence and stability.* Journal of Strategic Studies, 2019. **42**(6): p. 764-788.
99. Gabriel, I., *Artificial intelligence, values, and alignment.* Minds and machines, 2020. **30**(3): p. 411-437.
100. Kurakin, A., I. Goodfellow, and S. Bengio, *Adversarial examples in the physical world.* arXiv preprint arXiv:1607.02533, 2016.
101. OpenAI, *GPT-4 System Card*. 2023: Available at: https://cdn.openai.com/papers/gpt-4-system-card.pdf.
102. Akyürek, E., et al., *What learning algorithm is in-context learning? investigations with linear models.* arXiv preprint arXiv:2211.15661, 2022.
103. Yampolskiy, R.V., *Leakproofing Singularity - Artificial Intelligence Confinement Problem.* Journal of Consciousness Studies (JCS), 2012. **19(1-2)**: p. 194–214.
104. Babcock, J., J. Kramár, and R. Yampolskiy. *The AGI containment problem*. in *International Conference on Artificial General Intelligence*. 2016. Springer.
105. Babcock, J., J. Kramar, and R.V. Yampolskiy, *Guidelines for Artificial Intelligence Containment*, in *Next-Generation Ethics: Engineering a Better Society (Ed.) Ali. E. Abbas*. 2019, Cambridge University Press: Padstow, UK. p. 90-112.

106. Johnson, N., et al., *Financial black swans driven by ultrafast machine ecology.* arXiv preprint arXiv:1202.1448, 2012.
107. Schwarting, M., T. Burton, and R. Yampolskiy. *On the Obfuscation of Image Sensor Fingerprints*. in *Information and Computer Technology (GOCICT), 2015 Annual Global Online Conference on*. 2015. IEEE.
108. Christiano, P., *Inaccessible Information*. June 3, 2020: Available at: https://www.alignmentforum.org/posts/ZyWyAJbedvEgRT2uF/inaccessible-information.
109. Anonymous, *Cognitive steganography*, in *Arbital*. Retrieved April 27, 2023: Available at: https://arbital.com/p/cognitive_steganography/.
110. Turchin, A., *Catching Treacherous Turn: A Model of the Multilevel AI Boxing*. 2021: Available at: https://www.researchgate.net/profile/Alexey-Turchin/publication/352569372_Catching_Treacherous_Turn_A_Model_of_the_Multilevel_AI_Boxing.
111. Bowman, S.R., et al., *Measuring progress on scalable oversight for large language models.* arXiv preprint arXiv:2211.03540, 2022.
112. Scott, P.J. and R.V. Yampolskiy, *Classification schemas for artificial intelligence failures.* Delphi, 2019. **2**: p. 186.
113. Trazzi, M. and R.V. Yampolskiy, *Artificial Stupidity: Data We Need to Make Machines Our Equals.* Patterns, 2020. **1**(2): p. 100021.
114. Chen, F. and G. Roşu, *Towards monitoring-oriented programming: A paradigm combining specification and implementation.* Electronic Notes in Theoretical Computer Science, 2003. **89**(2): p. 108-127.
115. Christiano, P., A. Cotra, and M. Xu, *Eliciting latent knowledge: How to tell if your eyes deceive you*. 2021.
116. Burns, C., et al., *Discovering latent knowledge in language models without supervision.* arXiv preprint arXiv:2212.03827, 2022.
117. Belrose, N., et al., *Eliciting Latent Predictions from Transformers with the Tuned Lens.* arXiv preprint arXiv:2303.08112, 2023.
118. Ramamoorthy, A. and R. Yampolskiy, *Beyond Mad?: The Race for Artificial General Intelligence.* ITU Journal: ICT Discoveries, 2017.
119. Critch, A. and D. Krueger, *AI research considerations for human existential safety (ARCHES).* arXiv preprint arXiv:2006.04948, 2020.
120. Brcic, M. and R.V. Yampolskiy, *Impossibility Results in AI: a survey.* ACM Computing Surveys, 2023.
121. Yampolskiy, R., *The Singularity May Be Near.* Information, 2018. **9**(8): p. 190.
122. Yudkowsky, E. and R. Hanson, *The Hanson-Yudkowsky AI-foom debate*, in *MIRI Technical Report*. 2008: Available at: http://intelligence.org/files/AIFoomDebate.pdf.
123. Anderson, P.W., *More is different: broken symmetry and the nature of the hierarchical structure of science.* Science, 1972. **177**(4047): p. 393-396.
124. Yampolskiy, R. *On the Differences between Human and Machine Intelligence*. in *AISafety@ IJCAI*. 2021.