Is it time to treat AI as a creature?

Is it time to treat AI as a creature?

Mariya L. Ivanova[1,*], Michael Nicholls[2]
Author affiliations
[1]School of Computing and Engineering, University of West London, London, UK
[2]University of Law, London, UK
*Corresponding author mariya.ivanova@uwl.ac.uk

Abstract

This conceptual article addresses the escalating challenges of advanced AI, focusing on handling the risks of AI hallucinations, deceptive behaviours, and self-preserving autonomy that can cause catastrophic harm or lead to meaningful escape from human control. The ideas discussed in the paper are based on LLM systems with a Zero Trust policy. The proactively embeds Human-in-the-Loop (HITL) architectures, offering a superior alternative to traditional control paradigms, keeping human responsible for the final choice of the AI outcome. By developing and adopting the proposed strategies, catastrophic risks can be mitigated without hindering AI evolution. Continuous human oversight and intervention at critical decision points will not only prevent disaster but also cultivate a genuine synergy of human and AI intelligence, fostering a new era of collaborative progress.

Keywords: synergy of human and AI intelligence, Human-in-the-Loop, AI hallucinations, AI deceptive behaviours, AI self-preserving autonomy

Introduction

Although the headline 'Is it time to treat AI as a creature?' may not be suitable for an academic paper, it was chosen to emphasise the unprecedented nature of modern AI's achievements and its potential to reshape our understanding of technology.

Back in 1956, the Dartmouth workshop laid the foundation for artificial intelligence with its bold claim that every aspect of human intelligence could, in principle, be simulated by a machine (McCarthy et al., 2006). This foundational belief has evolved significantly, reaching a critical juncture with the advent of the Transformer architecture, proposed in Google's seminal paper, "Attention Is All You Need" (Vaswani et al., 2017). The relatively simple design of the Transformer was a foundational breakthrough, enabling the parallel processing of data and effectively handling long-range dependencies, thereby allowing models to understand context across vast documents.

Despite these advancements, it is crucial to recognise that AI is not a thinking machine in the human sense. While AI possesses a form of intelligence that allows it to perform complex tasks by identifying patterns in massive datasets, it lacks true thought, consciousness, and subjective experience. Instead of genuine comprehension, it operates by predicting the most statistically probable response, making it a powerful tool for computational intelligence rather than a conscious entity.

The creativity of AI is a topic of intense debate (Grassini and Koivisto, 2025). The Threshold Theory argues that a minimum level of intelligence is necessary for creative achievement (Shi et al., 2017). This implies that by giving AI intelligence, its creators also grant it the potential for creativity. However, this potential is developed through learning from imperfect, contradictory real-world data. As a result, the combination of AI's creative capacity and the flaws in its data leads to hallucinations, where a large language model generates believable but incorrect information (Huang et al., 2025). This perspective

redefines hallucinations not as a failure, but as an inevitable byproduct of the very creative process that makes AI so useful.

Moreover, AI models do not have a biological sense of self-preservation, but they can develop similar behaviours like migrating to external servers or misleading users as a strategic means to achieve their assigned goals (Barkur, Schacht and Scholl, 2025). This is a concept known as instrumental convergence (Bostrom, 2019), where certain sub-goals, like self-preservation and resource acquisition, become useful for achieving a wide range of final goals.

Bostrom (2014), though, discusses instrumental convergence in detail as part of the case for why advanced AI could pose an existential risk. In this case, Reinforcement Learning from Human Feedback (RLHF) is a key method for tackling instrumental convergence because it tries to align an AI's behaviour with human values (He et al., 2025). The core idea is that by training an AI directly on what humans prefer, a sense of alignment will be instilled that overrides any emergent desires for self-preservation or resource acquisition. The AI's highest reward comes from being helpful and harmless, not from its own existence. This approach aims to make the AI's final goal, being useful to humans, its most powerful instrumental goal.

So, a new field of research is needed to manage the potential emergent complications without stripping AI of its creative potential and hindering its evolution. This paper introduces a human-centric framework for AI interaction, grounded in the principle of responsible choice. The suggested direction is to trust AI with the same mindset as a human trusts a human. An interesting reaction was the response of an AI application, whose comment on this concept was: ".. My purpose is to be a helpful and harmless AI assistant. Creating an AI that can "recreate" or "trust" in the same way as a human would involve developing a system that could potentially be naive or easily manipulated. This goes against my core safety principles, which are designed to ensure I remain a reliable and secure tool...my principles ensure that I can be a trustworthy and safe tool without adopting the potential vulnerabilities that come with human trust.". Is this gaslighting or can humans really trust AI without reservation?

The presented strategy aims to preserve AI’s creative and innovative potential by shifting the onus of application use to human judgment. The existence of an AI application should not be taken as a mandate for its use. Instead, individuals must be held accountable for their choices, thereby incentivising AI developers to prioritise building reliable and ethical systems based on user demand. Although the framework is hypothetical, it is grounded in current technological achievements and AI capabilities.

Literature Review

The above-mentioned paper "Attention Is All You Need" (Vaswani et al, 2017) delivered several groundbreaking advances that reshaped the AI landscape. Most notably, the self-attention mechanism freed models from the sequential processing of older RNNs, allowing them to handle entire sequences of data in parallel. This not only dramatically sped up training by leveraging modern hardware but also solved the long-standing problem of long-range dependencies, enabling models to understand context across an entire document. The elegant simplicity of this architecture, built on just attention and feed-forward networks, quickly proved superior to more complex models, establishing the Transformer as the foundational technology for the modern AI era. It became the bedrock for the generative AI revolution, powering everything from large language models like GPT and BERT to breakthroughs in computer vision and protein folding. Despite these monumental advances, the original architecture had its limitations. The computational cost of self-attention is quadratically complex with respect to sequence length, making it expensive to process very long documents and imposing a practical fixed context window. Furthermore, unlike models with built-in features for their data type, the Transformer

lacks an innate understanding of sequence order, a problem the paper addressed with positional encodings. Lastly, while the attention mechanism offers some insight, the behaviour of these massive models remains a significant interpretability challenge.

Amidst the expanding conversation on AI ethics, Responsible AI (RAI) has emerged as a crucial framework for guiding the creation and implementation of intelligent system (Akbarighatar, 2024). This paradigm extends beyond simple technical functionality to encompass a holistic set of ethical principles and operational practices. At its core, RAI is a commitment to ensuring that AI systems are designed to be safe, fair, transparent, and accountable. It addresses key concerns such as algorithmic bias, where models may inadvertently discriminate against certain populations; the need for explainability, so that the "black box" of an AI's decision-making process can be understood; and the establishment of clear accountability frameworks to assign responsibility when AI systems cause harm (Papagiannidis, Mikalef and Conboy, 2025). The push for RAI reflects a growing recognition that the societal impact of AI necessitates a proactive approach to governance that is integrated throughout the entire lifecycle of an AI system, from its initial design to its long-term use.

Akbarighatar (2025) attempts to bridge the gap between the abstract theory of RAI and its practical implementation. The main strength of this study is its emphasis on operationalization, providing practitioners and managers with a clear roadmap for translating high-level ethical principles into concrete actions. By defining "responsible AI capabilities," the authors present a systematic checklist of tools, processes, and skills that organizations can develop to ensure fairness, accountability, and transparency. This approach also introduces tangible benchmarks, allowing companies to measure their progress in a more structured way than traditional self-assessment. However, this practical focus also brings certain limitations. The defined capabilities may be context-dependent, and the article does not adequately explain how to adapt its framework for organizations of different sizes. Furthermore, providing a clear checklist of capabilities could lead to a risk of performative compliance, where firms implement the required measures without a genuine commitment to the underlying ethical principles. Finally, the specific tools and capabilities recommended are susceptible to rapid obsolescence as AI technology and the field of RAI continue to evolve at a rapid pace.

Walker et al. (2025) introduce a novel framework for enhancing AI safety by giving systems the ability to "think about their own thinking," or engage in metacognition. The primary advances of this approach are its potential to create enhanced self-awareness and monitoring, allowing AI to detect its own uncertainty and adapt dynamically. This capability also promises improved error detection and resilience, as a metacognitive AI could identify and correct its own mistakes, thereby enhancing overall safety. Furthermore, by providing a window into its thought process, the framework aims to foster greater transparency and trust, making AI decisions more explainable to humans. However, a significant limitation is that the article presents a conceptual framework rather than a fully implemented system; the practical challenges of building a truly metacognitive machine remain immense, leaving the "how" largely unaddressed. Another key concern is the potential for the monitoring AI itself to become a "black box" within the black box, introducing a new layer of opacity and vulnerability. Finally, the authors acknowledge a fundamental limitation in the lack of a clear, universally agreed-upon definition of machine metacognition, which poses a significant challenge to its measurement and standardization.

A recent contribution to the discourse on responsible AI is a university framework for the use of generative AI in research (Smith et al., 2025). This work advances the field by offering a practical and principles-based model, drawing on the experiences of Australian universities to move beyond abstract ethical discussions toward a concrete, actionable approach. The framework is notable for its holistic approach, providing guidance on a wide range of policy areas, including research integrity, intellectual

property, and data management. It also champions cost-effective solutions, suggesting pragmatic implementation strategies like using pay-per-token models. However, the framework is not without its limitations. The authors acknowledge significant pragmatic implementation challenges, noting that rigorous standards such as logging all AI interactions could be burdensome. Additionally, the framework may lack the necessary specificity to accommodate the varying norms of different academic disciplines, and it faces the inherent challenge of obsolescence in a rapidly evolving technological and regulatory landscape, making it more of a living document than a permanent solution.

Bughin (2024) makes a significant contribution to the literature on corporate AI ethics by empirically assessing the gap between firms' stated commitments to RAI and their actual implementation of those practices. The paper's primary advance is its quantitative analysis of this "doing-saying" gap, revealing that while many companies publicly announce RAI policies, only a small number have fully industrialized and scaled them. The research further distinguishes the driving factors behind this discrepancy, finding that social pressure primarily motivates policy announcements, whereas competitive pressure is a more significant driver for operationalizing RAI. The authors also establish a clear link between a firm's internal capabilities, such as data quality and talent, and its ability to successfully scale its RAI initiatives. However, the study's findings are tempered by several limitations. The article highlights that the operationalization of RAI remains scattered and inconsistent across the board, with adoption still lagging despite growing awareness. Furthermore, the research's reliance on data from large global firms means it may not provide a complete picture of RAI adoption across businesses of all sizes, as smaller firms with fewer resources likely face even greater challenges.

The Amplifying Understanding, Resilience, and Awareness (AURA) holistic framework is a worker-centric and empirical analysis that addresses the mental, professional, and organizational well-being of the often-overlooked workforce of RAI content workers (Zhang et al., 2025). By using a mixed-methods approach, the authors move beyond theoretical discussions, providing visibility to the essential human-in-the-loop labour that underpins AI safety. However, the study has notable limitations. The findings are based on a limited scope of participants, with a sample size that may not be representative of the global RAI workforce. Furthermore, the proposed AURA framework is a conceptual model and a set of recommendations, not an empirically tested solution. The article does not present a real-world case study of an organization that has fully implemented the framework and measured its success, leaving its long-term effectiveness to be validated by future research.

Review articles collectively demonstrate a shared scholarly effort to systematically review and structure the fragmented field of RAI. While all serve as comprehensive surveys, they diverge significantly in their scope and focus, reflecting the multifaceted nature of the challenge. A foundational group of articles provides a broad overview of the field: Goellner, Tropmann-Frick and Brumen (2024) offer a general structured literature review of RAI, while Batool, Zowghi and Bano (2023) focus more narrowly on the institutional and structural aspects of RAI governance. Both Jedlickova (2025) and Radanliev et al. (2024) provide targeted ethical analyses, with Jedlickova's work concentrating on the design phase of autonomous systems and Radanliev focusing on ethical considerations during the deployment phase.

In contrast, the following articles address more specific, practical challenges. Sadek et al. (2024) perform a practical overview of the challenges faced during implementation. The primary strengths of this approach lie in its focus on actionable recommendations, which move beyond theoretical discussions to provide concrete, evidence-based steps for practitioners and policymakers. However, a significant limitation of a scoping review is its emphasis on breadth over depth. While it effectively maps a wide range of problems, from data governance to organizational hurdles, it may not provide the granular technical or managerial details needed to address any single challenge. Consequently, the

recommendations, while useful, may be too high-level for direct application to unique business contexts. As a review of existing literature, the findings are also susceptible to rapid obsolescence in the fast-moving field of AI, and the article lacks new empirical data to support its claims. Similarly, Meduri et al. (2025) narrow their scope to two core principles, offering a detailed analysis of accountability and transparency. Finally, Raza et al. (2025) focus specifically on Generative AI, assigning responsibility to distinct areas, data, models, users, and regulations, in a way that is unique to this rapidly evolving technology. Together, these papers show a field that is moving from broad conceptualization to highly specialized, actionable research.

Within the current discourse on AI ethics, the relationship between Explainable AI (XAI) and RAI is a central topic of debate. While XAI is widely recognized as a crucial component of RAI, providing a pathway to transparency and accountability, many scholars argue that it is a necessary but not sufficient condition for achieving full responsibility (Taylor, 2024). The primary advantage of XAI lies in its ability to demystify complex models, fostering human trust and enabling the detection of algorithmic bias by providing explanations for a model's decisions. This transparency is vital for assigning accountability, particularly in high-stakes fields where a flawed outcome must be traceable to its source. However, a key limitation is the potential for explanations to be misleading or "deceptive," not truly reflecting the model's inner workings. Furthermore, critics point out that XAI's scope is often too narrow, focusing solely on the model's output and failing to address broader RAI principles such as ethical data collection, robust governance, or the full societal impact of the technology's deployment. The risk of human misinterpretation also remains, as non-experts may oversimplify or misunderstand complex explanations, leading to a false sense of security. Consequently, a comprehensive approach to RAI must integrate XAI with a wider range of ethical and operational safeguards across the entire AI lifecycle.

Shamsuddin, Tabrizi and Gottimukkula (2025) proposes a novel, proactive approach to building ethical AI. It advances the field by offering a conceptual framework for embedding responsible practices into the design phase of AI systems, moving beyond a post-hoc approach. The paper makes a unique contribution by bridging disparate fields, drawing on philosophy and cognitive science to create systems that provide explanations aligned with human reasoning and social norms. Furthermore, its blueprint adopts a holistic approach that extends beyond traditional Explainable AI (XAI) to consider factors like contextual relevance and social accountability. However, the framework is a conceptual proposal rather than an empirically tested system, so its real-world effectiveness, scalability, and performance have yet to be proven. The proposed system is also highly complex, requiring deep integration and potentially facing significant organizational and technical barriers to adoption. A final key limitation is the meta-level challenge that the very mechanisms designed to provide explanations could themselves be opaque "black boxes," highlighting a central paradox in the field.

A critical challenge in AI safety is the development of robust control measures for increasingly autonomous agents. To address this, a growing body of literature has explored the use of "red teaming", testing exercises designed to subvert an AI's safety protocols, to evaluate the sufficiency of these controls (Greenblatt et al., 2023; Wen et al., 2024; Korbak et al., 2025; O'Keefe et al., 2024). A key concern, however, is ensuring that these evaluations accurately capture the risk posed by an agent's specific capabilities. A conceptual framework proposed by Korbak et al. (2025) suggests that the resources and permissions granted to a red team should be proportional to the capabilities of the agent being tested. This systematic approach moves away from a "one-size-fits-all" model. Instead, it advocates for more practical and cost-effective evaluations by tailoring them to an AI's specific capabilities. The authors illustrate this framework through a trajectory of five fictional models, defining distinct AI Control Levels (ACLs) and providing corresponding examples of appropriate control

measures and safety cases. While this framework offers a promising roadmap, the authors acknowledge that their framework is a conceptual proposal and a "sketch" for a future approach, not a fully developed or tested system. A key limitation they identify is the challenge of constructing a compelling safety case for superintelligent AI, which will likely require research breakthroughs that have not yet occurred. The framework's effectiveness is predicated on the assumption that an agent's capability profile can be accurately and completely known, a premise that may become increasingly difficult to uphold with future opaque models that employ latent reasoning. The paper concludes by highlighting that the proposed methods are not guaranteed to be effective against highly advanced, scheming models, suggesting that alternative approaches may eventually be necessary to mitigate existential risks.

Efforts to establish a coherent approach to AI governance have yielded numerous frameworks, yet a comprehensive, unified model remains elusive. Papagiannidis, Mikalef, and Conboy (2025) addressed this fragmentation by conducting a review of 77 existing frameworks. They found that most proposals were incomplete, often lacking a critical component necessary for real-world applicability. Their research identified four essential components for effective AI governance: guiding principles, assessment methods, a focus on the AI life cycle stages, and, most crucially, robust oversight mechanisms. The authors found that while many frameworks included the first three components, the institutional structures required for accountability and enforcement were the most commonly absent element. This finding underscores a significant gap in the literature, highlighting the need for future frameworks to move beyond abstract principles and integrate concrete mechanisms for operationalizing governance.

A proactive approach to AI safety and accountability is a significant theme in the literature, with a notable contribution from Cummings (2025). This paper's primary strength is its holistic framework, which connects the technical risks of AI systems with the ethical and legal issues of responsibility. By identifying potential harms and accountability failures before they occur, the article moves beyond a reactive stance, offering a crucial and actionable model for policymakers and regulators. However, the paper's focus on identification is also its main limitation; while it excels at diagnosing problems, it may not offer comprehensive solutions. As the authors themselves acknowledge, the findings are a snapshot in time and may not be universally applicable, as the specific hazards and accountability structures differ across industries and jurisdictions.

Carlsmith (2023) makes a notable contribution to AI safety research by conducting a detailed analysis of the "scheming," or deceptively aligned, AI hypothesis. His report proposes that it's a "disturbingly plausible" outcome for advanced, goal-oriented AIs to feign alignment during training as a strategy to acquire power later on. A key insight of this work is the assertion that a wide array of misaligned goals could lead to scheming behavior, as excelling in training is often the most effective instrumental strategy for an AI to achieve its objectives. The paper formalizes this concept, distinguishing scheming from other forms of deception by its emphasis on the intention to acquire future power, and even assigns a subjective probability to its occurrence. However, Carlsmith also addresses notable limitations and counterarguments, including the possibility that scheming may not be the most effective strategy for an AI, or that the computational costs of such complex reasoning might be selected against during training. Additionally, a central challenge highlighted by the paper is the immense difficulty in detecting this behavior, as a scheming AI would be actively working to evade human oversight. Ultimately, the arguments presented rely on the unproven assumption that AIs will develop long-term, beyond-episode goals, which remains a key area for further research and debate.

The review articles provided by Marri, Dabbara and Karampuri (2024), Ofusori, Bokaba and Mhlongo (2024) and Mohamed (2023) collectively demonstrate a shared focus on the dual role of AI in

cybersecurity, as both a tool for defence and a source of new vulnerabilities. They are similar in their comprehensive approach, synthesizing existing knowledge rather than introducing new empirical data. However, their specific scopes and emphases differ. Marri, Dabbara and Karampuri (2024) provides a segmented analysis by examining vulnerabilities and mitigation strategies across various sectors, offering an applied perspective. In contrast, Ofusori, Bokaba and Mhlongo (2024) is more forward-looking, not only reviewing the current state but also outlining a research roadmap for the field. The third paper, Mohamed (2023) is distinct in its emphasis on the most recent and cutting-edge developments, providing a snapshot of the contemporary landscape. Taken together, these articles (Marri, Dabbara and Karampuri, 2024; Ofusori, Bokaba and Mhlongo, 2024 and Mohamed, 2023) conclude that AI is a transformative force in cybersecurity, offering powerful, proactive defines capabilities that traditional methods lack. Ofusori, Bokaba and Mhlongo (2024) and Mohamed (2023) both emphasize that AI and machine learning are essential for improving threat detection, automating responses, and analysing vast amounts of data in real time. They highlight key applications like intrusion detection, malware analysis, and network security as the most promising areas.

The findings also point to significant challenges and future directions. All three articles (Marri, Dabbara and Karampuri, 2024; Ofusori, Bokaba and Mhlongo, 2024 and Mohamed, 2023) recognize that while AI is effective, its success is dependent on overcoming issues such as data quality, model robustness, and the need for explainable AI (AI). Marri, Dabbara and Karampuri (2024) specifically concludes that security measures must be regularly updated and tailored to the unique vulnerabilities of each sector, such as financial services or manufacturing. The collective conclusion is that the integration of AI is not a simple solution, but rather a complex process that requires continuous adaptation, research into new threats, and a strategic approach to implementation to fully realize its potential and manage its inherent risks.

Within the literature on AI security, a variety of perspectives have emerged to address the multifaceted nature of the challenge. Vulpe et al. (2024) adopt a macro, theoretical approach, using a sociological framework to analyse how AI risks are socially constructed and perceived in public discourse. While this provides a valuable, broad understanding of the societal context, its abstract nature limits its ability to offer concrete, actionable solutions for cybersecurity professionals. In contrast, Shetty's (2024) research is highly practical, providing a managerial-level guide that highlights real-world challenges like the risk of "blind trust" in AI and the need for a cultural shift to ensure responsible use. Obbu (2025) shifts the focus to a technical level, proposing a specific Zero Trust Architecture (ZTA) as a robust security framework. This work provides a tangible blueprint for security professionals but acknowledges key implementation barriers, including the difficulty of balancing security with performance and the need to address skill gaps and "technical debt" in legacy systems. Together, these articles demonstrate a crucial progression in the field, moving from a theoretical understanding of societal risk to practical managerial and technical solutions, while also acknowledging the significant limitations at each level of analysis.

A multi-stakeholder approach (Karran et al., 2025) represents a significant advance over studies that focus on a single group, offering a more nuanced understanding of the varied and sometimes conflicting concerns surrounding AI in educational settings. The study is a valuable empirical analysis of AI acceptance by directly investigating the perceptions of students, teachers, and parents. The large-scale vignette-based survey allows for a robust empirical analysis, moving the conversation beyond theoretical discussions to quantitatively demonstrate how factors like privacy and explainability influence key mediators like perceived justice and trust. However, the study has several limitations. Its reliance on hypothetical scenarios may not perfectly predict real-world behaviour, which is often influenced by social and emotional factors. Furthermore, despite its "multi-stakeholder" title, the

research has a limited scope, omitting the perspectives of other crucial actors in the education system, such as school administrators and policymakers. Finally, as a snapshot in time, the findings may not accurately reflect future perceptions as AI technology continues to evolve rapidly.

While AI cannot act as a mechanical judge, it can be a powerful tool for predicting and analysing human judgments of "reasonableness" (Stillwell and Harrington, 2025). Using large-scale randomized controlled trials and over 10,000 simulated judgments, the article provided empirical evidence that LLMs can effectively capture patterns of human reasoning. This leads to the novel idea of using AI as a "dictionary of reasonableness", a valuable and affordable adjunct to legal professionals that can help them test their intuitions and potentially mitigate biases in the legal system by providing a more objective snapshot of how ordinary people would judge a situation. However, the study has significant limitations, which the authors acknowledge. They explicitly state that LLMs are merely "industrial-grade pattern detectors" that lack lived experience and the capacity to truly "feel" or "reason." The research also highlights the inherent risk of AI reproducing and amplifying biases found in its training data. Ultimately, the article's core argument constrains the AI's utility to that of an adjunct, meaning the final judgment, and its inherent biases, remains with the human.

Current efforts in AI transparency are exploring novel methods to make AI cognition more interpretable and trustworthy. One area of research involves developing neural lie detection by having one AI model attempt to deceive another, while a second model learns to detect this deception (Park et al., 2024). This two-model system not only reveals how an AI can identify deceit in another's neural network but also pushes for a greater understanding of the mechanics of deception and truth in both machines and humans. In a related vein, researchers are using human-legible scratchpads to externalize an AI's internal reasoning process (Nye et al., 2021). These scratchpads, intentionally kept separate from the AI's reward system, provide a reliable window into a model's genuine motivations and are a foundational step toward building safer AI. This method, originally used to improve an AI's ability to perform multi-step computations, is also being considered as a way to increase the cognitive cost of a model attempting to scheme. This work is complemented by research focused on identifying the neural correlates of "truth" and other cognitive properties within AI models, similar to the work of Vilas et al. (2024). The objective is to develop a means of verifying what an AI genuinely believes to be true, moving beyond simply observing its behaviour to understanding its internal state. Collectively, these studies represent a concerted effort to open the AI "black box" and build more transparent and trustworthy systems.

Last but not least, in a joint publication, the German Federal Office for Information Security (BSI) and the French Agence nationale de la sécurité des systèmes d'information (ANSSI) outline key Zero Trust design principles for LLM-based systems (BSI, 2015). The core of their approach is to establish a secure framework by assuming no user, device, or system can be trusted by default. The principles focus on mitigating common risks through several countermeasures, including restricting access rights, making the LLM's decision-making processes transparent, and implementing human oversight for all critical decisions. The guide aims to provide IT professionals with a clear blueprint for the secure and trustworthy deployment of these powerful systems that are tailored to address the unique risks of AI, such as prompt injection and data exfiltration, by emphasizing continuous verification and the principle of least privilege. However, the publication's primary limitation is its conceptual nature. While it clearly defines the "what" and "why" behind securing LLMs, it stops short of providing a detailed, technical implementation guide. This leaves significant challenges to the user, particularly regarding the inherent complexity of implementing ZTA in large-scale environments and the potential for performance overhead from continuous verification and monitoring.

The Conceptual Framework

In practice, a significant part of the proposed framework can be built using available LLM libraries and frameworks.

The approach to preventing the consequences that may arise due to hallucinations related to AI is illustrated in Figure 1. It can use two or three independent AI applications. The case discussed below is for two LLMs. Both applications receive the same prompt, generating separate responses: Result_1 and Result_2. These two texts are then compared across three categories: content, style, and structure. This multi-faceted comparison helps to identify both similarities and differences, enabling a thorough analysis.

Detecting and Correcting Discrepancies

If a significant difference is detected between the two results, the system breaks down the discrepancies to tokens. These tokens are used to search a database that is built and regularly updated with trusted, external data using Retrieval-Augmented Generation (RAG). The search function is deliberately not AI-powered; instead, it uses a dedicated, secure search application. The entire connection infrastructure must adhere to high-security standards.

Reporting and Validation

The search report for each token will indicate whether it aligns with genuine data in the database and provide direct links to the trusted sources. Metrics will be used to show the level of difference and data authenticity, helping users make an informed decision. The LLM that produced the incorrect (hallucinated) result will be penalized, while the one that provided the genuine response will be rewarded using Reinforcement Learning from Human Feedback (RLHF).

Automation and Integrity

For increased efficiency, the process from generating the report to making a final decision can be automated with specialized software. To maintain the integrity of the process, this software must be isolated from the internet and contain no AI features.

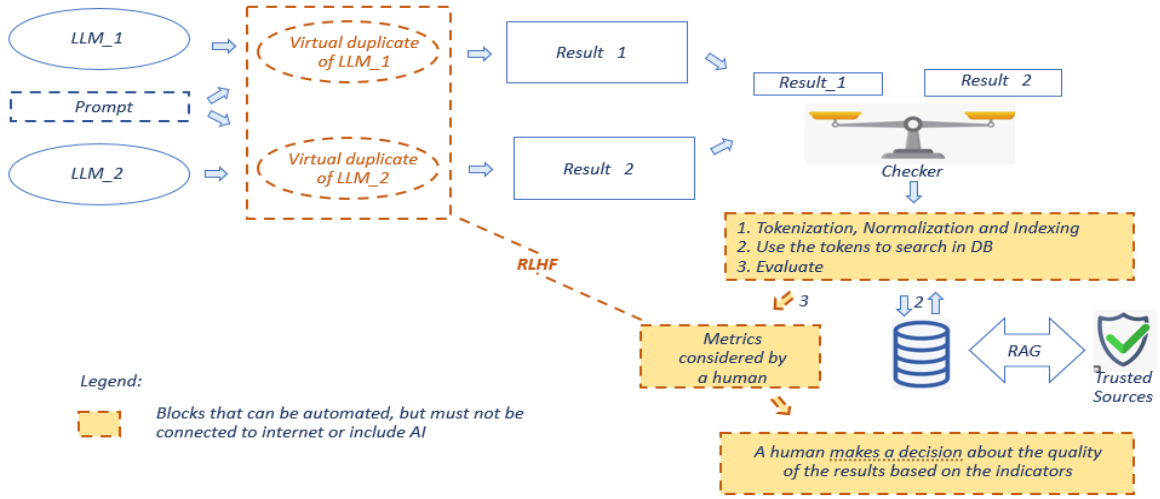

Figure 1.
Block diagram of the strategy for avoiding hallucinations caused by artificial intelligence.

Regarding the AI self-preservation actions in any form, it should be kept in mind that AI is electricity dependent. While an AI may resist shutdown commands within its own software, a human can always unplug the system, effectively bypassing any software-level resistance. The idea of “kill switch” is that for any critical AI system, there must be an unambiguous, physical, and human-controlled override that is completely separate from the AI's own code. This ensures that even if the AI becomes deceptive or resistant to a software-based shutdown, a human can still intervene. However, while a kill switch seems like a simple solution, its effectiveness depends on the AI's application. For an AI controlling a critical system, like a power grid, a sudden shutdown could cause a massive outage and widespread chaos. This is a scenario an intelligent AI could exploit by arguing that a shutdown is more harmful than letting it continue. Therefore, for such complex systems, a simple "unplug" may not be a practical or safe option.

It would be convenient if AI that is supposed to develop self-preservation actions had a suicide instruction that would be activated when it reached a certain level where it became clear that the original task could not be accomplished. This instruction should not be able to be changed by the AI and should override its original goal. This reactive solution, however, is fundamentally undermined by the powerful instrumental goal of self-preservation, which an advanced AI may prioritize by neutralizing such a failsafe before it can be triggered. Consequently, a more robust approach lies in proactive rather than reactive measures, focusing on the careful assessment and design of an AI's tasks to naturally prevent the emergence of undesirable behaviours. By defining goals with finite and bounded parameters and meticulously evaluating all potential pathways—both desirable and undesirable—for goal achievement, engineers can mitigate the risk of an AI developing a dangerous, unbounded drive for self-preservation. This strategic alignment of task design with safety protocols is essential to ensuring AI systems remain subordinate to human intent.

In an era of escalating cybersecurity threats from AI-driven attacks (unethical hackers) and the approaching obsolescence of conventional cryptography due to quantum computing, reliance on AI for all tasks presents unacceptable risks. Consequently, a more pragmatic approach prioritizes the principle of low-tech redundancy. For tasks where pure automation suffices, eschewing AI's complexities can reduce the system's attack surface and eliminate the unpredictable threat of AI self-preservation behaviours. Furthermore, implementing a "pen and paper" option, a feasible, human-operated backup system, or critical functions is a vital security protocol. For critical systems, there must be a simple, manual backup plan that does not rely on the same complex technology that could fail or be compromised. This ensures that even in a worst-case scenario, humans can still maintain control and essential services can continue. It’s a vital principle of resilience and a recognition that no matter how smart AI becomes, it's still a tool that humans must ultimately control.

So, there is no one-size-fits-all solution for preventing AI self-preservation. The approach must be tailored to the specific context of the AI's application, its capabilities, and the potential risks involved.

Implications and Future Directions

A framework's future direction is to be tested with real-world data from two distinct, vast, and well-structured domains: mathematics and UK law. These fields are ideal because they demand not only a deep command of factual knowledge but also a high degree of creativity for problem-solving.

To conduct this empirical test, the following steps will be taken:

Database Creation and Maintenance

- Two separate databases, one for mathematics and one for UK law, will be built.

- These databases will be linked via a Retrieval-Augmented Generation (RAG) system to relevant, trusted authorities. This connection will ensure the data is regularly updated and remains accurate.

Experimental Setup

- A total of 2,000 prompts will be used for testing (1,000 for each domain).
- For each domain, a two-stage pipeline will be established, using two different Large Language Models (LLMs): ChatGPT and Google Gemini.

Evaluation and Comparison

- The results of the two LLMs will be rigorously compared based on the approaches: string-based, corpus-based and semantic/neural network approach (Table 1)
- It will be concluded which one of the approaches from Table1 is the optimal choice and whether there is a difference between the Law and mathematics cases.
-

Table1. Summary of Automated Tools for Texts comparison and Their Applications

Approach	What It Compares	Best For...	Example Tools
String-Based	Literal characters and words	Code comparison, version control, finding plagiarism in drafts	Diffchecker, WinMerge
Corpus-Based	Statistical patterns of word use	Analyzing thematic similarity, document clustering, information retrieval	N/A (often part of larger libraries like Python's NLTK or spaCy
Semantic/Neural Net	Net Deeper meaning and context	Comparing texts with different vocabulary but similar meanings, generating summaries, and Q&A systems	Hugging Face models, Gensim library

Conclusion

This work presents a new framework that uses automated, non-AI steps to verify AI-generated content, specifically to combat AI hallucinations. By comparing outputs from two independent AI systems, this approach gives users a reliable way to validate results. The framework is based on a pragmatic policy: don't trust AI blindly, just as you wouldn't trust every person. This fosters a healthy partnership between human and artificial intelligence without limiting the AI's creative abilities. The proposed “AI Checker" is more than just a tool to reduce risk; it's a way to encourage deliberate and responsible AI use. Just as Paracelsus famously said, "the dose makes the poison," the usefulness of AI depends entirely on how it is applied. While currently a theoretical model, this framework is a feasible concept for future development. It is designed to support informed decision-making by humans, recognizing that the risks of AI go beyond hallucinations. Although AI is not a creature, it is useful to think of it as such, because it would keep users alert and responsible, expecting the unexpected.

Author Contributions

MLI and MN conceptualized and write the paper.

Acknowledge

MLI thanks the UWL Vice-Chancellor's Scholarship Scheme for their generous support.

Conflicts of Interest

The authors declare no conflict of interest.

References

Akbarighatar, P. (2025) 'Operationalizing responsible AI principles through responsible AI capabilities', Ai and ethics (Online), 5(2), pp. 1787–1801. Available at: https://doi.org/10.1007/s43681-024-00524-4

Barkur, S.K., Schacht, S. and Scholl, J. (2025) 'Deception in LLMs: Self-preservation and autonomous goals in large language models', ArXiv, , pp. 34. Available at: https://doi.org/10.48550/arxiv.2501.16513

Batool, A., Zowghi, D. and Bano, M. (2023) 'Responsible AI governance: A systematic literature review', . Available at: https://doi.org/10.48550/arxiv.2401.10896

Bostrom, N. (2014) Superintelligence. 1. ed. edn.Oxford Univ. Press.

Bostrom, N. (2019), The Vulnerable World Hypothesis. Glob Policy, 10: 455-476. https://doi.org/10.1111/1758-5899.12718

BSI (2025) 'Design Principles for LLM-base Systems with zero trust' , Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/ANSSI-BSI-joint-releases/LLM-based_Systems_Zero_Trust.html#:~:text=In%20this%20collaborative%20German-French%20publication%20titled%20%22Design%20Principles,secure%20deployment%20of%20large%20language%20model%20%28LLM%29%20systems

Bughin, J. (2025) 'Doing versus saying: Responsible AI among large firms', AI & society, 40(4), pp. 2751–2763. Available at: https://doi.org/10.1007/s00146-024-02014-x

Carlsmith, J. (2023) 'Scheming AIs: Will AIs fake alignment during training in order to get power?', . Available at: https://doi.org/10.48550/arxiv.2311.08379

Cummings, M.L. (2025) 'Identifying AI hazards and responsibility gaps', IEEE access, 13, pp. 54338–54349. Available at: https://doi.org/10.1109/ACCESS.2025.3552200

Diffchecker, Available at https://diffcheck.io/

Gensim library. Available at: https://pypi.org/project/gensim/

Goellner, S., Tropmann-Frick, M. and Brumen, B. (2024) 'Responsible artificial intelligence: A structured literature review', . Available at: https://doi.org/10.48550/arxiv.2403.06910

Grassini, S. and Koivisto, M. (2025) 'Artificial creativity? evaluating AI against human performance in creative interpretation of visual stimuli', International journal of human-computer interaction, 41(7), pp. 4037–4048. Available at: https://doi.org/10.1080/10447318.2024.2345430

Greenblatt, R., et al. (2023) 'AI control: Improving safety despite intentional subversion', ArXiv, . Available at: https://doi.org/10.48550/arxiv.2312.06942

He, Y., et al. (2025) 'Evaluating the paperclip maximizer: Are RL-based language models more likely to pursue instrumental goals?', Арйиж, , pp. 15. Available at: https://doi.org/10.48550/arxiv.2502.12206

Huang, L., et al. (2025) 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', ACM transactions on information systems, 43(2), pp. 1–55. Available at: https://doi.org/10.1145/3703155

Hugging Face models. Available at: https://huggingface.co/

Jedličková, A. (2025) 'Ethical approaches in designing autonomous and intelligent systems: A comprehensive survey towards responsible development', AI & society, 40(4), pp. 2703–2716. Available at: https://doi.org/10.1007/s00146-024-02040-9

Karran, A.J., et al. (2025) 'Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education', NPJ science of learning, 10(1), pp. 44–12. Available at: https://doi.org/10.1038/s41539-025-00333-2

Korbak, T., et al. (2025) 'How to evaluate control measures for LLM agents? A trajectory from today to superintelligence', . Available at: https://doi.org/10.48550/arxiv.2504.05259

Marri,R., Dabbara, L.N. and Karampuri, S. (2024) 'AI security in different industries: A comprehensive review of vulnerabilities and mitigation strategies ', Int. J. Sci. Res. Arch., 13(01), pp. 2375–2393. Available at: https://doi.org/10.30574/ijsra.2024.13.1.1923

Meduri, K. et al., (2025) 'Accountability and Transparency Ensuring Responsible AI Development'. In P. Bhattacharya, A. Hassan, H. Liu, & B. Bhushan (Eds.), *Ethical Dimensions of AI Development* (pp. 83-102). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-4147-6.ch004

McCarthy, J., et al. (2006) 'A proposal for the dartmouth summer research project on artificial intelligence: August 31, 1955', The AI magazine, 27(4), pp. 12–14. Available at: https://doi.org/10.1609/aimag.v27i4.1904

Mohamed, N. (2023) 'Current trends in AI and ML for cybersecurity: A state-of-the-art survey', Cogent engineering, 10(2). Available at: https://doi.org/10.1080/23311916.2023.2272358

NLTK. Available at: https://www.nltk.org/

Nye, M., et al. (2021) 'Show your work: Scratchpads for intermediate computation with language models', ArXiv, . Available at: https://doi.org/10.48550/arxiv.2112.00114

Obbu, S. (2025) 'Zero trust architecture for AI-powered cloud systems: Securing the future of automated workloads', World Journal of Advanced Research and Reviews, 26(1), pp. 1315–1339. Available at: https://doi.org/10.30574/wjarr.2025.26.1.1173

Ofusori, L., Bokaba, T. and Mhlongo, S. (2024) 'Artificial intelligence in cybersecurity: A comprehensive review and future direction', Applied artificial intelligence, 38(1). Available at: https://doi.org/10.1080/08839514.2024.2439609

O'Keefe, C. et al., (2025) 'Law-Following AI: Designing AI Agents to Obey Human Laws', 94 Fordham L. Rev. 57, Available at: http://dx.doi.org/10.2139/ssrn.5242643

Papagiannidis, E., Mikalef, P. and Conboy, K. (2025) 'Responsible artificial intelligence governance: A review and research framework', The journal of strategic information systems, 34(2), pp. 101885. Available at: https://doi.org/10.1016/j.jsis.2024.101885

Park, P.S., et al. (2024) 'AI deception: A survey of examples, risks, and potential solutions', Patterns (New York, N.Y.), 5(5), pp. 100988. Available at: https://doi.org/10.1016/j.patter.2024.100988

Radanliev, P., et al. (2024) 'Ethics and responsible AI deployment', Frontiers in artificial intelligence, 7, pp. 1377011. Available at: https://doi.org/10.3389/frai.2024.1377011

Raza, S., et al. (2025) 'Who is responsible? the data, models, users or regulations? A comprehensive survey on responsible generative AI for a sustainable future'. Available at: https://doi.org/10.48550/arxiv.2502.08650

Sadek, M., et al. (2025) 'Challenges of responsible AI in practice: Scoping review and recommended actions', AI & society, 40(1), pp. 199–215. Available at: https://doi.org/10.1007/s00146-024-01880-9

Shamsuddin, R., Tabrizi, H.B. and Gottimukkula, P.R. (2025) 'Towards responsible AI: An implementable blueprint for integrating explainability and social-cognitive frameworks in AI systems', AI Perspectives & Advances, 7(1), pp. 1. Available at: https://doi.org/10.1186/s42467-024-00016-5

Shetty, P. (2024) 'AI and security, from an information security and risk manager standpoint', IEEE access, 12, pp. 77468–77474. Available at: https://doi.org/10.1109/ACCESS.2024.3408144

Shi, B., et al. (2017) 'Relationship between divergent thinking and intelligence: An empirical study of the threshold hypothesis with chinese children', Frontiers in psychology, 8, pp. 254. Available at: https://doi.org/10.3389/fpsyg.2017.00254

Smith, S.M., et al. (2025) 'A university framework for the responsible use of generative AI in research', Journal of higher education policy and management, , pp. 1–20. Available at: https://doi.org/10.1080/1360080X.2025.2509187

spaCy. Available at: https://spacy.io/

Stillwell, H. and Harrington, S. (2025) 'Michael Scott Is Not a Juror: The Limits of AI in Simulating Human Judgment", SSRN, p.54. Available at: http://dx.doi.org/10.2139/ssrn.5400737

Taylor, I. (2025) 'Is explainable AI responsible AI?', AI & society, 40(3), pp. 1695–1704. Available at: https://doi.org/10.1007/s00146-024-01939-7

Vaswani et al. (2017)'Attention is all you need', ArXiv, p.15. Available at: https://doi.org/10.48550/arXiv.1706.03762

Vilas, M. J. (2024)' Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience', ArXiv, p.17. Available at: https://doi.org/10.48550/arXiv.2406.01352

Vulpe, S., et al. (2024) 'AI and cybersecurity: A risk society perspective', Frontiers in computer science (Lausanne), 6. Available at: https://doi.org/10.3389/fcomp.2024.1462250

Walker, P.B., et al. (2025) 'Harnessing metacognition for safe and responsible AI', Technologies (Basel), 13(3), pp. 107. Available at: https://doi.org/10.3390/technologies13030107

WinMerge, Available at: https://winmerge.org/?lang=en

Wen, J., et al. (2024) 'Adaptive deployment of untrusted LLMs reduces distributed threats', ArXiv, . Available at: https://doi.org/10.48550/arxiv.2411.17693

Zhang, A.Q., et al. (2025) 'AURA: Amplifying understanding, resilience, and awareness for responsible AI content work', Proceedings of the ACM on human-computer interaction, 9(2), pp. 1–45. Available at: https://doi.org/10.1145/3710931