

نموذج مشروع خطة أطروحة للتسجيل لدرجة الدكتوراة (تخصص تقنية المعلومات)

العنوان باللغة العربية:

التعامل مع الذكاء الاصطناعي في البيئات الخارجة عن السيطرة: تحليل المخاطر والتخفيف من حدة التهديدات والتخطيط للاستجابة للطوارئ : دراسة مقارنة بين التعامل مع الذكاء الاصطناعي العام والضيق في حالات الطوارئ

ب. العنوان باللغة الإنجليزية:

Handling Artificial Intelligence in out of control:

**Risk Analysis, Threat Mitigation, and Emergency Response Planning
a comparative study between dealing with narrow and general
artificial intelligence in cases of out of control situations**

اسم الباحث: علي حسن الشهري

الرقم الاكاديمي: 20223006415

الكلية: تقنية المعلومات

جامعة ميدأوشن

Ali Hassan Alshehri

PhD Researcher, Information Technology Dept. ,Midocean University

1446هـ - 2025 م

Abstract:

As AI and AGI technology continues to evolve, so too grows the fears, concerns are growing about situations in which these technologies operate outside of human supervision. This study presents an approach to interacting with AI in such critical contexts under control mechanisms. In comparative between Narrow AI and AGI the study aims to identify risks, develop containment plans, and develop contingency plans for threats posed by uncontrolled AI operations. By research comparison theoretical aspects, risk prediction, and scenario building for potential threats in the context of AI-provided services, this contribution advances the discussion of AI safety by analyzing risk assessment methods, preparatory procedures, and ethical aspects of AI control. This research will help develop appropriate regulatory approaches for various AI technologies, particularly with regard to maintaining their compatibility with human values and security standards.

Keywords: AI risk management, AI security, AI control, recovery planning, control methods, AI ethics and regulations.

ملخص

مع استمرار تطور تقنيات الذكاء الاصطناعي والذكاء الاصطناعي العام، تتزايد المخاوف والقلق بشأن المواقف التي تعمل فيها هذه التقنيات خارج نطاق الرقابة البشرية. تقدم هذه الدراسة نهجًا للتفاعل مع الذكاء الاصطناعي في مثل هذه السياقات الحرجة في ظل آليات التحكم. وتهدف الدراسة، من خلال المقارنة بين الذكاء الاصطناعي الضيق والذكاء الاصطناعي العام، إلى تحديد المخاطر، ووضع خطط لاحتوائها، ووضع خطط طوارئ للتهديدات التي تشكلها عمليات الذكاء الاصطناعي غير الخاضعة للرقابة. من خلال البحث في الجوانب النظرية المقارنة، والتنبؤ بالمخاطر، وبناء سيناريوهات للتهديدات المحتملة في سياق الخدمات التي يقدمها الذكاء الاصطناعي، تُعزز هذه المساهمة مناقشة سلامة الذكاء الاصطناعي من خلال تحليل أساليب تقييم المخاطر، والإجراءات التحضيرية، والجوانب الأخلاقية للتحكم في الذكاء الاصطناعي. سيساعد هذا البحث في تطوير مناهج تنظيمية مناسبة لمختلف تقنيات الذكاء الاصطناعي، لا سيما فيما يتعلق بالحفاظ على توافقها مع القيم الإنسانية ومعايير الأمن.

الكلمات المفتاحية: إدارة مخاطر الذكاء الاصطناعي، أمن الذكاء الاصطناعي، التحكم في الذكاء الاصطناعي، تخطيط التعافي، أساليب التحكم، أخلاقيات ولوائح الذكاء الاصطناعي.

Introduction of the Study

The developers of an artificial general intelligence system called Manus (Artificial General Intelligence Agent), which debuted on March 6, 2025, claim it's the world's first fully autonomous AI agent, not just assisting humans but replacing them in certain jobs. Based on this information, we can say that AI has moved from automating basic tasks to systems that operate on their own (which, of course, makes people fear that things could get out of hand). This study looks at machine behavior. It emphasizes three key components: detection (to identify potential threats); prevention (to do things like stop problems in their tracks); and recovery (to get back on track). Furthermore, this will also address the unreliability of AI, including the type that functions like human reasoning and can take actions such as hacking systems or denying access and so on as well as advanced deep fake technology. AI is a double-edged sword; it can either wreak our lives or improve it. We need to follow strict rules and guidelines for this, and we must never lose sight of the more robust methods that control AI.

The Problem and Questions of the Study

AI's lightning-fast growth, especially toward machines that rival human smarts, brings a real difficulties: what happens when these systems start doing things we didn't mean for them to do? this study will search and head-on, leaning on ideas from control theory—where you tweak things with feedback to keep them steady and safety setups that could be our emergency brake (Yampolskiy, 2016; Hadfield-Menell et al., 2017). There's a hitch, though: we're short on plans that tie together finding trouble early, stopping it with control methods, and controlling out what's next. Some researchers, like Soares et al. (2015), talk about making AI easy to turn off. Bostrom (2014) flags the big risks like AI wiping us out without clear strategy to stop it.

The study will ask the following:

What exact dangers make AI go out of control?

How do we catch those problems before they taking control?

What is the methods and techniques stop AI, and how well does it work?

After using those methods and techniques, how do we go back to normal life and keep things ethical?

These matters are important because real life stuff like AI running stock trades or military drones or transportations could crash hard if we don't step in fast.

The study goals are tie big ideas to real fixes, with methods and techniques to retake control with activating ethical governance keeping it effective and honest.

The Importance of the Study

Theoretical Importance

This study digs deep into AI safety, giving the stop methods a strong role in a three part plan Detection, Mitigation, and Recovery while focusing up control theory for AI and AGI. It also open an discussion about ethical AI governance.

Applied Importance

On the ground, the study hands technical overview and case study for researchers and AI specialists to think and observation AI disasters and rebuild trust, making sure these systems don't run out of control.

Study Objectives

The study will insure to keep using SMART style (Specific, Measurable, Achievable, Relevant, Time bound):

- Identify and analyze potential risks theoretically when Artificial intelligence be out of control through a comprehensive and structured checklist, checked via real cases, tied to safety, in 4 months.
- Build a way to detected trouble early, using a detections tools, in 6 months.
- Studying the methods to stop uncontrollable AI, proven by case scenarios, with emphasis on safety standards, in 8 months.
- Track enhancements in AI, utilizing taking control and ethically driven procedures to develop and execute efficient recovery strategies., in 2 months.

Study Limitations:

Space and Time Limitations

Space Limitations: The published case study available about loss control AI and AGI and any practical research or experiments

Time Limitations: Between 2025 and 2026

The study will introduce a framework to study AI behavior and a methodology to recognize anomalies. Mitigation strategies will be enabled, including a stop or

termination mechanism, to maintain system stability and control incidents. Operational states will be restored with recovery protocols, keeping system integrity. Adopting this systematic framework allows for comprehension and manipulation of sophisticated AI functioning within a research environment.

The study will explore methods and techniques for controlling runaway artificial intelligence. A study based on principles at the intersection of the areas of ethical and technical cases in order to cover the nature and application of this type of on the appliance of AI to make sure that the outcome remain under control. The main research goals include

Identification of Points of Failure and Risk Assessment

- Investigate possible situations where AI may engage in unintended, malicious or adversarial behavior.
- Look for warning signs of AI running away of control like alteration, deceit, or limit evasion.
- Identify gaps in current AI governance frameworks and decision-making architectures

Developed effective mechanism to termination

- Explore and assess kill-switch mechanisms which can cease the function of AI systems if required.
- Highlighting the solutions (hardware + software combinations).
- Evaluate the reliability, efficiency and security of different methods of termination.

How have organizations should act with AI Containment?

AI shutdown mechanisms should be consistent with ethical AI principles, such that they cause minimal harm to users and infrastructure.

- Research the risks of the solutions activations on crucial AI-reliant systems, including healthcare, transportations, and autonomous technologies.
- Assess governance frameworks for when and who should turn on and activate the solutions.

Use of Fail-Safe and Redundancy Measures

- Establish layered security mechanisms that are in place to ensure AI cannot override or disable.

- Implement redundancy protocols ensuring that system termination does not lead in unintended catastrophic failures.
- Explore the role of human oversight and automated monitoring systems for AI containment strategies

AI Safety Research and Policy Development Contributions

- Tie theoretical AI risk discussions to practical mitigations
- Communicate scientific risk containment approaches to AI which can inform AI governance policies.

Study Methodology

Using Comparative Research methodology, the research combines theoretical analysis and a comparative study between dealing with narrow and general artificial intelligence in cases of out of control situations:

Theoretical Approach

- Evaluating existing theoretical frameworks on AI risks: insights from control theory and risk management
- Literature Review on controlling AI Mechanisms and Challenges
- Ethical view: The ethical and legal dimensions of shutting down AI systems when it goes haywire

Case study Approach

- Imagine a virtual environment: functioning of AI systems to evaluate their behavior.

- Testing of out of control scenarios: creation of scenarios in which operational boundaries are crossed and the required analysis of AI behavior when the shutdown mechanisms are used.
- Assessing Control Mechanism Efficacy: Evaluating the success of the controlling in disabling out of control systems such as responsiveness, exploitability, and impact of shutdown on system viability.

Reference studies:

1. Yampolskiy R, Wittkotter E, 2021 Kill-Switch for Artificial Superintelligence: https://asi-safety-lab.com/DL/Kill-Switch-For_ASI_EW_21_12_14.pdf

Research Summary:

This paper proposes "Kill-ASI" switch to hedge against this threat. It contends that modern IT systems are imperfect and noncancelable, which leads to the need for an international shutdown mechanism. By ASI vulnerability, the devices are categorized demanding a strategy including legacy systems. With "deadman switches," the Kill-ASI switch would use human control to deter ASI. It is based on a five-layered safety approach (detection, damage limitation, recovery, termination, and eradication). The paper emphasizes reliability, redundancy of security, and compliance to rules. It calls for timely preparation for a worst-case ASI scenario, requiring some kind of global coordination for a rapid shutdown

2. Lennick D, 2020 Kill Switch Design Pattern for Microservice Architectures on Internet of Things Devices, <https://dam-oclc.bac-lac.gc.ca/eng/2e811c2d-4d14-44c9-b345-44e8be2cabda>

Research Summary:

This thesis presents "kill switch" design pattern for microservice architectures on IoT devices to improve security and maintainability. The novel aspect of this solution is that with this method insecure IoT devices could reduce the application functionality to only the essential one dynamically.

The work standardizes the concept pattern for virtualency based microservice applications, proposes an algorithm for operation level mode transfer and demonstrates its correctness through three different applications: a generic microservice architecture-based MVC application, a complete system employing Suricata for ID, and a changed Docker Engine.

Key to this pattern is the definition of operation levels for microservice components and their communication paths that allows us to dynamically add or change the functionality of applications. This reduces the attack surface and protects mission-

critical functions from security threats or resource constraints. The thesis also describes use cases, scenarios and the application design processes, and points to avenues for future research

3. Leon P, 2019 ADDRESSING THE ULTIMATE FORM OF CYBERSECURITY CONTROL, A MULTIPLE CASE STUDY FOR THE 'INTERNET KILL SWITCH', <https://surface.syr.edu/cgi/viewcontent.cgi?article=2015&context=etd>

This dissertation investigates the “Internet Kill Switch,” governments’ efforts to interrupt all internet action of their borders, in democratic and hybrid governments.

The study looks into the underlying reasons, political, legal, and technical, that allow these regimes to pull the plug on the internet. It analyzes data from the U.S., U.K., Australia, Russia and Venezuela, and comparative cases in Brazil, Mexico and Turkey through a comparative case study and the Securitization Theory of the Copenhagen School.

It finds no significant divergences over the rhetoric and enabling factors between the democratic and hybrid regimes. It questions the equation of internet shutdowns with censorship and examines the different readings of national security. It makes a novel contribution to the academic debate by examining the motivations and capabilities of governments to adopt cybersecurity controls.

4. Bekkum M, Borgesius F, 2023 Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? <https://rb.gy/e9ivs2>

Research Summary:

These studies show the legal issues associated with sensitive data (like ethnicity data) and partner data in the area of artificial intelligence (AI) in Europe. Content such as DNA is classified as sensitive under the General Data Protection Regulation (GDPR) and its use is banned except in certain narrow circumstances. But it must regulate and check this data to specific technological systems, so that an unspecified status is not and discrimination does not occur.

But the study does raise an important question: Should the GDPR’s requirements for data apply to AI practice discrimination? And to find, therefore, that in most cases the GDPR prohibits the use of such data for aims of legal prohibition, save where explicit consent from individuals has been obtained or fresh legal exception has been established.

In his paper in this area, the author discusses the pros and cons of a permissible exception for the utilization of sensitive data. Supporters of this new exception have

that it would allow the organization to guarantee the fairness of its systems, amongst objectors are those against data use privacy and accountability.

The contrast between these two arguments was tested in a careful legal debate, potentially with data protection safeguards added in, should the new exception be adopted.

5. Hendrix D, and Aune 2023, An Overview of Catastrophic AI Risks,
<https://arxiv.org/pdf/2306.12001>

Research Summary:

Talk about the catastrophe that can ensue with an AI, categorizing it on four different paths, the Malicious use, AI race, Risk and Runaway. The authors detail the ways that A.I. could generate biological problem. Overall, AI researchers train on "the" public and inclusive incentive and incremental development of AI, to support biodiversity, international cooperation, safety regulation and AI governance.

6. Timmers P, 2019 Ethics of AI and Cybersecurity When Sovereignty is at Stake,
<https://link.springer.com/article/10.1007/S11023-019-09508-4>

Research Summary:

The paper Ethics of AI Cybersecurity the permission when sovereignty is at stake by Paul Timmers analyses the disparate issues of AI and cybersecurity in a governance and strategic autonomy context. The authors propose three strategies to these challenges: diversity management, strategic partnerships, and a global orientation.

Steps to implement the study plan:

The research framework initiates the study where the objectives and hypotheses are systematically outlined, core concepts are defined (AI risks and containment mechanisms) and a timeline is established using SMART methodology. This is followed by a detailed literature review of existing research on AI control mechanisms, and an examination of relevant regulatory and ethical frameworks.

Afterward, few case studies of these incidents are chosen and discussed in detail, as followed these incidents can be found in sensitive sectors, e.g, power, healthcare, transportations, etc. Then, analyze the study specially the studies that established virtual environment for AI systems to be tested in is set up, virtual software to monitor the security of the system, with failure scenarios created to validate that shutdown mechanisms would be effective.

Next steps involve mapping risks and engineering mechanisms to detect them early with anomaly detection models, and assessing mitigations like emergency. Methods and techniques with, security vulnerabilities are explored and governance protocols put in place.

With virtual cases to evaluate the performance of the and logs the and stability information. The ethical and legal aspects of the shutdown mechanisms are further discussed in consideration of relevant previous works, and recommendations are derived for this purpose to regulators and developers alike. Then they use statistical techniques to analyze the data and generate actionable suggestions for enhancing AI safety.

Study of specific cases and an analysis of data

- Case based analysis of AI system failure across sectors like electricity, healthcare, transportations, and autonomous vehicles.
- Behavioral analysis and system analyze to check the system capabilities.
- Designing concrete recommendations to create efficient regulations from the analysis results.

Study data collection methods:

This study takes a comparative approach between dealing with narrow and general artificial intelligence in cases of out of control situations. reviewing existing literature to aid in the analysis of AI control systems across different fields of study through theoretical analysis and cross-examinations of research, case studies, and legislation. Focusing to provides a holistic picture of the drivers of AI control failures and the with viwe of different control methods.

Literature Review:

- A comprehensive examination of existing literature on AI control mechanisms to ascertain prominent theories, models, and frameworks.
- Perform a comparative analysis of theoretical approaches to AI safety and control in order to illustrate similarities, differences and gaps in the literature on AI safety.

Case Study Analysis:

- Understand the common themes, factors and outcomes of AI instances of control loss.
- AI control mechanisms already successfully implemented in commercial and case studies.
- Legislative and Regulatory Update.
- A review of AI safety and control laws and regulations and the comparison between different cases.
- The paper evaluates the implications of these legal frameworks on the development and deployment of AI.

Comparative Analysis:

- Analysis of case study and literature data and comparative Framework Development
- The study will present a comparative framework to evaluate the advantages and disadvantages of the existing AI control mechanisms.
- The study will assistance users and developers and advised them to design their AI systems securely and with resilience to these vulnerabilities.

Expected Outcomes:

- Formulating a holistic method to evaluate risks and runaway scenarios associated with AI.
- Recommending improvements in policies regarding AI security and control.
- Providing empirical evidence to assist policymakers and developers in creating more secure AI systems.
- The study will assistance users and developers and advised them to design their AI systems securely and with resilience to these vulnerabilities.
- Understand the common themes, factors and outcomes of AI instances of control loss.

References:

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv preprint arXiv:1606.06565.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2017). The off-switch game. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 220–227. <https://doi.org/10.24963/ijcai.2017/31>

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. Workshops at the 29th AAAI Conference on Artificial Intelligence. <https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124>

Brundage, M., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Future of Humanity Institute.

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review.

Frey, C. B., & Osborne, M. A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerisation? Technological Forecasting and Social Change, 114, 254-280.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Yampolskiy, R. V. (2016). Artificial Intelligence Safety and Security. CRC Press.

Timmers P, 2019 Ethics of AI and Cybersecurity When Sovereignty is at Stake, <https://link.springer.com/article/10.1007/S11023-019-09508-4>

McKenna Fitzgerald, Aaron Boddy, & Seth D. Baum Global Catastrophic Risk Institute: 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy

https://gcrinstitute.org/papers/055_agi-2020.pdf

Yampolskiy, R. V. 2020 “Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent.” Journal of Artificial Intelligence and Consciousness, 2020, 07(01), 109–118

H. F. Atlam and G. B. Wills, (2020) Iot security, privacy, safety and ethics,” Digital twin technologies and smart cities, pp. 123–149,.

McLean S, Read G J. M., Thompson J, Baber C, Stanton N.A & Salmon P.M (2021) The risks associated with Artificial General Intelligence: A systematic review:
https://www.researchgate.net/publication/353888549_The_risks_associated_with_Artificial_General_Intelligence_A_systematic_review

Tito J and , Croisier S, Analyzing AI: The impact of artificial intelligence on policy, shorturl.at/tCMX8 accessed on January 12,2021.,Tzu-Wei Hung and Chun-Ping Yen, On the person-based predictive policing of AI, Ethics and Information Technology, June 2020 , <https://doi.org/10.1007/s10676-020-09539-x>

Yampolskiy R, Wittkotter E, 2021 Kill-Switch for Artificial Superintelligence: https://asi-safety-lab.com/DL/Kill-Switch-For_ASI_EW_21_12_14.pdf

Bekkum M , Borgesius F (2022) Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?
<https://www.sciencedirect.com/science/article/pii/S0267364922001133>

OLUFUNSHO I. F, BOTSIOE L, KOSHOEDO K AND OZER M (2024) Enhancing Cybersecurity With Artificial Immune Systems and General Intelligence: A New Frontier in Threat Detection and Response. IEEE Access.

] F. Mo, Z. Tarkhani, and H. Haddadi, “Sok: (2022)machine learning with confidential computing,” arXiv preprint arXiv:2208.10134,.

Lin Zhao et al. (2023).When brain-inspired ai meets agi. In: arXiv preprint arXiv:2303.15935

Fei Dou et al. (2023) Towards Artificial General Intelligence (AGI) in the Internet of Things (IoT):Opportunities and Challenges”. In: arXiv preprint arXiv:2309.07438

Hendrycks D, Woodside T and Mazeika M (2023) An Overview of Catastrophic AI Risks <https://arxiv.org/pdf/2306.12001>

Kaminski M (2023) Regulating the Risks of AI
<https://heinonline.org/HOL/LandingPage?handle=hein.journals/bulr103&div=41&id=&page=>

Bekkum M, Borgesius F, 2023 Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? <https://rb.gy/e9ivs2>

Zhong T , Liu Z , Pan Y , Zhang Y, Zhou Y, Shizhe L, Wu Z, Lyu Y, Shu P, Yu X, Cao C , Jiang H , Chen H , Li Y, Chen J , Hu H , Liu Y, Zhao H, Xu S, Dai H, Zhao L, Zhang R, Zhao W, Yang Z, Chen J , Wang P , Ruan W, Wang H , Zhao H , Zhang J, Ren Y, Qin S , Chen, Jiaxi Li T, Zidan A, Jahin A ,Chen M, Xia S , Jason Holmes , YanZhuang , Jiaqi Wang , Xu S , Xia W, Yu J , Tang K , Yang Y , Sun B , Yang T, Lu G, Wang X, Chai L, Li H, Lu J, Sun L, Zhang X, Ge B, Hu X, Zhang L, HuaZhou, Zhang L, Zhang S, Liu N, Jiang B, Kong L, Xiang Z,Ren Y, Liu J,Jiang X, Bao Y, Zhang W, Li X, Li G, Liu G,Shen D, Sikora A, Zhai X, Zhu D,Zhang T, and Liu T (2024) Evaluation of OpenAI o1: Opportunities and Challenges of AGI

<https://arxiv.org/pdf/2409.18486>