NPRC Journal of Multidisciplinary Research
Vol. 1, No. 6, Special Issue-I 2024 Pages: 1-12
ISSN: 3059-9148 (Online)
DOI: https://doi.org/10.3126/nprcjmr.v1i6.71734

A Chronology of AI Failures in Safety and Cybersecurity

Ashish Gautam*
PhD Scholar
Lincoln University College, Malaysia
ashish.gautam@texascollege.edu.np

Suman Thapaliya, PhD
IT Department
Lincoln University College, Malaysia
mailsumanthapaliya@gmail.com
https://orcid.org/0009-0001-1685-1390

Corresponding Author*

Received: August 07, 2024; Revised & Accepted: November 09, 2024

Abstract

Background: Artificial intelligence (AI) has rapidly evolved, leading to significant technological advancements and raising important questions about its implications for society. This study examines the progress and potential risks associated with AI development, drawing on historical milestones and expert predictions to outline both achievements and failures. Ray Kurzweil's forecasts highlight a future where AI could surpass human intelligence, potentially leading to new and challenging scenarios. **Aim:** This research aims to explore the safety and security challenges posed by AI, particularly the risks associated with malicious AI and AI failures, and to propose strategies for developing safe and reliable AI systems. **Methodology:** The study reviews key milestones in AI development, analyzes documented AI failures, and compares AI safety approaches to cybersecurity principles. It also examines Roman Yampolskiy's contributions to AI safety engineering and the broader implications of AI's integration into various sectors. **Results:** Historical analysis reveals numerous AI failures, from misidentifying objects to causing financial market disruptions, and highlights the challenges in ensuring AI safety. The study identifies intentional and unintentional failures and emphasizes the potential dangers of malevolent AI. A comparison with human safety efforts underscores the complexity of creating inherently safe AI systems. **Findings:** Ensuring AI safety requires a multidisciplinary approach, incorporating techniques from cybersecurity, software engineering, and ethics. The study stresses the importance of proactive measures and adversarial testing to mitigate risks. It concludes that while current AI systems pose significant challenges, the development of comprehensive safety mechanisms is crucial to prevent catastrophic outcomes in future AI advancements.

Keywords: Artificial Intelligence, Cyber security, Safety Engineering, breaching networks

Introduction

Every single day, there is a news article that describes some remarkable achievement in artificial intelligence [1]. This is something that never fails to happen. In point of fact, artificial intelligence has advanced at such a quick pace that futurologists like Ray Kurzweil are now able to predict future occurrences by projecting present tendencies into the future. Consider some of the following developments in technical innovation:

2004 A driverless car grand challenge is sponsored by DARPA. In the end, the technology used by the participants makes it possible for Google to develop an autonomous vehicle and alter existing traffic laws.

2005 In a restaurant setting, Honda's ASIMO humanoid robot delivers platters to patrons while walking at a human pace. Robots employed in the military currently use the same technology.

2007 When computers mastered the game of checkers, they paved the path taken by algorithms capable of searching through enormous informational libraries.

2011 In Jeopardy, IBM's Watson triumphs over the best human players. It is presently receiving training to provide medical advice to clinicians. It can become an authority in any academic subject.

2012 A semantic search information base called information Graph, unveiled by Google, could be the first step towards actual artificial intelligence.

2013 Facebook releases Graph Search, a search engine with semantics that is deeply aware of its users. This effectively means that there is nothing we can hide from the sophisticated algorithms.

2013 The White House funds the BRAIN initiative, which intends to use $3 billion in US dollars to reverse engineer the human brain. This funding follows an earlier European plan worth billions of euros to achieve the same goal.

2014 A modified version of the Turing Test was passed by the chatbot, which persuaded 33 percent of the judges that it was human.

2015 A single general-purpose piece of software gains the ability to beat lots of Atari games with human players.

2016 World champion is defeated by a deep neural network in one game.

The instances that were presented earlier make it very evident that breakthroughs in artificial intelligence are taking place and even accelerating as a consequence of the technology feeding off of itself. Despite the fact that the majority of research endeavors are geared towards advancing society, each technology that is generated has the potential to be utilized in either a positive or negative manner.

Ray Kurzweil made hundreds of precise forecasts for the near and distant future based on his observations of technology's exponential advancement. He predicted in 1990 that, among other things, between 2010 and 2020, we will witness:

- Project Gl**ass** eyewear, which project images onto users' retinas to create virtual reality.
- Computers equipped with "virtual assistant" software that can aid the user with a range of everyday duties (Siri).
- E-textiles, or cell phones embedded into garments that broadcast sound waves straight into users' ears.

However, his predictions for a not too distant future are genuinely amazing and terrifying. According to Kurzweil, by the year:

2029 Machines are consistently able to pass the Turing Test, which assesses a machine's ability to mimic human behavior.

2045 The moment when machines surpass humans considered the most sophisticated animals on the planet, and maybe even beyond, is known as the technological singularity.

These long-term projections, if they turn out to be accurate, as Kurzweil has been demonstrated to be on multiple occasions, would give rise to fresh and unsettling questions about our future in an era where machines are capable of feeling and communicating. The fundamental goal of the work that is being done by around 10,000 scientists all over the world is to maximize the potential of intelligent machines through the various components of their production. Taking into account the progress that has been made in artificial intelligence over the course of the previous decade, it is becoming increasingly important to make certain that the technology that we develop is beneficial to humanity. The introduction of robotic personal digital assistants, self-driving autos, and financial advisors has resulted in the emergence of a multitude of problems that have yet to be resolved. For example, self-driving cars have already been responsible for accidents, intelligent trading software has been responsible for the destruction of markets, and catboats that have grown racist and spread hate speech have embarrassed individuals. According to our projections, the frequency and severity of these catastrophes will rapidly increase as artificial intelligences continue to gain power. As soon as we successfully construct comprehensive artificial intelligence with cross-domain performance capabilities, the least of our concerns will be the possibility of hurting people's sentiments. The limits of the narrow domain AIs that are currently available are only a caution.

We recently suggested a Taxonomy of Routes Leading to Perilous AI[2], with the following justification: "It is important to understand how a potentially dangerous artificially intelligent system came to be in such a state in order to properly handle it." Science fiction literature and films in popular culture present AIs and robots as self-aware, rebelling against humans and ultimately deciding to wipe it out. It's not impossible, but it's by far the least likely path towards the development of dangerous AI. We proposed that far more plausible explanations include the willful acts of unethical individuals (referred to as "on purpose"), the unintended consequences of subpar design (referred to as "engineering mistakes"), and lastly a variety of incidents involving the influence of the system's surroundings (referred to as "environment"). Artificial intelligence (AI) that has been purposefully created to be malevolent is the most dangerous kind and the most difficult to combat since It is likely to provide the worst outcomes and is equally likely to have all other kinds of safety issues.

A follow-up paper [3] examined the construction of a malevolent AI and the significance of researching and comprehending harmful intelligent software. Similar to a doctor researching the spread of diseases, the emergence of new ones, and the effects they have on a patient's body, an AI researcher studying malevolent AI studies these same concepts. Learning how to combat diseases is the main objective, not spreading them. The authors note that the field of cybersecurity research publishes material on both how to create tools to defend cyberinfrastructure and publications concerning dangerous attacks. The information sharing that takes place between security specialists and hackers is what creates a healthy cyber-ecosystem. Hundreds of papers[4] on various suggestions aimed at creating a safe machine have been published in the field of AI safety engineering; however, nothing has been written on how to construct a malicious computer. The availability of such data would be especially beneficial to mathematicians, computer scientists, and other experts interested in creating safe artificial intelligence. These individuals are working to prevent the unintentional or intentional development of dangerous AI, which could have a detrimental impact on human activities or, in the worst case, cause the extinction of the human race. The study made the implication that an AI safety mechanism cannot be deemed functional if it is not built to withstand attacks by malicious human actors!

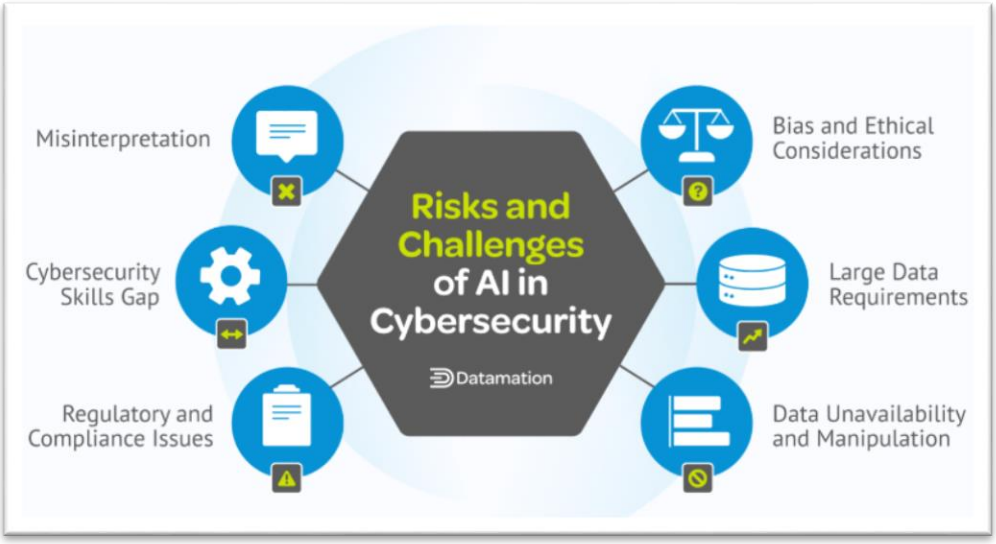

Risks and Challenges of AI in Cybersecurity
Datamation
Misinterpretation
Cybersecurity Skills Gap
Regulatory and Compliance Issues
Bias and Ethical Considerations
Large Data Requirements
Data Unavailability and Manipulation

Fig 1: The risks and challenges of using AI for cybersecurity.

Literature Review

Singh et al. (2020) The expansion of industrial information technology has resulted in the emergence of new threats, such as cyber attackers entering into industrial control systems (ICS) networks in order to steal information, harm equipment, and put people in risk. The strategies for managing cyber risks in industrial settings were discussed in this chapter, with an emphasis placed on the use of ICS-specific security metrics. Recent advancements in artificial intelligence have demonstrated that they have the potential to mitigate threats through the use of automatic reaction and real-time monitoring.

Ansari et al. (2022) evaluated the benefits that artificial intelligence offers across industries. AI's influence on cybersecurity was investigated in this study. As the field of cybersecurity continues to expand, artificial intelligence is becoming increasingly important for the protection of data and information. Today's cybersecurity demands the application of machine learning. The study investigated the effects of artificial intelligence on cybersecurity in published works. It demonstrated the increasing significance of artificial intelligence in terms of security. Both the constraints and the opportunities of AI cybersecurity were investigated. Artificial intelligence (AI) security concerns have been thoroughly researched. According to a study, artificial intelligence and cybersecurity are making progress together. AI's significance in the future of cybersecurity is highlighted in a specific study.

Viganò et al. (2020) criticized critical infrastructure cybersecurity philosophically and politically. The chapter explored national security cybersecurity limits, AI, and monitoring ethics for critical infrastructures. A literature review until 2016 discovered cybersecurity value conflicts and national security ethical difficulties. Also reviewed was the newest research on AI's role in national infrastructure cyberthreats. The chapter used four case studies to demonstrate how digital and non-digital infrastructure connections increased value trade-offs from previous years.

Radanliev et al. (2020) the most effective way using survey risk models, deep learning algorithms, and Internet of Things cybersecurity to build a dynamic and self-adapting predictive cyber risk analytics solution for Mars colony cyber risk in harsh environments. Their mathematical technique automates anomaly identification using edge computing, AI/ML, and cognitive engine design. This method employed IoT network edge AI/ML for real-time predictive cyber risk assessments. It improved risk analytics and helped identify edge computing and AI/ML strengths and shortcomings in difficult situations.

Kuzlu et al. (2021) collected data from surveys and academic publications to study AI in IoT cybersecurity. They discussed the exponential rise in IoT adoption and cybersecurity risks. They stressed the importance of AI in designing complex algorithms to defend networks and systems, including Internet of Things frameworks. Cybercriminals are utilizing adversarial AI to undertake cybersecurity attacks, they said. The review presented and summarized relevant material on IoT, AI, and AI-related assaults to understand their relationships.

1. AI Failures

History will not be unique if people fail to draw lessons from it. Many things have been compromised with disastrous consequences, including locks, signatures, bank vaults, laws, guarded leaders, blackmailing judges, buying off police officers, creating counterfeit money, brute-forcing passwords, breaching networks, computers, spoofing biometric systems, misappropriating cryptocurrency, taking advantage of planes, breaking cryptographic protocols, cracking CAPTCHAs, and even scholarly peer review. Millions of attempts have been made over millennia of human history to develop technological and logistical solutions that would increase security and safety. But there's not a single one that hasn't failed in the end. Although software and industrial robots have been involved in accidents, some fatal ones, since their inception, these incidents are not directly related to the specific intelligence that these systems possess. Nonetheless, errors committed by the intelligence that these systems are supposed to demonstrate are directly linked to AI failures. These kinds of failures can be broadly categorized as errors made during the learning phase and errors made during the performance phase. Instead of learning what its human designers intended, the system can wind up learning a different but similar purpose. One commonly used example is a computer vision system that was designed to distinguish pictures of tanks but instead learned to identify the backgrounds of such photos. [10]. Additional instances include issues brought on by shoddy utility functions that only partially compensate agents for partially desirable behaviors, like circling a target on a bicycle[11], pausing a game to prevent a loss[12], or constantly touching a football to receive credit for possession [13]. The system may fail during the performance phase for a variety of reasons, any of which could result in an AI Failure[14].

Numerous instances of AI failure are reported in the media, yet upon deeper inspection, the majority of these cases can be traced back to other factors. Only intentional intelligence failures are included in the list below. Furthermore, the instances that follow only show the initial instance of a certain failure; in spite of this, it's typical to see the same issues resurfacing in subsequent years. Lastly, the list excludes AI failures caused by intentional hacking or other means. However, the duration of AI Failures follows an exponential pattern:

1959 AI intended to be a General Problem Solver was unable to resolve practical issues.
1982 Instead of making discoveries, software learned how to cheat.
1983 The attack's occurrence was misreported by the nuclear attack early warning system.
2010 The trillion-dollar flash crash was caused by sophisticated AI stock trading software.
2011 Once instructed to "call me an ambulance," the E-Assistant started referring to the user as Ambulance.
2013 Neural networks for object recognition observed ghostly objects in specific noisy photos.
2015 Unsuitable responses were generated using an automated email reply generator.
2015 A guy was slain by an auto part-grabbing robot that grabbed him.
2015 Software for image tagging identified black people as gorillas.
2015 Software designed to filter adult content was unable to get rid of unsuitable stuff.
2016 Racist AI was used to predict recidivism.
2016 Superweapons created by unapproved NPCs in games.
2016 The patrol robot and the toddler collided.
2016 level of a world champion Playing AI made you lose a game.
2016 A deadly collision included a self-driving automobile.
2016 An AI created to communicate with Twitter users started using foul language.

There are fewer examples of artificial intelligence (AI) that function perfectly. For instance, important emails are blocked by spam filters, incorrect instructions are given by GPS, machine translation skews sentence meanings, autocorrect replaces the wrong word with the intending word, biometric systems misidentify people, and transcription software struggles to capture spoken words accurately. These are just some of the examples. According to the criteria that determine what makes a valid example of an issue with intelligent software, the list might never come to an end. Narrow Artificial Intelligence (NAI) can be applied to any software that has even a single "if statement," and any errors that occur in such software are examples of AI failure. This is the most extreme version of NAI.

By examining the compilation of Narrow AI Failures, spanning from the field's origin to contemporary systems, we can derive a straightforward generalization: An AI created to perform X will ultimately be unable to perform X. Despite its seeming insignificance, it is a potent generalization tool that can be used to forecast NAI failures in the future. For instance, by examining state-of-the-art AIs now and in the future, we can forecast that:

- Joke-generating software sometimes fails to produce humorous jokes.
- Orgasms and timely stops will not be produced by sex robots.
- Sincere and caustic remarks will be misinterpreted by sarcasm detecting software.
- Movie storylines will be misinterpreted by video description software.
- Software-generated virtual worlds are unlikely to be particularly engaging.
- In some circumstances, AI doctors may misdiagnose patients in ways that a human doctor would not.
- Systematic bias in employee screening algorithms will result in the hiring of underachievers.
- The robot explorer on Mars will misread its surroundings and plummet into a crater.

As a result of the fact that AGI is a superset of all NAIs, it will display a superset of failures in addition to more complex failures that are the result of the union of new super failures with individual NAI failures. Given that AGI is a superset of all NAIs, this is the reason why. It is not out of the question that this could pose a danger to the very life of humanity. In other words, flaws in the artificial general intelligence (AGI) have the potential to influence everything. In general, we are of the opinion that the capabilities of artificial intelligences will directly correlate to a rise in the number and intensity of deliberate hostile AI events as well as AI failures. The occurrence of this is something that we are looking forward to.

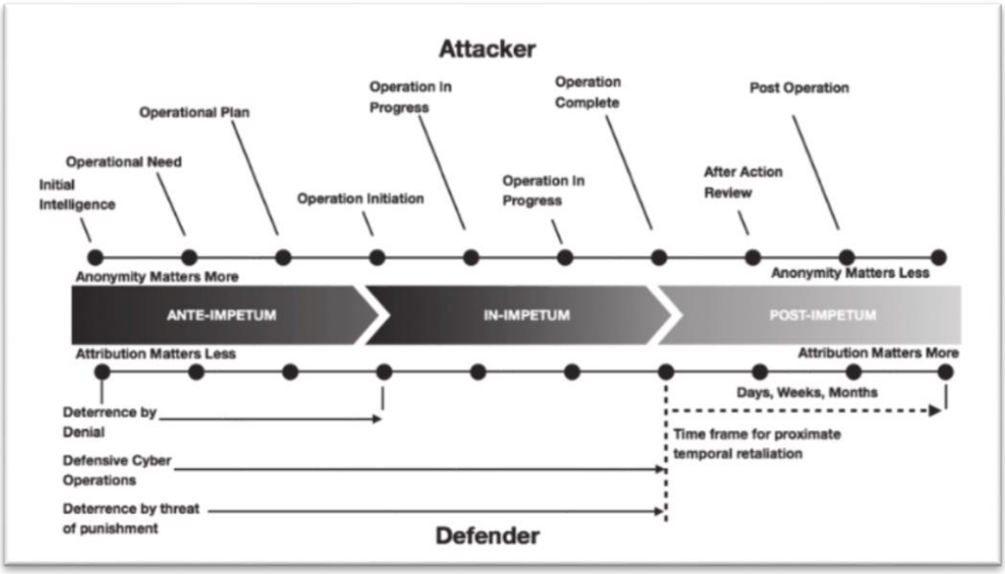


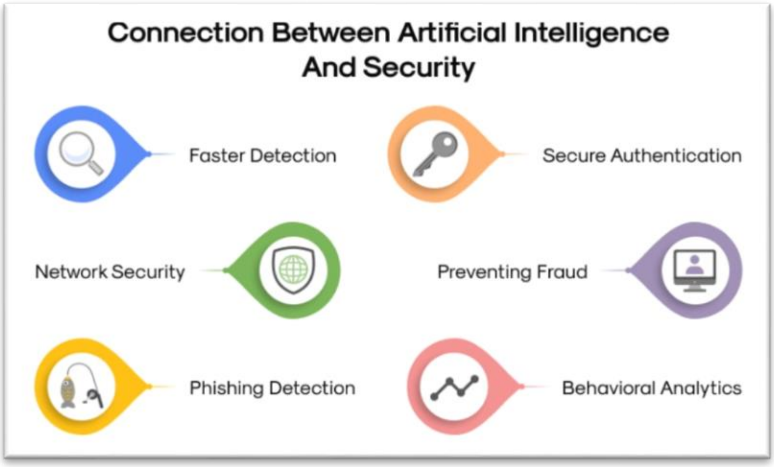
Fig 2: TIMELINE OF CYBER ATTACKS AND DEFENSE

AI Safety and Security

Roman Yampolskiy named a new line of inquiry he was promoting in 2010 by coining the terms "Artificial Intelligence Safety Engineering" and its abbreviation " AI Safety." He formally presented his ideas on AI safety in 2011 at a symposium that was subject to peer review. He then published papers on the subject in 2012, 2013, 2014, 2015, and 2016. Although the phrase may have been used informally before, Yampolskiy is the first known to have used it in a peer-reviewed publication, which is what made it popular. Prior to then, "Machine Ethics" and "Friendly AI" were the terms most frequently used to refer to the pertinent ideas. Nowadays, the bulk of eminent researchers seem to refer to the topic by the title "AI safety"[16]. In its early days, the field was considered either science fiction or pseudoscience, but it is now becoming mainstream.

Our legal system is not as advanced as our technological capabilities, and machine morality is still in its infancy. The challenge of managing sentient machines is only now being acknowledged as a significant issue, and many scientists remain dubious about the concept itself. Even worse, there are just a hundred or so people who are fully committed to filling in the gaps in our knowledge and expertise in this field everywhere. In computer science, cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology, and other related subjects, only a small fraction of them possess formal competence. It is easy to see that creating a machine that is both capable and safe is a far bigger challenge than creating a machine that is only capable. However, just 1% of researchers are working on that issue at the moment, and their funding levels aren't even close to that threshold. Since AI safety is still a very new and underfunded area of research, it can gain from incorporating techniques and concepts from more mature scientific disciplines. Attempts have been made to transfer the originally designed cybersecurity professionals' methods for protecting software systems to the new domain of protecting intelligent machines[17]. Software engineering and software verification are two more areas that can offer important techniques.

When it comes to developing code that is both safe and reliable throughout the software development process, iterative testing and debugging are absolutely necessary. Software developers now have access to a wide range of sophisticated techniques that enable for the detection and rectification of the majority of significant problems, resulting in software that is suitable for its intended use. This is despite the fact that it is reasonable to anticipate that complex software may on occasion contain bugs. Many of the modular development and testing processes that are used in the software industry can surely be used to the design of intelligent agents; however, it is not apparent how the protocols that are used for testing a finished software package will be transferred. The testing and debugging of super intelligent software would not be a suitable fit for alpha and beta testing, which entails making nearly-finished software available to experienced users for the purpose of reporting flaws discovered in real-world settings. In a similar vein, it is not feasible to just run the software and see how it operates when dealing with a super intelligent agent.



2. Cybersecurity vs. AI Safety

According to Bruce Schneier, "you don't understand the problems and you don't understand the technology if you think technology can solve your security problems." In a more general comment, Salman Rushdie said, "There is no such thing as perfect security, only varying degrees of insecurity." We put out the following theory, which we refer to as the Fundamental Theorem of Security: Every security mechanism eventually fails, thus none is ever totally safe. Simply wait longer if your security system hasn't malfunctioned.

Reduction to a different, occasionally more well-analyzed problem is a frequent technique in theoretical computer science for extracting the essential elements of challenging problems[18]. If it is possible and computationally effective to achieve such a decrease, it suggests that if the problem that is more thoroughly studied is resolved in any way, the problem that we are presently addressing will also have a functional solution. One may simplify the challenge of AGI safety to the problem of ensuring the safety of a specific person. This issue is known as the Safe Human Problem (SHP). Formally, this reduction can be achieved in a manner similar to applying a restricted Turing test to assess the AI-Completeness of a problem in the safety domain. All we wish to point out is that in both cases, we have an intelligent agent that is at least as clever as a human, and we would like to make sure that the agent is kept safe and under control; however, our work does not cover such formalism. Although it is not as easy to alter a human's DNA in practice as it is to alter an artificial intelligence's source code, it is theoretically equally feasible.

Humans have been shown to be unsafe around one another and oneself. Success is unattainable despite millennia of attempts to create safe beings through eugenics, relationships, family, oaths, education, laws, ethics, punishment, and reward. People kill and commit suicide, steal and cheat, lie and betray, and generally do these things in proportion to how much they get away with. Strong enough tyrants will violate all human rights, breach all the laws, and enslave people. They will also commit genocide. The reality that there isn't such a thing as a human without sin. Reducing these dangerous tendencies to a level where our civilization can endure is the best we can hope for. The most we can hope for, even with sophisticated genetic engineering [19], is a further decrease in the level of human danger. People are inherently dangerous as long as we allow them to have free will, which means they can be bought off, will lie, and will put their own interests ahead of those they are told to serve. Humans are everything but safe, even though they are simple instances of a solution to the Value Learning Problem. This calls into doubt our current optimism that resolving VLP will lead to Safe AI. This is a crucial matter. Bruce Schneier once said, "Professionals target people; amateurs only attack machines." As a result, we consider research on AI safety to be adversarial in nature, akin to security or cryptography, at least in part.

In most circumstances, the damage caused by a cybersecurity system failure is unpleasant but bearable: someone loses money, someone loses privacy, or even someone loses their life. Safety lapses are as important for narrow AIs as they are for general cybersecurity, but they are essentially different for artificial general intelligence. An existential danger event could result from a single super intelligent system malfunction. Everybody could lose everything and the possibility of the extinction of all biological species in the cosmos exists if an AGI Safety mechanism fails. You will get another chance—or at least a better one—if you use security systems. With the AGI Safety system, you only get one chance at success, so failing ahead is not an option. Even worse, a typical security system is likely to have some failures; perhaps not much data will be compromised, for instance. Whether an AGI Safety system succeeds or fails depends on whether it has a controlled and secure superintelligence or not. While preventing attacks from successfully breaching the system is the aim of AI safety, cybersecurity aims to lower the amount of successful attacks on the system. Because of this, the ability to distinguish between NAI and perhaps AGI programmes is a fundamentally important outstanding topic in the field of AI safety.

There are numerous issues. It is not possible for us to track, visualize, or evaluate the actions of super intelligent agents. Even more trivially, we have no idea what will happen when such programme is put into use. Should our surroundings change right away? Are we supposed to see nothing? What is the time frame within which something ought to be detectable? Will it happen too quickly for us to notice, or will we notice it too slowly? Will the effects be felt locally or in far-off places of the globe? How are standard tests conducted? Which data sets are they applied to? What does a general intelligence "edge case" consist of? There are a lot of questions, but as of right now, no one has the answers. The interplay between intelligent software and security measures intended to maintain AI's safety and security may cause further difficulties. Additionally, we will need to evaluate every AI safety feature that is presently in development. Even if AI is still at the human level, tests can be conducted where the artificial agent is replaced by a human agent. It appears that adversarial testing is not possible using current technology at levels above human capabilities. More importantly, there would only ever be one opportunity for testing.

Conclusion

In summary, there are a lot of risks and prospects connected to the rapid development of artificial intelligence (AI) for humanity. While breakthroughs in AI technology have led to remarkable achievements across various domains, from driverless cars to advanced gaming algorithms, the accelerating pace of progress also raises concerns about the potential consequences of unchecked development. As visionary forecasts predict the eventual surpassing of human intelligence by machines, the need for rigorous AI safety engineering becomes increasingly urgent. The literature review underscores the importance of learning from past AI failures, establishing incident databases, and integrating AI with cybersecurity measures. Addressing the challenges of AI safety requires interdisciplinary collaboration, rigorous testing, and proactive measures to mitigate risks. Ultimately, ensuring the responsible development and deployment of AI is crucial to harnessing its transformative potential while safeguarding against unintended consequences and potential existential threats.

Abbreviation

SHP - Safe Human Problem
AI – Artificial Intelligence
AGI – Artificial General Intelligence
NAI – Narrow Artificial Intelligence

References

1. Yampolskiy, R. V. (2015). *Artificial superintelligence: a futuristic approach*. cRc Press.
2. R. V. Yampolskiy, "Taxonomy of Pathways to Dangerous Artificial Intelligence," in Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
3. F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," presented at the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016), New York, NY, July 9, 2016.
4. K. Sotala and R. V. Yampolskiy, "Responses to Catastrophic AGI Risk: A Survey," Physica Scripta, vol. 90, 2015.
5. Singh, S., Karimipour, H., HaddadPajouh, H., & Dehghantanha, A. (2020). Artificial intelligence and security of industrial control systems. *Handbook of Big Data Privacy*, 121-164.
6. Ansari, M. F., Dash, B., Sharma, P., & Yathiraju, N. (2022). The impact and limitations of artificial intelligence in cybersecurity: a literature review. *International Journal of Advanced Research in Computer and Communication Engineering*.
7. Viganò, E., Loi, M., & Yaghmaei, E. (2020). Cybersecurity of critical infrastructure. *The Ethics of Cybersecurity*, 157-177.
8. Radanliev, P., De Roure, D., Page, K., Van Kleek, M., Santos, O., Maddox, L. T., ... & Maple, C. (2020). Design of a dynamic and self-adapting system, supported with artificial intelligence, machine learning and real-time intelligence for predictive cyber risk analytics in extreme environments–cyber risk in the colonisation of Mars. *Safety in Extreme Environments*, *2*, 219-230.
9. Kuzlu, M., Fair, C., & Guler, O. (2021). Role of artificial intelligence in the Internet of Things (IoT) cybersecurity. *Discover Internet of things*, *1*(1), 7.

10. E. Yudkowsky, "Artificial intelligence as a positive and negative factor in global risk," Global catastrophic risks, vol. 1, p. 303, 2008.
11. J. Randløv and P. Alstrøm, "Learning to Drive a Bicycle Using Reinforcement Learning and Shaping," in ICML, 1998, pp. 463-471.
12. T. M. VII, "The first level of Super Mario Bros. is easy with lexicographic orderings and time travel," The Association for Computational Heresy (SIGBOVIK) 2013, 2013.
13. A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in ICML, 1999, pp. 278-287.
14. F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," arXiv preprint arXiv:1605.02817, 2016.
15. R. V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," presented at the Philosophy and Theory of Artificial Intelligence (PTAI2011), Thessaloniki, Greece, October 3-4, 2011.
16. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565, 2016.
17. R. Yampolskiy, "Leakproofing the Singularity Artificial Intelligence Confinement Problem," Journal of Consciousness Studies, vol. 19, pp. 1-2, 2012.
18. R. M. Karp, "Reducibility Among Combinatorial Problems," in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, Eds., ed New York: Plenum, 1972, pp. 85-103.
19. R. V. Yampolskiy, "On the Origin of Samples: Attribution of Output to a Particular Algorithm," arXiv preprint arXiv:1608.06172, 2016.