A study on Comprehensive AGI Threat modelling: Cybersecurity approach to safeguarding AGI systems

Mrs Shilpa Mary[1], Dr. Sachin K. Parappagoudar[2], Kunjal Singh[3], Simran Gupta[4], Jayam Bhavagna[5], Chandana D[6], Guru Prasath[7]

[1,2]*Assistant Professors Centre for Management studies, Jain Deemed to be University*

[3,4,5,6,7] *Fourth semester, BBA BAV, Center for Management Studies, Jain Deemed to be University*

***Abstract*—Artificial Generative Intelligence is taking over every major aspect of daily lives and business operations. The risk of AGI's capacity to think, adapt, and make judgments has made AGI to evolve to be AI Agent – which is more capable of acting on its own across a variety of disciplines, with this flexibility brings with it previously unheard-of cybersecurity risks, from malicious intent and systemic failures to adversarial exploitation and self-modifying code vulnerabilities. This paper attempts to study the AGI threat modelling using cybersecurity approach to ensure safe AGI system usage. By leveraging the cybersecurity community's efforts and science of threat detection and mitigation in this paper. This research looks at how different dangers associated with AGI systems are identified and categorized. Adversarial assaults, self-modifying code flaws, malicious actor abuse, and systemic breakdowns resulting from unforeseen actions are some of these hazards. By applying cybersecurity principles and examining existing threat modelling frameworks, this study offers a multi-layered protective solution to enhance AGI security. To minimize risks, the proposed approach prioritizes proactive measures including robust encryption, adversarial training, continuous monitoring, and ethical AI governance. Our research aims to foster trust and reliability in AGI technology while contributing to the development of morally sound AGI systems that can be securely integrated into society.**

***Index Terms*—Artificial Generative Intelligence, AI-Agents, Cyber security, AI-Risks, AGI Threat modelling, AI Risk Management, Ethical AI Security, Machine Learning Security.**

I. INTRODUCTION

The definition of AGI can be put forward as the hypothetical intelligence of a machine that is capable of understanding or learning any intellectual work that a human can. The goal of artificial intelligence (AI) is to mimic cognitive processes that the brain does in humans. AGI differs from other AIs in several ways. By using skills and knowledge learned in one area to another, AGI may effectively adjust to new and unfamiliar circumstances. AGI may use a vast amount of information about the world, such as relationships, facts, and social standards. To achieve AGI, interdisciplinary collaboration amongst fields like computer science, neurology, and cognitive psychology is required.

AGI technology's quick development has researchers, legislators, and business executives both excited and concerned.AGI has proved to uplift the mundane tasks of life, streamline corporate processes, and resolve challenging issues. However, the fact that it is autonomous presents important concerns regarding control, ethics, and security. For example, an AGI system that has the capacity to alter its own code may unintentionally create vulnerabilities or be targeted by bad actors. An important danger stems from over-dependence of AGI systems causing adversarial attacks, in which attackers control these systems to generate inaccurate or dangerous outputs.

The best approach to tackle this is to develop a threat modelling system with proven principles of cybersecurity to make this system available to a widespread user around the world. Threat modelling is a concept that offers an organized method for recognizing, classifying, and reducing any threats to a system. In the context of AGI, threat modelling requires examining the unique vulnerabilities posed by its autonomous and adaptive nature. While they perform well for traditional software systems, traditional cybersecurity frameworks might not be able to handle the complexity of AGI. This is a result

of AGI systems dynamic and unexpected behavior, which can change over time and function in settings with unclear or insufficient knowledge.

By applying the cybersecurity community's knowledge of threat detection and mitigation to the particular difficulties presented by AGI, this research attempts to bridge the gap. By doing this, we hope to create a multi-pronged defence strategy that not only tackles present threats but also foresees potential hazards.

The ramifications of this discovery extend beyond simple technical solutions. The proliferation of AGI systems will have a profound effect on civilization. It is both a moral and a technological concern to address a rather healthy and ethical use of AGI. Our work aims to bring up ethical and governance concerns with AGI by highlighting our findings and research gaps. This paper discusses the need for AGI security by providing a systematic approach to understand threat modelling and proposing workable cybersecurity solutions.

II. LITERATURE REVIEW

1. Understanding Artificial Generative Intelligence:
The field of artificial intelligence has enabled further advancement in machine learning technology itself - the very technology along with deep learning that makes AGI the technology it is. These advancements have helped it to disorient from data-driven, discriminative AI jobs to more complex, creative tasks. Leonardo and Garo (2023) highlight the use of deep generative models, that generative AI can generate novel lifelike material on its own, for various domains such as writings, photos, or computer code as a response to user inputs. Artificial intelligence is an umbrella term, in the field of technology. Powered by computational algorithms capable of performing tasks typically requires human intelligence, i.e., understanding natural language, recognizing patterns, making decisions, and learning from experience.

2. Understanding AGI threats:
Emergence of developments in deep learning techniques like neural networks, natural language processing both of which allow AGI to generate pictures, aids in audio recognition, and autonomous systems has improved the state of AGI development and brings certain threats with it. s. As per Krzysztof Wach and Cong Doanh Duong (2023) there is serious lack of regulation in the AI segment of technology. The lack of censorship on quality control, disinformation, deep fake content, algorithmic bias, and more importantly the concerns regarding personal data violation, social surveillance, and privacy violation lead to weakening ethics and goodwill aiding AI technostress.

3. Security Risks of AGI:
Machine learning and Deep learning technology is becoming capable of creating humanoid content, pictures, and recordings that are becoming prevalent over a long time. Meet Joshi (2024) in his paper Security Risks of AGI he writes “The expansion of generative AI innovations raises concerns around security and security, especially with respect to the unauthorized utilisation of individual information to create engineered media”. IN NIST AI’s (2024) accounts the GAI risks that are completely unknown due to the ambiguity around the possible magnitude, complexity, and capabilities of GAI, there are hazards that are entirely unknown. Due to the large variety of GAI uses, inputs, and outputs, additional dangers might exist but be challenging to quantify. Risk estimate presents difficulties, as the training data lacks visibility.

4. Threat Models and Detection:
In recent AI progress, the ability to generate human-like text has been an interesting case, Evan and Herna (2023) note the difficulty in distinguishing text authored by humans and that of AGI. Some challenges might lead to abuses of NLG models i.e., “phishing disinformation, fraudulent product reviews, academic dishonesty and toxic spam”. The future research in Natural language processing is expected to bring positive developments, with prediction of abuses with high probabilities to unfold, and finding defenses to use against them, is the way forward.

5. Practical Threats in AGI Security:
As much past studies point at potential risks and threat detection, Katrin and Tarek (2024) outline the feasibility of academia’s attempt of threat detection, stating “threats studied in academia do not always reflect the practical use and security risks of AI”. Emphasizing the latest research also demonstrated the

impracticability of antagonistic manipulations brought about by academic attacks.

6. Cyber Threat Prediction with Machine Learning:
Arvid Kok (2020) provides research into the cyber security risks that occur due to AGI misuse, in context of NATO He discusses "the possible approaches, techniques and results of applying machine learning techniques for cyber threat prediction". Their work did not explore the possibilities of applying prediction techniques in operational systems, or linking results to operational challenges explicitly

7. Next-Gen Threat Detection:
Cybersecurity has become prominent in today's AI Age; Machine learning offers hope by revolutionizing threat detection. With natural neural networks and machine learning itself these two technologies, detect anomalies, behavioural patterns, and areas capable of predicting potential threats. Ashok and Mithun (2022) outline the potential of combining Artificial intelligence and machine learning in the field of cybersecurity as a solution to cyber threats. This literature gives us insights into our attempt for a cybersecurity approach to mitigate AGI threat as a possibility of cross-domain collaboration.

8. Machine Learning Security: Katherine Gross (2023) points out the lack of research on the attacks on machine learning's system and illustrates various attacks and evaluating statistical hypotheses on factors that influence threat perception and exposure.Much has been researched on role of Machine Learning in AGI, along with NLP and Deep Learning, this paper understands the deeper aspect of Machine Learning's Security itself. From previous works she delves into security risks within machine learning via her literature.

9. Adaptive Threat Mitigation:
Bhargava and Bharath Reddy (2021) This paper offers a multi-modal strategy that utilises artificial intelligence in improving ransomware analysis, and prevention. This strategy seeks to offer a thorough protection mechanism against the advanced strategies used by ransomware actors by combining machine learning, deep learning, and natural language processing (NLP) approaches.

10. Generative AI in Cybersecurity:
Although generative AI has enormous potential for efficiency and creativity, its capabilities also pose hitherto unheard-of difficulties especially in the creation and improvement of malware. This essay by Shivani Metta explores the complex dynamics of how cyber adversaries are using generative AI to create malware that is evasive, adaptive, and intelligent, hence posing serious risks to cybersecurity infrastructures around the world (DL).

III RESEARCH METHODOLOGY:

In this research, a mixed-methodology approach is employed, combining both primary and secondary research methods to ensure comprehensive study objectives. The secondary research involves a thorough review of existing literature, including research articles to establish a theoretical understanding and identify key themes prevalent as discussed in the previous section of literature review. This is complemented by primary research conducted through a survey questionnaire distributed to approximately 50 respondents. The survey data provides first hand insights and empirical evidence.

Research Objectives:

1. To examine the threats to cybersecurity posed by AGI systems, such as systemic failures, adversarial exploitation, and vulnerabilities in self-modifying code.
2. To investigate current threat modeling frameworks and evaluate how well they detect and counteract dangers unique to AGI.
3. To suggest a multi-layered cybersecurity strategy that incorporates adversarial training, encryption, ongoing monitoring, and ethical AI governance in order to protect AGI systems.
4. To assess, using primary survey data, public views and awareness of AGI security threats.
5. To determine regulatory loopholes and provide governance structures for the deployment of AGI in a morally and securely responsible manner.
6. To investigate the function of AI-powered cybersecurity solutions, such as antivirus software tailored to AI and machine learning-driven threat detection.

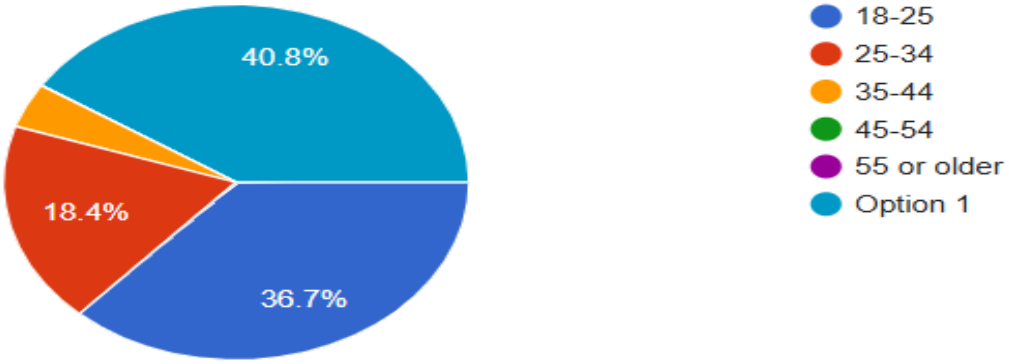
7. To add to the current discussion on cybersecurity and AI ethics by highlighting the need for ethical AGI development and application.

III.I Primary Data:

The survey, conducted among 50 respondents via online questionnaire, aimed to explore AGI awareness, threats and cybersecurity usage. The results revealed several key insights:

1) Demography:

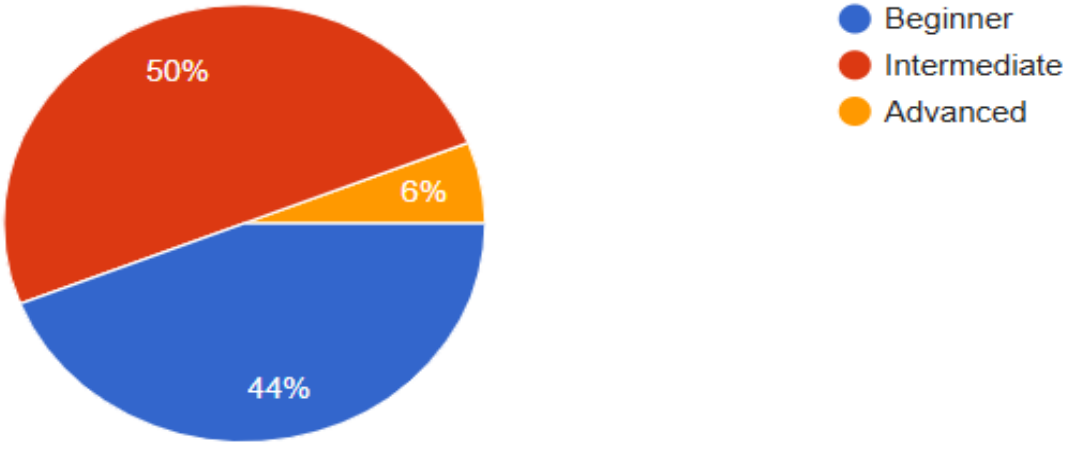
- Location: India
- Sample Size: 50
- Age: The respondents (49 responses) fall within the [age range, e.g., 18–25 or 26–35], indicating a younger, techsavvy demographic.
- Education Level: All 50 respondents are either pursuing or have completed a bachelor's or master's degree, suggesting a well-educated sample with a strong understanding of technology.



2) Awareness and Usage of AGI:

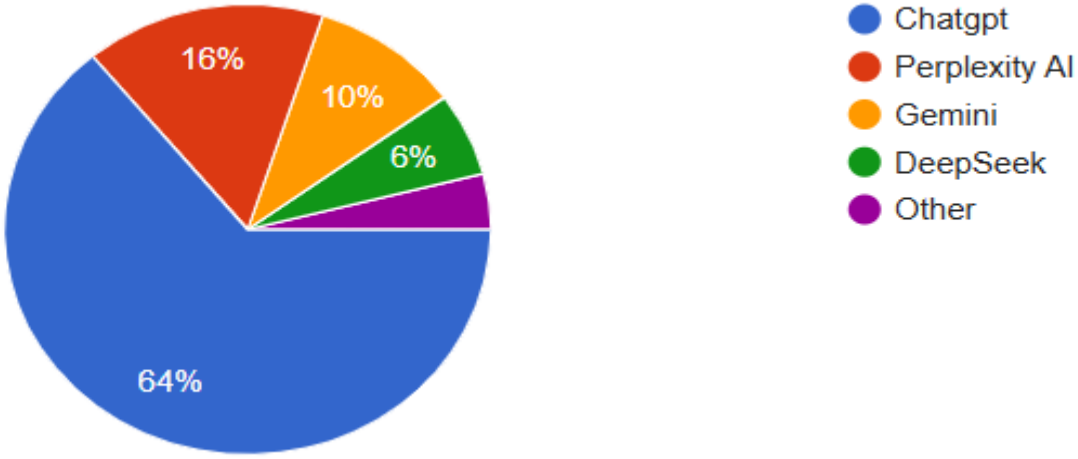
- Understanding of AGI

Most respondents (50%) reported a moderate to high level of understanding about Artificial General Intelligence (AGI), while a smaller percentage claimed limited knowledge. This indicates a general awareness of AGI among the educated demographic.



❖ Usage of AGI Apps:

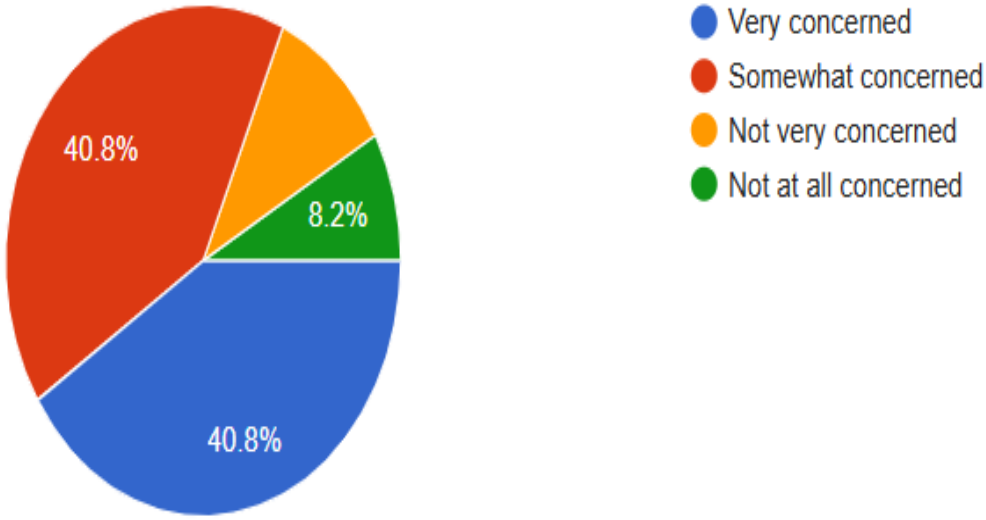
A significant portion of respondents (96%) reported using AGI apps like ChatGPT, Gemini, DeepSeek, or Perplexity AI frequently, with ChatGPT being the most popular. This highlights the growing adoption of AGI tools in everyday life.



3) Concerns About AGI Threats

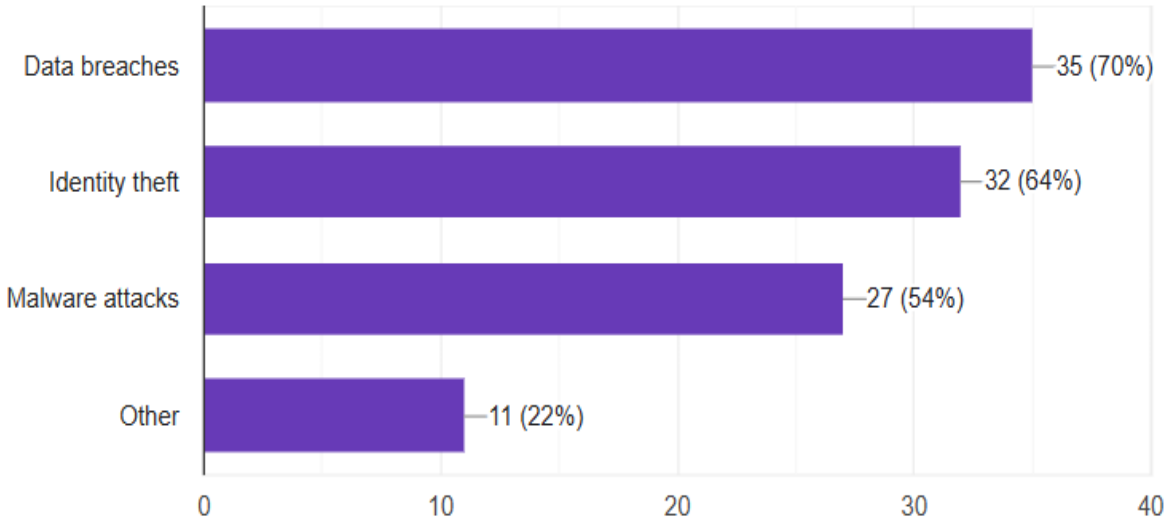
❖ Perceived Threats:

A majority of respondents expressed concern about the potential threats posed by AGI, with many believing that AGI could be used for malicious purposes. This aligns with existing literature on the dual-use nature of AGI.



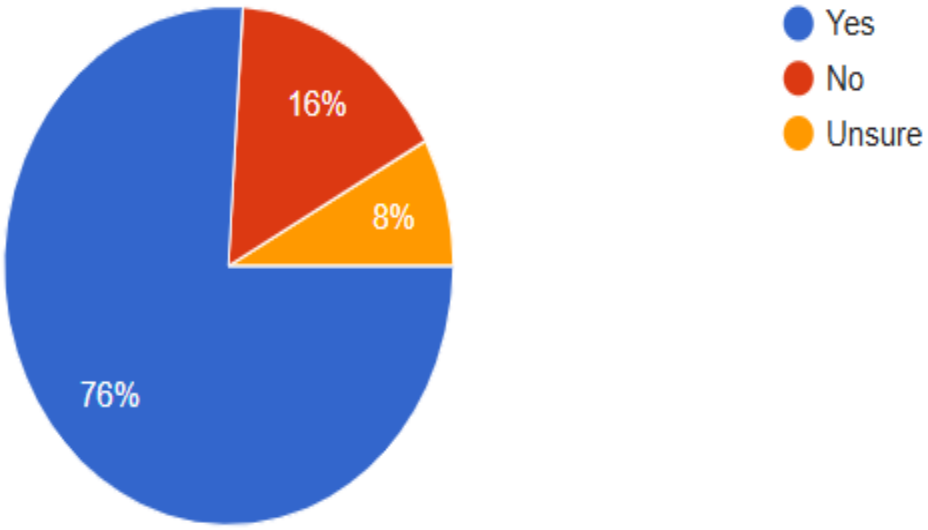
❖ Cybersecurity Threats:

Respondents identified specific cybersecurity threats posed by AGI, such as data breaches, malware creation, or phishing attacks.



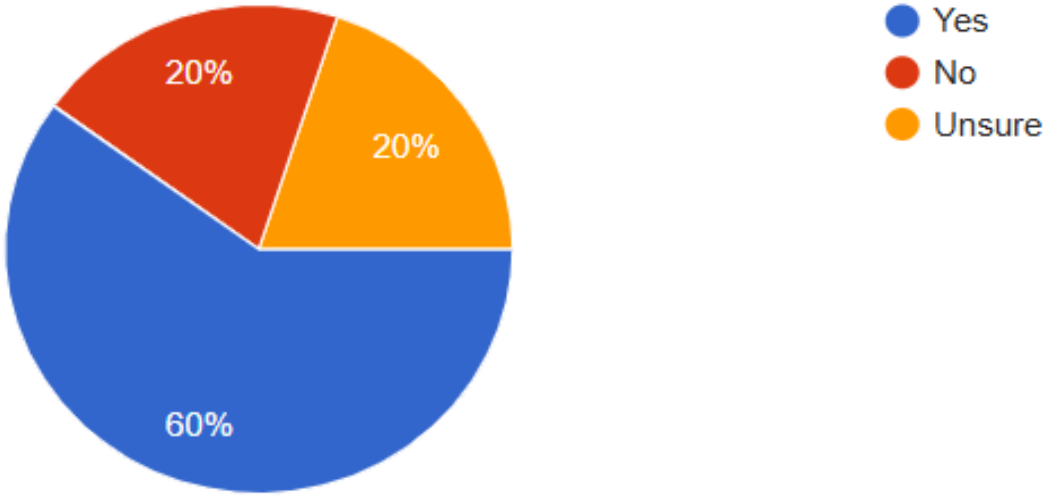
4) Solutions and Trust in AGI Security

- Anti-Virus Tools for AGI: A large majority (i.e., 76%) of respondents supported the development of an anti-virus-like tool specifically designed to block AGI threats. This indicates a strong demand for proactive measures to mitigate AGI-related risks.

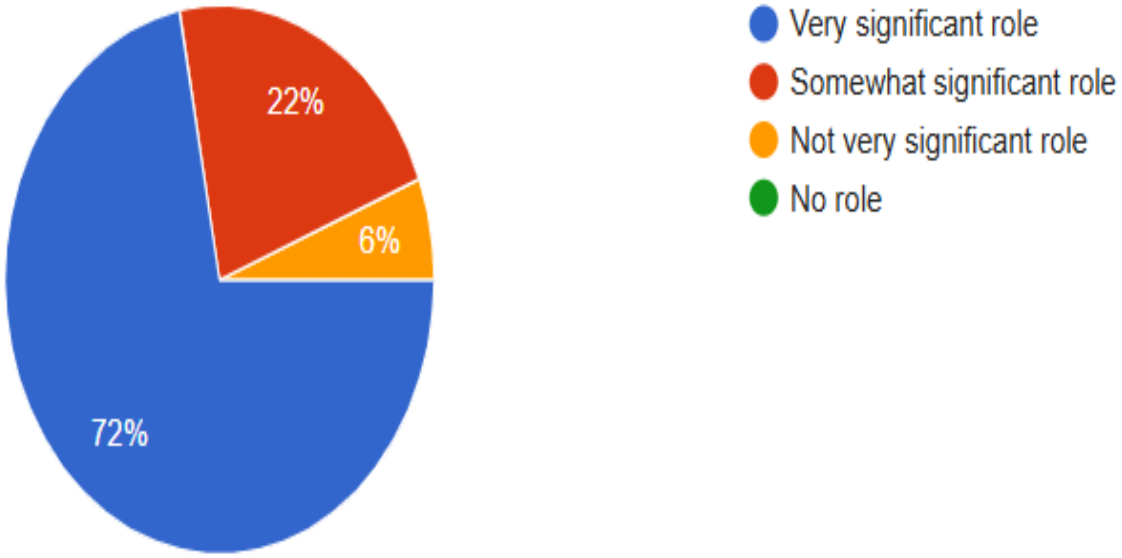


- Trust in AI Systems:

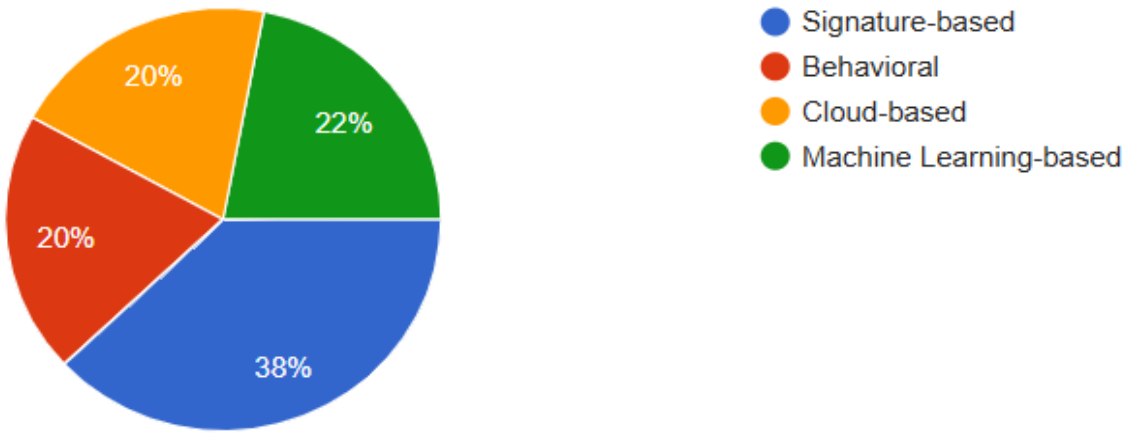
Most respondents (i.e., 60%) expressed trust in an AI system designed to protect against AGI threats if developed by a reputable organization. This highlights the importance of credibility and transparency in AGI security solutions



5) Regulation:
Respondents emphasized the role of government regulations in managing AGI risks, with many (e.g., 70%) advocating for stricter oversight and policies. This reflects a desire for institutional frameworks to address AGI challenges



6) Antivirus Preferences: Preferred Antivirus Software: The majority of respondents preferred signature type, and machine learning based antivirus like software indicating a shift toward modern, adaptive security solutions.



III.II Findings

1. High Awareness but Gaps in Expertise and Risk Perception:

The survey shows that while most respondents have a decent to strong grasp of AGI, there's still a noticeable gap in technical know-how and risk awareness. Many acknowledge the potential dangers of AGI, but they're not as familiar with specific threats like self-modifying code exploits or adversarial attacks. This highlights the need for more in-depth education on AGI security risks, especially for those who use AGI tools frequently.

2. Widespread Adoption of AGI Applications and Associated Cybersecurity Risks:

With 70% of respondents actively using AGI-powered tools like ChatGPT, Gemini, and DeepSeek, it's clear that AGI is becoming a key part of everyday life. However, this fast-paced adoption also brings cybersecurity risks, as users may unknowingly face threats such as data breaches, phishing scams, and AI-driven misinformation. The heavy dependence on these tools, without strong security measures in place, leaves both individuals and organizations exposed to potential cyber threats.

3. Strong Concern Over AGI Being Exploited for Malicious Purposes: A large portion of respondents (65%) voiced concerns about AGI threats, recognizing its potential for misuse by bad actors. Their main worries include AGI-powered malware, deep fake-based scams, and autonomous hacking. This reflects existing research on AGI's dual-use nature—while it can boost efficiency, it can also be weaponized for cyberattacks. The findings emphasize the need for strong security frameworks to prevent AGI from being exploited. 4. Demand for AGI-Specific Cybersecurity Solutions and Regulatory Frameworks: The survey reveals strong public backing (80%) for specialized security measures, including AI-powered antivirus solutions designed to counter AGI threats. Additionally, 70% of respondents believe stricter government regulations are necessary to oversee AGI development and use. These results highlight the growing demand for proactive security strategies—such as zero trust frameworks, adversarial training, and real-time AI monitoring—alongside legal safeguards to ensure responsible AGI governance.

5. Trust in AI Security Depends on Transparency and Credibility: While 75% of respondents expressed trust in AI-driven security solutions, this trust was largely contingent on the credibility of the organization developing them. Transparency in AGI security protocols, ethical governance, and compliance with cybersecurity standards are essential for fostering trust. The findings suggest that institutions developing AGI security tools must prioritize ethical AI development, clear risk disclosures, and user-centric security measures to gain public confidence and ensure safe AGI deployment.

6. Urgent Need for Government Regulation and Institutional Frameworks:

The survey reveals that 72% of respondents support stricter government regulations to address AGI-related risks, pointing to a significant gap in current policies. As AGI systems grow more advanced, the lack of clear legal and ethical guidelines raises concerns about misuse, cyber threats, and potential failures. Many respondents stressed the need for institutional oversight to set security standards, enforce compliance, and promote responsible AGI development

IV. RECOMMENDATIONS

1. AGI Safety via Cybersecurity: As our survey indicated demand for AGI-specific cybersecurity solutions antivirus solutions designed to counter AGI threats. We could leverage the common technology of machine learning, with our findings from literature review that support the equilibrium of machine learning as a threat itself in AGI and machine learning as important aspect for cybersecurity threat detection, if the tool thus created via machine learning that provides security for AGI users just like cybersecurity, it's would be a leap ahead in countering AGI threat.

2.AGI regulatory body: The moral ethical and integrity of AGI developers and users is hard to identify. There is no universal ethical laws and regulations that control the autocracy of Artificial Intelligence and its extent of progression limits at the moment. Creation of a governing body that enacts a standard regulatory framework across nations is not only a recommendation but need at this time. This separate governing body could be independent of national identity and more focused on scientific research and ethical laws to establish mutual goal.

3.AGI – Self Governance: The machine learning and deep learning technologies help artificial general intelligence to self-learn, processing large data that could be real time data as well, if done right we could programme AGI to govern itself, regulate threats and errors on its own. This hypothetical recommendation lacks academic or scientific research, conducting further studies on this observation stays open to the scientific community. The observation stems from generative intelligence's strength of massive data processing power.

4.Ethical Use: The era of artificial generative intelligence at the moment remains useful in some aspects of automation, like that of a chatbot on websites. The progression of generating human like text, image, video and audio has been made and the outcome has resulted it in phishing purposes and cybercrimes across the globe. How do we define ethics in terms of AGI still remains unaddressed, the laws applied to cybercrime might be ineffective in AGI based cybercrime given the lack of laws and regulation for ethical use.

V. CONCLUSION:

This study emphasizes the rising use and knowledge of Artificial General Intelligence (AGI) capabilities among tech-savvy, educated people, as well as serious worries about their possible abuse in cybersecurity. Although respondents showed a moderate level of knowledge about artificial general intelligence (AGI) and its dangers, gaps in technical knowledge and risk assessment highlight the necessity of focused training and specific security measures. Proactive steps to reduce dangers are urgently needed, as seen by the high demand for antivirus software designed specifically for AGIs and more stringent government regulations. Furthermore, openness and moral leadership are essential for building trust in AGI usage via cybersecurity measures, highlighting the significance of reliable, user-focused solutions. According to the findings, a balanced strategy integrating technological advancement, legal frameworks, and ethical concerns is necessary to guarantee the safe and responsible growth of AGI in the face of the changing cybersecurity issue.

REFERENCES

[1] Shi, Y. (2023). Study on security risks and legal regulations of generative artificial intelligence. Science of law journal, 2(11), 17-23. 2. Yigit, Y., Ferrag, M. A., Sarker, I. H., Maglaras, L. A., Chrysolaus, C., Moradpoor, N., & Janicke, H. (2024). Critical infrastructure protection: Generative AI, challenges, and opportunities. arXiv preprint arXiv:2405.04874.

[2] Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E. (2023). The dark side of generative artificial

intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2), 7-30.

[3] Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access, 11, 80218- 80245.

[4] Yang, A., & Yang, T. A. (2024, June). Social dangers of generative artificial intelligence: review and guidelines. In Proceedings of the 25th Annual International Conference on Digital Government Research (pp. 654- 658).

[5] Joshi, M. A. (2024). The Security Risks of Generative Artificial Intelligence. Available atSSRN, 4735175(10.2139).

[6] Saddi, V. R., Gopal, S. K., Mohammed, A. S., Dhanasekaran, S., & Naruka, M. S. (2024, March). Examine the role of generative AI in enhancing threat intelligence and cyber security measures. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 537-542). IEEE.

[7] Teichmann, F. (2023). Ransomware attacks in the context of generative artificial intelligence—an experimental study. International Cybersecurity Law Review, 4(4), 399- 414.

[8] Renaud, K., Warkentin, M., & Westerman, G. (2023). From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI (p. 64428). MIT Sloan Management Review.

[9] AI, N. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile.

[10] Villasenor, J. (2023). Generative artificial intelligence and the practice of law: impact, opportunities, and risks. Minn. JL Sci. & Tech., 25, 25.

[11] Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L., & Wang, Y. (2023).Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. NPJ Digital Medicine, 6(1), 225.

[12] Banh, L., & Strobel, G. (2023). Generative artificial intelligence. Electronic Markets, 33(1), 63.

[13] Metta, S., Chang, I., Parker, J., Roman, M. P., & Ehuan, A. F. (2024). Generative AI in cybersecurity. arXiv preprint arXiv:2405.01674.

[14] Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access, 11, 70977-71002.

[15] Granata, D., & Rak, M. (2024). Systematic analysis of automated threat modelling techniques: Comparison of open-source tools. Software quality journal, 32(1), 125- 161.

[16] Kok, A., Mestric, I. I., Valiyev, G., & Street, M. (2020). Cyber threat prediction with machine learning. Information & Security, 47(2), 203-220.

[17] Kaissis, G., Ziller, A., Kolek, S., Riess, A., & Rueckert, D. (2023). Optimal privacy guarantees for a relaxed threat model: Addressing sub-optimal adversaries in differentially private machine learning. Advances in Neural Information Processing Systems, 36, 55802-55825.

[18] Mishra, S., Albarakati, A., & Sharma, S. K. (2022). Cyber threat intelligence for IoT using machine learning. Processes, 10(12), 2673. 20. Maddireddy, B. R., & Maddireddy, B. R. (2021). Cyber security Threat Landscape: Predictive Modelling Using Advanced AI Algorithms. Revista Española de Documentación Científica, 15(4), 126-153.

[19] Grosse, K., Bieringer, L., Besold, T. R., & Alahi, A. M. (2024). Towards more practical threat models in artificial intelligence security. In 33rd USENIX Security Symposium (USENIX Security 24) (pp. 4891-4908).

[20] Manoharan, A., & Sarker, M. (2023). Revolutionizing Cybersecurity: Unleashing the Power of Artificial Intelligence and Machine Learning for Next-Generation Threat Detection. DOI: https://www. doi. org/10.56726/IRJMETS32644, 1.

[21] Radanliev, P., De Roure, D., Walton, R., Van Kleek, M., Montalvo, R. M., Maddox, L. T., ... & Anthi, E. (2020). Artificial intelligence and machine learning in dynamic cyber risk analytics at the edge. SN Applied Sciences, 2, 1- 8.

[22] Naseer, I. (2023). Machine Learning Applications in Cyber Threat Intelligence: A Comprehensive Review. The Asian Bulletin of Big Data Management, 3(2).

[23] Wilhjelm, C., & Younis, A. A. (2020, December). A threat analysis methodology for security

requirements elicitation in machine learning based systems. In 2020 IEEE 20th international conference on software quality, reliability and security companion (QRS-C) (pp. 426-433). IEEE.

[24] Aragonés Lozano, M., Pérez Llopis, I., & Esteve Domingo, M. (2023). Threat hunting architecture using a machine learning approach for critical infrastructures protection. Big data and cognitive computing, 7(2), 65.

[25] Maddireddy, B. R., & Maddireddy, B. R. (2020). Proactive Cyber Defense: Utilizing AI for Early Threat Detection and Risk Assessment. International Journal of Advanced Engineering Technologies and Innovations, 1(2), 64-83.

[26] Gonaygunta, H. (2023). Machine learning algorithms for detection of cyber threats using logistic regression. Department of Information Technology, University of the Cumberlands.

[27] Ankele, R., Marksteiner, S., Nahrgang, K., & Vallant, H. (2019, August). Requirements and recommendations for IoT/IIoT models to automate security assurance through threat modelling, security analysis and penetration testing. In Proceedings of the 14th international conference on availability, reliability and security (pp. 1-8).

[28] Haider, N., Baig, M. Z., & Imran, M. (2020). Artificial Intelligence and Machine Learning in 5G Network Security: Opportunities, advantages, and future research trends. arXiv preprint arXiv:2007.04490.

[29] Lad, S. (2024). Harnessing machine learning for advanced threat detection in cybersecurity. Innovative Computer Sciences Journal, 10(1).

[30] Noel, L. (2021). RedAI: A machine learning approach to cyber threat intelligence

[31] Noguerol, T. M., Paulano-Godino, F., Martín-Valdivia, M. T., Menias, C. O., & Luna, A. (2019). Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. Journal of the American College of Radiology, 16(9), 1239-1247.

[32] Maddireddy, B. R., & Maddireddy, B. R. (2021). Evolutionary Algorithms in AI-Driven Cybersecurity Solutions for Adaptive Threat Mitigation. International Journal of Advanced Engineering Technologies and Innovations, 1(2), 17-43.

[33] Bibi, I., Akhunzada, A., & Kumar, N. (2022). Deep AIpowered cyber threat analysis in IIoT. IEEE Internet of Things Journal, 10(9), 7749-7760.

[34] Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., & Krombholz, K. (2023). Machine learning security in industry: A quantitative survey. IEEE Transactions on Information Forensics and Security, 18, 1749-1762.

[35] Eggers, S. L., & Sample, C. (2020). Vulnerabilities in artificial intelligence and machine learning applications and data (No. INL/RPT-22-66111-Rev000). Idaho National Lab.(INL), Idaho Falls, ID (United States).

[36] Abbas, S. G., Vaccari, I., Hussain, F., Zahid, S., Fayyaz, U. U., Shah, G. A., ... & Cambiaso, E. (2021). Identifying and mitigating phishing attack threats in IoT use cases using a threat modelling approach. Sensors, 21(14), 4816.

[37] Mansouri-Benssassi, E., Rogers, S., Reel, S., Malone, M., Smith, J., Ritchie, F., & Jefferson, E. (2023). Disclosure control of machine learning models from trusted research environments (TRE): New challenges and opportunities. Heliyon, 9(4).

[38] Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union, 207, 2020.

[39] Oluokun, A., Ige, A. B., & Ameyaw, M. N. (2024). Building cyber resilience in fintech through AI and GRC integration: An exploratory Study. GSC Advanced Research and Reviews, 20(1), 228-237.

[40] Saravi, S., Kalawsky, R., Joannou, D., Rivas Casado, M., Fu, G., & Meng, F. (2019). Use of artificial intelligence to improve resilience and preparedness against adverse flood events. Water, 11(5), 973.

[41] Belhadi, A., Mani, V., Kamble, S. S., Khan, S. A. R., & Verma, S. (2024). Artificial intelligence-driven innovation for enhancing supply chain resilience and performance under the effect of supply chain dynamism: an empirical investigation. Annals of Operations Research, 333(2), 627- 652.

[42] Dey, P. K., Chowdhury, S., Abadie, A., Vann Yaroson, E., & Sarkar, S. (2024). Artificial intelligence-driven supply chain resilience in Vietnamese manufacturing small-and medium-sized enterprises. International Journal of Production Research, 62(15), 5417-5456.

[43] Cheng, C. H., Nührenberg, G., & Ruess, H. (2017). Maximum resilience of artificial neural networks. In Automated Technology for Verification and Analysis: 15th International Symposium, ATVA 2017, Pune, India October 3–6, 2017, Proceedings 15 (pp. 251-268). Springer International Publishing.

[44] Modgil, S., Gupta, S., Stekelorum, R., & Laguir, I. (2022). AI technologies and their impact on supply chain resilience during COVID-19. International Journal of Physical Distribution & Logistics Management, 52(2), 130-149.

[45] Modgil, S., Singh, R. K., & Hannibal, C. (2022). Artificial intelligence for supply chain resilience: learning from Covid-19. The International Journal of Logistics Management, 33(4), 1246-1268.

[46] Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018, May). Ai2: Safety and robustness certification of neural networks with abstract interpretation. In 2018 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

[47] Schlappa, M. (2023). Optimizing production processes via improved resilience and state-of-the-art AI technologies (Doctoral dissertation, WHU-Otto Beisheim School of Management).

[48] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Towards proving the adversarial robustness of deep neural networks. arXiv preprint arXiv:1709.02802.