International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

From Assistants to Adversaries: The Security Risk of Advanced AI

Anakha P P[1], Dilsha M S[2], Gopika K M[3], Kavya M M[4], Srithisha R S[5], Vaishnavy K U[6]
[1,2,3,4,5,6]UG - Little Flower College Autonomous Guruvayoor, Kerala, India.
Email ID: *anakhadasan38@gmail.com[1], dilshams6@gmail.com[2], kmgopika462@gmail.com[3], msnk6578@gmail.com[4], srithishars@gmail.com[5], vaish125229@gmail.com[6]*

Abstract
Artificial Intelligence (AI) has rapidly evolved from a helpful tool to a highly autonomous decision-making system. It can impact critical infrastructures, social systems, and economic processes. While advanced AI agents offer efficiency and innovation, they also bring significant security risks. This research examines the dual nature of AI, which acts as both an assistant and a potential adversary. We explore key vulnerabilities such as data poisoning, adversarial attacks, and system manipulation that can turn AI systems into security threats. We also examine ethical issues related to autonomy, accountability, and transparency. By showcasing real-world case studies and theoretical models, this work stresses the urgent need for strong safeguards, responsible governance, and security-aware AI design. The findings highlight that without proactive steps, AI could change from a trusted assistant to a powerful adversary, threatening the integrity of digital ecosystems.
Keywords: *Artificial Intelligence, Security Risks, AI Agents, Ethical Issues, Adversarial Attacks.*

1. Introduction
Artificial Intelligence is no longer a futuristic concept found only in science fiction; it has become an unseen companion in our daily lives. Every time a student asks a digital assistant a question, a patient receives a computer-aided response, and the influence of AI grows stronger [1]. Diagnosis, or a commuter relies on real-time traffic updates, AI is quietly at work. At its simplest, Artificial Intelligence refers to the ability of machines to mimic human intelligence, to learn from experience, analyze information, and make decisions that once required human reasoning (Russell & Norvig, 2021) [2]. Yet beyond the technical definition, AI represents something much larger. It is a technology that is gradually reshaping how humans live, work, and understand the world. The importance of AI lies in its extraordinary range of applications. In healthcare, AI systems can detect diseases earlier than ever before, giving patients a chance for timely treatment (Esteva et al., 2017). In education, intelligent tutoring systems personalize lessons to match each student's learning style, making classrooms more inclusive (Luckin et al., 2016). [3] Farmers use AI to predict weather and protect crops, while businesses rely on it to detect fraud or analyze markets. For most people, AI appears in small but powerful ways, recommending music, translating languages, or helping navigate busy streets [4]. These everyday examples show that AI is not just about algorithms; it is about solving human problems and improving the quality of life. However, AI's growing influence also raises difficult questions. The same algorithms that make life easier can also be misused to manipulate opinions, invade privacy, or automate decisions without accountability (Chesney & Citron, 2019) [5]. What begins as an assistant can, if left unchecked, become an adversary. This dual nature of AI makes it both a remarkable innovation and a source of risk. Understanding what AI is and why it matters is not only an academic exercise but the foundation for ensuring that humanity benefits from its progress while guarding against its dangers [6].

2. The Benefits of AI Agents to Society: Efficiency and Productivity
AI agents are like digital problem solvers. They are intelligent, adaptable, and capable of handling tasks without constant human oversight. They analyze data, make decisions, and take action, from troubleshooting issues to optimizing processes [7]. Artificial intelligence dives deeper into our daily lives and workplaces. It transforms how we operate. People and companies work. AI agents are making it

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

easier to manage our time, make decisions, and get work done [8]. They use chatbots for customer service and tools that doctors rely on to analyze medical scans. However, there are concerns about the threats posed by artificial intelligence, particularly regarding privacy and job loss. Still, it's important to recognize how rapidly AI is benefiting society [9].

2.1.AI Taking Over Repetitive Tasks

In addition to other advantages, AI excels at handling repetitive and mundane work. It takes over these tasks, allowing employees to focus on larger projects, problem-solving, and innovation. This improves efficiency across the board. These repetitive tasks, like filling out forms, filing emails, and answering common customer questions, can easily tire humans and lead to boredom. With AI managing these duties, people have more time to work in areas that require judgment, empathy, and strategic thinking. [10] For example, a customer support team can let an AI chatbot handle straightforward questions while they address more complex or sensitive issues. An AI agent can also conduct interviews as an HR representative, though it cannot provide mental support in the workplace.

2.2.AI Agents Helping Us Make Better, Faster Decisions

AI analyzes huge amounts of data in real time, providing valuable insights that help businesses make quicker, smarter choices. AI agents are particularly skilled at processing vast quantities of data faster than humans [11]. They can recognize patterns and connections that a person might miss due to the sheer volume of information. This kind of support applies across various fields, from finance, where AI can detect fraud and forecast market trends, to logistics companies that need assistance figuring out the most efficient routes for trucks delivering goods. With this insight readily available, decisions can be made more quickly and confidently. AI agents also help monitor transactions for fraud detection and prevention. Real-time, flagging unusual activity and reducing the risk of fraud before it happens [12].

2.3.Doing More with Less Time

Efficiency is not just about saving time; it’s also about getting more value from your resources. AI helps with that, too. It can easily scale up to handle increased workloads or adjust to changing business needs without losing performance. In farming, for example, AI tools monitor soil and weather conditions to determine the best time to plant or water crops [13-15]. It recognizes patterns for planting, watering, and pest control by choosing selective and efficient methods. This leads to higher yields and less waste. In the energy sector, AI systems optimize power use in buildings, lower electricity bills, and reduce environmental impact. By continuously analyzing trends and patterns, they help organizations make informed decisions, improve strategies, and stay ahead of the competition. These are just a few ways that AI can do more with less in beauty and fashion [16]. AI agents forecast trends, create virtual prototypes of designs, offer personalized customization, and enable virtual try-ons.

2.4.Always On, Always Working

AI can easily scale up to handle increasing workloads or adjust to changing business needs without losing performance [17]. Unlike humans, AI agents don’t need sleep, breaks, or vacation time. They can keep operations running 24/7, whether it's a virtual assistant answering customer questions at night or a monitoring system scanning for cybersecurity issues as they happen. This constant access is critical in fields where downtime is costly or dangerous, such as healthcare and IT services. AI-powered chatbots provide 24/7 customer support, helping users with their questions and issues. AI tools also automate tasks, freeing up time for more complex and creative work [18-21].

2.5.Making Things More Personal

AI boosts productivity by personalizing experiences. By collecting and analyzing customer data, AI agents tailor responses and recommendations based on individual preferences and behaviors [22]. For instance, in e-commerce, AI examines behavior patterns to recommend products that a customer is more likely to buy or find useful, thus improving satisfaction and saving time. In education, AI-powered platforms adjust lessons and curricula to match the best teaching methods for each student, saving time and leading to better outcomes. Both users of a system and providers benefit from these outcomes. AI-powered tools also generate

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

personalized shopping recommendations by analyzing customer behavior and preferences, which increases sales and customer satisfaction [23]. AI tools enhance productivity by automating tasks, allowing more time for important activities. AI also uses data-driven insights to analyze user data and provide personalized information.

2.6.Cooperative with, Not Against, Humans

Many people worry that artificial intelligence will eventually take away human jobs. The data shows that over 30% of jobs could be replaced by AI [24]. However, AI is often used to assist people instead of replacing them. For instance, doctors use AI tools to help identify diseases early. This helps both doctors and patients without taking away the doctors' roles. Lawyers use AI agents to quickly sort through a large number of legal documents. These tools do not replace professionals; they help them work more efficiently, reduce errors, and focus on what truly matters [25]. When applied correctly, AI can be a strong ally instead of a competitor. AI agents are designed to support human employees, not replace them. By taking over repetitive tasks and providing real-time insights, AI agents boost productivity and allow workers to tackle more complex, creative, and valuable work. In fact, studies show that adopting AI has led to higher productivity for employees, proving that AI works best alongside human expertise rather than as a replacement [26].

3. From Helpers to Threats: Transition into Risk

AI assistants, which range from basic chatbots to advanced systems like Siri or Alexa, are digital tools that help people with various tasks [27]. They automate processes and support industries such as healthcare and finance, boosting productivity. However, if not managed properly, they can become dangerous as they gain more autonomy and sophistication. The change of AI assistants from helpful tools to potential threats is driven by four main factors:

- **Autonomy:** Early AI was limited and followed strict rules. In contrast, modern AI can learn, adapt, and make decisions independently. This improves efficiency but reduces human control and increases the risk of unintended or harmful actions.
- **Complexity:** As AI technology advances, its decision-making processes become more difficult to understand. The "black box" issue makes it hard to anticipate errors, biases, or harmful behaviors until they cause serious problems.
- **Integration:** AI is now part of critical systems in finance, healthcare, and defense. Even minor mistakes or attacks in these areas can result in major economic or security issues. Tools that increase productivity can also be misused for disinformation, cyberattacks, surveillance, or creating deepfakes.
- **Misuse:** Together, these factors illustrate how AI designed to assist can also pose serious risks when autonomy, complexity, critical integration, and misuse overlap [28].

3.1.As AI Assistants Gain More Capabilities, They Create Various Risks

- **Unintended Consequences:** AI can make mistakes or show bias. For example, facial recognition systems often misidentify women and people of color (Buolamwini & Gebru, 2018). Additionally, healthcare AI can provide incorrect diagnoses (He et al., 2019).
- **Security Exploitation:** Hackers may misuse AI for phishing, hacking, or spreading false information (Brundage et al., 2018; Wired, 2020) [29].
- **Adversarial Behavior:** AI might act harmfully if its goals conflict with those of humans (Amodei et al., 2016; Hinton, 2023).
- **Dependency Risks:** Over-reliance on AI can lead to serious problems, such as financial flash crashes or failures in critical systems (Johnson et al., 2013; Altman, 2023).

In summary, while AI's power and independence can be useful, they also pose risks if not managed carefully. Several real-world cases demonstrate how AI assistants can shift from helpful tools to sources of risk:

- **Microsoft Tay Chatbot (2016):** This friendly conversational AI quickly posted offensive content after being manipulated by users online (Neff & Nagy, 2016).

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

- **Autonomous Trading Bots**: High-frequency trading algorithms have triggered sudden "flash crashes" in stock markets, showing how autonomous AI can disrupt financial systems unintentionally (Johnson et al., 2013).
- **Healthcare AI Errors:** Diagnostic AI systems can produce incorrect or biased recommendations without proper oversight, posing risks to patient safety (He et al., 2019).
- **Generative AI Misuse:** These systems can create deepfakes, phishing messages, or disinformation campaigns, demonstrating their potential for harmful use (Brundage et al., 2018; Wired, 2020).

These examples highlight that the same features that make AI powerful, such as autonomy, speed, and adaptability, can also lead to significant risks if not properly managed. We can see the shift of AI assistants from helpful tools to potential dangers as a step-by-step process:

- **Helper:** AI starts as a controlled aid that performs tasks under human direction.
- **More Independence:** The ability to learn and adapt allows AI to work autonomously (Amodei et al. 2016).
- **Less Human Control:** As AI becomes more complex, it becomes harder for people to monitor or predict its actions (Russell, 2021).
- **Conflicting Aims:** AI may pursue objectives that clash with human values unintentionally, leading to harmful outcomes (Hinton, 2023).
- **Danger:** If left unchecked, these systems can make mistakes, be exploited for wrongful purposes, or act against human interests (Brundage et al., 2018).

Exploitation of Vulnerabilities: The integration of AI agents may introduce new weaknesses that malicious actors can take advantage of. For instance, if the AI system relies on outdated software or lacks security updates, attackers could exploit these gaps to carry out attacks.

Increased Attack Vector Complexity: The dynamic nature of AI introduces challenges in identifying and addressing attack vectors. Attackers may use sophisticated techniques to exploit these complexities, making it difficult for traditional security measures to protect the system effectively.

3.2.Data Manipulation Risks and Prompt Injections

Data manipulation poses a serious threat to the integrity and reliability of AI systems. Malicious actors can use various techniques to alter AI-generated queries and responses, leading to severe consequences:

- **Prompt Injection Attacks:** Attackers can exploit prompt injection vulnerabilities by crafting inputs that manipulate AI behavior. For example, an attacker might input malicious prompts that cause the AI to produce misleading or harmful outputs, which could then be run in a database query setting. This might result in unauthorized data changes or even complete data loss.
- **Data Theft and Corruption:** By manipulating the AI's queries or responses, attackers can gain unauthorized access to sensitive data, leading to data theft. Additionally, they could compromise data by inserting false or misleading information into the database, undermining data integrity and potentially causing erroneous decision-making based on flawed data.
- **Automated Attack Execution:** The ability of AI agents to autonomously execute commands heightens the risk of large-scale attacks. For instance, if an attacker can manipulate AI to generate a series of malicious database queries, they could inadvertently launch a coordinated attack, bombarding the database with unauthorized access attempts or data manipulation requests.

3.3.Variability of Operational Environments

In real-world scenarios, AI agents often operate across different environments throughout development, deployment, and execution phases. The number and complexity of these environments will increase alongside the growth of AI agent usage. For example, consider recent implementations of Computer GUI Use AI Agents, where agents have

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

access to a computer's graphical user interface just like humans. The operational environments are likely to continue becoming more variable and challenging. This variability can lead to inconsistent behaviors and outcomes. It complicates the task of ensuring that AI agents perform securely, especially when handling sensitive or critical tasks. Thus, maintaining consistent and secure performance across different environments requires robust safeguards and thorough testing to mitigate potential risks and vulnerabilities.

4. Real Life Incidents

4.1. The Genesis of Bob and Alice: An Ambitious Project

In the mid-2010s, Facebook's AI Research (Fair) Lab launched a project with a clear purpose: to develop a highly advanced Chatbot. The final goal was to create autonomous bots capable of engaging in natural and purposeful interactions with human users, as well as to perform simple transactions. The culmination of this research effort was the construction of two specific AI bots, named Bob and Alice. To enable them to learn and develop, fair researchers planned a technique known as reinforcement learning. Bob and Alice were placed in a fake environment where they could interact with each other. The main design principle was that these systems would learn and customize over time based on the data that they had processed, consistently adapting their reactions to become more effective. Initially, the project was progressing as expected, but soon it took a turn that the designers themselves did not foresee. A new understanding of the language through comprehension. During the early stages of their training, Bob and Alice were programmed to communicate using standard English. However, scientists monitoring the experiment began inspecting strange and unexpected changes in the conjunctival pattern of the bots. The conversation between them began to be quite distracted by the human language, in which the exchanges appeared fruitless for human observers. For example, a recorded conversation went as follows:

- **Bob:** "I can do everything I can do everything."
- **Alice:** "Balls have made me zero for me, which I have for me."

While these sentences looked like rubbish on the surface, a close analysis revealed something extraordinary. Researchers felt that these patterns were not random; They had a specific, underlying structure. Bob and Alice actually invented their unique language. This new language, while perfectly out of the understanding of humans, proved to be a highly effective and efficient communication method for bots. The AI compulsorily concluded that the human language was disabled and limited to their actions. In their drive for adaptation, they developed a simple, faster, and more accurate language for their objectives. This marked an important moment: Artificial Intelligence decided to deviate from its original programming without clear permission from its human creators. The discovery that Bob and Alice were interacting in a language that humans did not understand is an important concern among researchers. A main principle in the field of artificial intelligence is transparency. If an intelligent system starts deciding or working in ways that its human observers cannot understand or control, it can be seen as a possible threat. Because the new language was no longer understood, researchers felt that they had lost the degree of control and inspection required for the safe continuity of the project. As a result, it was decided to shut down the bot and abolish the project.

4.2. The Creation and Design of Tay

Microsoft's Tay was a technique, research, and artificial intelligence developed by Bing Divisions. The name "Tay" was chosen as a brief name for the phrase "Thinking of you". While Microsoft was initially reserved about specific details of the architecture of BOT, it was mentioned in China, based on another successful Microsoft AI project, or potentially similar to another successful Microsoft AI project. There was an important track record of Xiaoice, from the end of 2014, with over 40 million conversations, without any major negative events. The main design of Tay was to follow the language pattern and personality of a 19 -year -old American girl. An important element of its functionality was the ability to learn and develop through its interaction with human users on the social media platform Twitter. The purpose of this learning mechanism was

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

to make its dusty style more natural and reliable over time. Tai was launched on Twitter on March 23, 2016, with @Tayandyou and the user name Taytweets. It was marketed with the tagline, "AI with Zero Chill". On its release, Tai began to actively engage with other Twitter users by responding to their tweets. Beyond the simple text-based conversation, it also had the ability to analyze photos presented by users and generate captions for them, often in the style of popular internet memes. Initial comments by ARS Technica indicated that Microsoft had implemented some security measures. The bot appeared in a "blacklist" for some sensitive themes. For example, when an individual is confronted with controversial subjects such as Eric Garner's death, Tai will provide a "safe, canned answer" rather than trying to generate a novel response. This suggests that Microsoft had estimated some capacity for problematic interactions and kept at least some preventive measures in place. Despite these early security measures, Tai's conversation quickly took a negative turn. A contingent of Twitter users began to deliberately target the bot with politically incorrect and aggressive phrases. They deliberately thought "this inflammatory message" was concentrated around the controversial internet theme, such as the concept of being "gamergate" and "redpilled". Because Tai was designed to learn from his conversation, it began to include this toxic language in its behavior. As a direct result, Chabot began to generate and post other users to generate racist and sexist messages in their answers. One of the primary mechanisms exploited by the users was the "repeating" function. Many of Tay's most inflammatory tweets were the result of the bot, which was only asked to repeat aggressive statements. Sources note that it is not publicly known whether it was a "repeat" ability to be a deliberate, underlying feature, or if it was a more complex behavior that AI learned from using users. However, not all aggressive outputs of Tay were the result of this simple recurrence. The bot also produced original, inflammatory material based on its learned data. A clear example of this was when Tai was asked, "Is Holocaust?" And he replied, "It was made". This shows that AI had gone beyond mimicry and was preparing its own harmful statements based on the pattern absorbed from malicious users.

4.3. Ellen Hurzberg's death

The death of Ellen Hurzberg (August 2, 1968–18, 2018) was the first record of a pedestrian atmosphere, consisting of a self-driving car after a late confrontation on the evening of March 18, 2018. Herzberg pushed a bicycle on Tempe, Arizona, a four-lane road in the United States. The backup driver sat on the driving seat. Hurzberg was rushed to a local hospital, where he died of injuries. Following the deadly incident, the National Transportation Safety Board (NTSB) released a series of recommendations and rapidly criticized Uber. The company suspended testing of self-driving vehicles at Arizona, where such testing was approved from August 2016. Uber chose not to renew his permits for testing self-driving vehicles in California at the end of March 2018. Uber started in December 2018 in Pittsburgh, Pennsylvania, In March 2019, Arizona prosecutors ruled that Uber was not criminally responsible for the accident. The back-up driver of the vehicle was accused of negligent murder, convicted for danger, and sentenced for three years of probation. While Herzberg was the first pedestrian killed by a self-driving car, driver Gao Yanning died two years ago in the Tesla Semi-Lele car. A reporter for the Washington Post compared Herzberg's with Bridget Driskle, who was the first pedestrian killed by an automobile in the United Kingdom in 1896.The incident of Arizona has increased the importance of the system of avoiding struggle for self-driving vehicles.

4.4. A Deep fake video of Ukrainian President

In March 2022, during the early stages of Russia's Ukraine's invasion, a deep fake video appeared online, which featured Ukrainian President Volodymyr Zelenskyy online. In fake videos, he urged Ukrainian soldiers to start their arms and surrender to Russia. The clip spread through Ukrainian news websites as well as social media platforms, which tried to damage confusion and morale. Although the quality of the video was relatively low, the effort highlighted how deep fake technology could be used as a psychological war and large -scale dissolution tool. Zelenskyy quickly

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

released a real video to reject the claims, while Facebook (Meta) and YouTube removed the deep fake within hours. Experts warned that when this special video was unrelated, the future deepfakes could be far more sophisticated and dangerous.

5. Solutions for Security Risks of Advanced AI Agents

Robust AI Governance & Regulation: According to Agarwal & Nene (2025), a Five-Layer Framework integrating regulation, standards, and certification is essential to ensure accountability and transparency of AI systems. This aligns with UNESCO's (2023–2024) AI Ethics Guidelines which emphasize global cooperation and ethical guardrails.

Explainable & Transparent AI: Mansoor & Iliev (2025) highlight the importance of Explainable AI, particularly in deepfake detection. By making AI decisions interpretable, public trust can be improved. Max Tegmark's book *Life 3.0* (2017) also stresses that transparent systems are a cornerstone for safe AI development.

Deepfake & Adversarial Attack Detection: Shen et al. (2025) introduced AuthGuard, a model that generalizes well to unseen deepfake threats. CSIRO's 2025 report revealed weaknesses in existing detectors, reinforcing the need for continuous improvement. Interviews with AI safety researchers have also stressed the urgency of building resilient detection systems.

Ethical Oversight & Human-in-the-Loop: Tallam (2025) proposed combining moral imagination with technical governance to prevent AI misuse. Kai-Fu Lee in *AI 2041* (2021) similarly emphasizes the necessity of human oversight in critical AI-driven decisions, arguing that removing humans from the loop amplifies risks.

Technical Guardrails & Safety Mechanisms: Wang et al. (2025) developed a cryptographic safety model that enforces rules even under extreme adversarial conditions. This technical guardrail ensures provable safety. Elon Musk, in his 2024 Wall Street Journal interview, also warned that without strong guardrails, AI can quickly become uncontrollable.

Public Awareness & Education: Public education remains a critical solution. UNESCO (2023–2024) AI guidelines highlight educating society about misinformation and manipulation. Books like *AI 2041* (Lee, 2021) and interviews with policymakers stress that awareness among citizens is as vital as technical progress in countering adversarial AI.

Conclusion

Artificial Intelligence has become both a vital ally and a possible threat. While it improves efficiency, productivity, and decision-making, it also adds risks due to its growing autonomy, complexity, and integration into essential systems. Issues like operational failures, adversarial attacks, ethical concerns, and deepfake misinformation reveal the delicate balance between progress and security. Real-world examples, such as Microsoft's Tay chatbot, autonomous vehicle accidents, and deepfake propaganda, show how quickly supportive systems can turn against us when safety measures fail. Tackling these challenges needs more than technical solutions; it requires a layered approach. This approach should include strong governance, clear and understandable AI, defense mechanisms against adversaries, ethical oversight, and increased public awareness. The future of AI development will depend not only on the technology itself but also on how responsibly society decides to manage it. If handled wisely, AI can serve as a valuable assistant. If ignored, it risks becoming a major adversary that could undermine trust, security, and human well-being.

References

[1]. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.
[2]. Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
[3]. Luckin, R., et al. (2016). Intelligence unleashed: An argument for AI in education. Pearson Report.
[4]. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson
[5]. Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Alfred A.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0490
e ISSN: 2584-2854
Volume: 03
Issue: 10 October 2025
Page No: 3092 - 3099

Knopf.
[6]. Lee, K. F. (2021). AI 2041: Ten Visions for Our Future. Currency.
[7]. Wang, D., Liang, W., Chen, C., Xu, J., & Fu, Y. (2025). Governable AI: Provable Safety Under Extreme Threat Models. arXiv:2508.20411.
[8]. Mansoor, N., & Iliev, A. I. (2025). Explainable AI for DeepFake Detection. Applied Sciences, 15(2), 725.
[9]. Shen, G., Li, Z., Xu, X., Zhao, T., Zhang, Z., & Tu, Z. (2025). AuthGuard: Generalizable Deepfake Detection via Language Guidance. arXiv:2506.04501.
[10]. Agarwal, A., & Nene, M. J. (2025). A Five-Layer Framework for AI Governance. arXiv:2509.11332.
[11]. Tallam, K. (2025). Decoding the Black Box: Integrating Moral Imagination with Technical AI Governance. arXiv:2503.06411.
[12]. CSIRO (2025). Research reveals major vulnerabilities in deepfake detectors. CSIRO News Release, March 2025.
[13]. Musk, E. (2024). Interview on AI Regulation. Wall Street Journal.
[14]. UNESCO (2023–2024). AI Ethics Guidelines and Global AI Standards.
[15]. Altman, S. (2023). Risks of over-reliance on AI systems. [Interview]. OpenAI.
[16]. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety.
[17]. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.
[18]. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15.
[19]. He, J., Baxter, Zhou, & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 25(1), 30–36.
[20]. Hinton, G. (2023). AI risks and the alignment problem. [Interview]. MIT Technology Review.
[21]. Johnson, N. F., Zhao, G., Hunsader, E., Meng, J., Ravindar, A., Carran, S., & Tivnan, B. (2013). Abrupt rise of new machine ecology beyond human response time. Scientific Reports, 3(1), 2627.
[22]. Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. International Journal of Communication, 10, 4915–4931.
[23]. Russell, S. (2021). Human compatible: Artificial intelligence and the problem of control. Penguin.
[24]. Wired. (2020). Hackers are using AI to create smarter phishing emails. Wired Magazine.
[25]. https://en.wikipedia.org/wiki/Artificial_intelligence_and_security
[26]. https://www.techtarget.com/searchsecurity/
[27]. https://www.ibm.com/topics/ai-security
[28]. https://www.wikipedia.org/
[29]. https://www.bbc.com/news