IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2

AI is Powerful But Risky: Security, Reliability and Ethical Challenges in Modern AI Systems

Sonu Prajapati
MCA Department
Institute of Distance and Open Learning, Mumbai

Abstract: *Artificial Intelligence (AI) has emerged as a transformative technology that enhances automation, efficiency, and decision-making across industries such as healthcare, finance, cybersecurity, and education. AI systems are capable of processing large volumes of data, identifying complex patterns, and providing intelligent outputs with minimal human intervention. However, despite its advantages, AI introduces several critical risks that raise concerns about its widespread adoption. The study follows a qualitative research methodology by reviewing recent research papers and reports from 2024–2025. It also includes real-world case studies related to cyberattacks in financial systems, healthcare institutions, and government infrastructures..*

Keywords: Artificial Intelligence, AI Risk, Cybersecurity, Model Drift, Explainable AI, Data Privacy, AI Ethics

1. INTRODUCTION

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are designed to think, learn, and make decisions. These systems are capable of performing tasks that typically require human intelligence, such as problem-solving, pattern recognition, language understanding, and decision-making.

Over the past decade, AI has experienced rapid growth due to significant advancements in machine learning, deep learning, and big data technologies. The availability of large datasets, increased computational power, and improved algorithms has accelerated the development and adoption of AI across various domains.

Today, AI has become an integral part of modern digital systems and plays a critical role in transforming industries. In the banking sector, AI is widely used for fraud detection, risk assessment, and customer service automation. In healthcare, AI supports disease diagnosis, medical imaging analysis, and predictive healthcare systems, improving both accuracy and efficiency. In cybersecurity, AI enables real-time threat detection, anomaly identification, and automated response systems, helping organizations defend against increasingly complex cyberattacks. These applications demonstrate the powerful capabilities of AI in enhancing productivity, reducing human effort, and enabling intelligent decision-making.

Despite its numerous advantages, the rapid adoption of AI introduces several significant challenges and risks. One of the major concerns is security, where AI systems can be targeted by adversarial attacks, data poisoning, and hacking attempts that manipulate system outputs. Another critical issue is reliability, as AI models may produce incorrect or inconsistent predictions due to factors such as overfitting, biased data, or changing real-world conditions (model drift). Additionally, ethical concerns have emerged, including bias in decision-making, lack of transparency, and misuse of AI technologies such as deepfakes and automated surveillance systems. These challenges highlight the need for careful design, monitoring, and regulation of AI systems.

Given these concerns, it is essential to study both the strengths and risks associated with AI. The primary objective of this research is to analyze the power of AI technologies while identifying key risks and challenges related to security, reliability, and ethics. Furthermore, the study aims to examine real-world cyber threats involving AI systems and propose effective mitigation strategies to ensure safe and responsible AI deployment.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, April 2026 Impact Factor: 8.2

In conclusion, while AI offers transformative benefits and has the potential to revolutionize multiple industries, it also presents serious risks that must be addressed. A balanced approach that combines innovation with strong security measures, ethical considerations, and regulatory frameworks is necessary to ensure that AI technologies are used safely and effectively for the benefit of society.

II. LITERATURE REVIEW

The literature on Artificial Intelligence (AI) highlights both its transformative capabilities and the growing concerns regarding its security, reliability, and ethical implications. Several foundational and recent studies provide insights into the risks associated with AI systems.

Dario Amodei (2016) identified several concrete problems in AI safety, including unintended consequences, reward hacking, and scalability issues in oversight. His work highlighted that AI systems may behave unpredictably if not properly aligned with human intentions, raising concerns about long-term reliability and control.

Recent studies from 2024–2025 indicate a significant shift in the threat landscape:

• AI-generated cyberattacks have increased, with attackers using AI to automate phishing, generate malicious code, and conduct large-scale attacks.

• AI security has become a major research focus, with increasing efforts toward developing robust, secure, and explainable AI systems.

• Studies also show that generative AI models can unintentionally produce harmful or biased outputs, further emphasizing the need for governance.

III. RESEARCH METHODOLOGY

This research adopts a structured qualitative approach to analyze the capabilities and risks associated with Artificial Intelligence (AI), with a focus on security, reliability, and ethical challenges.

A. Research Type

The study is based on Qualitative Research Methodology, which involves analyzing existing literature, reports, and case studies rather than conducting numerical or experimental analysis.

This approach is suitable for understanding complex issues such as AI risks, where interpretation and comparison of findings are required.

B. Data Sources

The research is conducted using secondary data collected from reliable and peer-reviewed sources to ensure accuracy and credibility. The primary data sources include:

• IEEE Xplore Digital Library o Provides high-quality research papers on AI, cybersecurity, and machine learning.

• Springer Journals o Contains academic articles focusing on AI security, reliability, and emerging technologies.

• ResearchGate o Used to access recent research publications and preprints related to AI risks and applications.

• Government Reports (CERT-In) Official reports from the Indian Computer Emergency Response Team provide real-world data on cyber threats, malware attacks, and security incidents.

These sources align with academic standards and ensure compliance with research quality guidelines.

IV. POWER OF AI

Artificial Intelligence (AI) has become a transformative technology that enhances efficiency, productivity, and decision-making across various industries. Its power lies in its ability to

automate tasks, process large volumes of data, make intelligent decisions, and scale globally. This section discusses the major strengths of AI systems in detail.

A. Automation

AI-based automation is widely used in:

• Industrial manufacturing (robotic assembly lines)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2

• IT operations (automated testing, deployment)
• Customer service (chatbots, virtual assistants) By automating routine processes, organizations can:
• Reduce human errors Improve productivity
• Save time and operational costs

B. Data Processing Capability

AI systems are capable of processing large volumes of structured and unstructured data, often referred to as Big Data. Traditional systems struggle with such scale, but AI can efficiently analyze and extract meaningful insights.

Key capabilities include:
• Pattern recognition
• Data classification
• Predictive analytics AI improves accuracy by:
• Reducing human bias in data analysis
• Identifying hidden patterns
• Providing data-driven insights Applications:
• Fraud detection in banking
• Recommendation systems (e-commerce, streaming platforms) Medical data analysis

C. Intelligent Decision Making

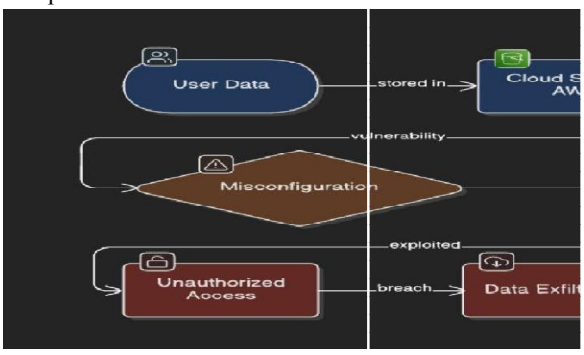
AI-driven decision-making is used in:
• Healthcare: Disease diagnosis, treatment recommendations
• Finance: Credit scoring, fraud detection
• Cybersecurity: Threat detection and prevention Advantages:
• Faster decision-making
• Improved accuracy
• Real-time analysis

V. AI SYSTEM ARCHITECTURE

Artificial Intelligence (AI) systems follow a structured lifecycle that ensures efficient data handling, model development, deployment, and continuous improvement. The architecture consists of multiple stages, each playing a critical role in building a reliable and scalable AI system.

A. Overview of AI Architecture

The AI system architecture is a pipeline that transforms raw data into actionable insights through a sequence of well-defined steps.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2

B. Data Collection

Data collection is the first and most important step in the AI lifecycle. AI systems rely heavily on data to learn patterns and make predictions.

Sources of Data:

- Databases (SQL, NoSQL)
- Sensors and IoT devices
- APIs and web scraping
- User-generated data

C. Data Preprocessing

Raw data is often incomplete, inconsistent, or noisy. Preprocessing ensures that the data is clean and suitable for model training.

Key Steps:

- Data cleaning (removing duplicates, handling missing values)
- Data transformation (normalization, encoding)
- Feature selection (choosing relevant attributes)

Outcome:

- Improved data quality
- Better model performance

A. Financial Sector & Data Theft

Case Study 1: Banking Data Breach (2025 Trend)

In 2025, financial institutions experienced large-scale cyberattacks where attackers targeted cloud- based infrastructures and APIs to steal customer data.

Attack Method:

- Exploitation of cloud misconfigurations (e.g., unsecured storage buckets)
- API vulnerabilities
- Credential theft

Data Compromised:

- Customer records (KYC data)
- Bank account details
- Transaction history

Impact:

- Identity theft
- Financial fraud
- Loss of customer trust

AI Risk Link:

- AI systems processing financial data became targets
- Lack of secure AI pipelines increased vulnerability

B. Android Malware Campaigns

Case Study 2: Fake “RTO Challan.apk” Scam (CERT-In Alert 2025)

A large-scale phishing campaign targeted Android users using fake APK files disguised as official government applications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, April 2026 Impact Factor: 8.2

Attack Flow:

1. Victim receives SMS/WhatsApp link
2. Downloads fake APK ("RTO Challan.apk")
3. App requests permissions (SMS, contacts)
4. Malware steals OTP and banking data

AI Role:

• AI used to generate realistic phishing messages
• Automated targeting of victims

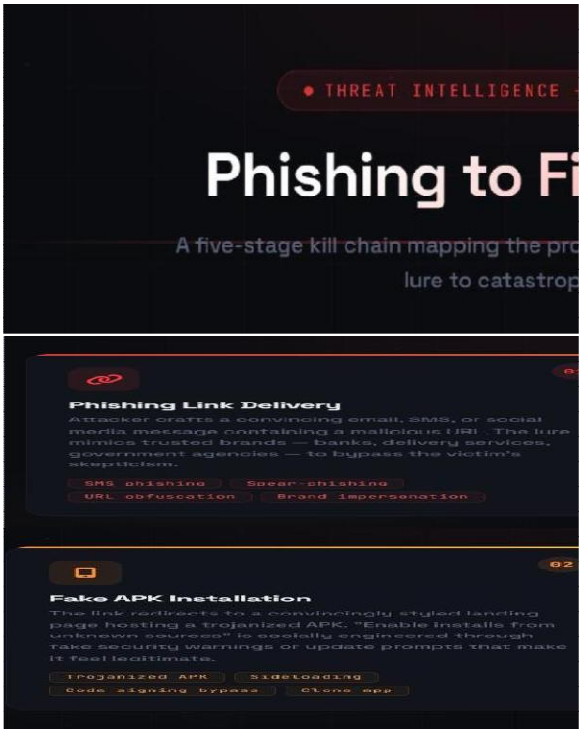
Impact:

• Unauthorized bank transactions
• Financial loss
• Personal data theft

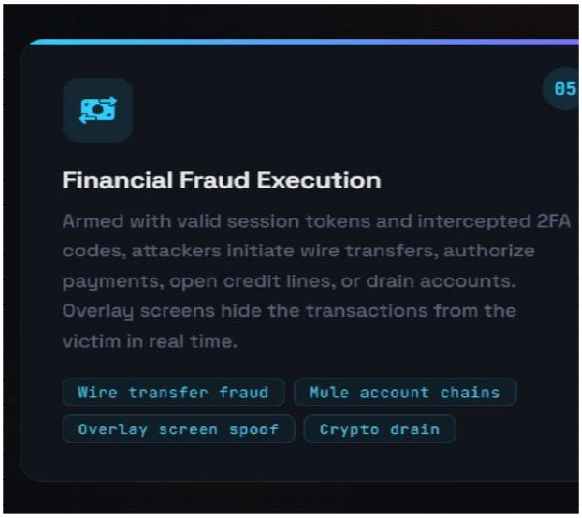
Attack Diagram

Phishing Link → Fake APK Install → Permission Access → Data Theft → Financial Fraud

Phishing Link → Fake APK Install → Permission Access → Data Theft → Financial Fraud



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2



C. Ransomware Attacks

Case Study 3: Hospital Ransomware Attack (India, 2025)

Healthcare institutions faced ransomware attacks where attackers encrypted patient data and demanded ransom payments.

Attack Process:

1. Phishing email or vulnerability exploited
2. Malware deployed in hospital systems
3. Files encrypted
4. Ransom demanded

Impact:

• Patient data inaccessible
• Delay in medical services
• Risk to human lives

AI Risk Link:

• AI-based systems in healthcare disrupted
• Critical dependency on automated systems increased risk

Ransomware Flow

System Access → Malware Execution → File Encryption → System Lock → Ransom Demand

D. Critical Infrastructure Attacks

Case Study 4: Government & Infrastructure Attacks (2025)

Cyberattacks targeted government websites and critical infrastructure such as power grids and telecom systems.

Attack Types:

• DDoS (Distributed Denial of Service)
• Website defacement
• Network probing Impact:
• Service disruption
• National security threats
• Public panic

AI Risk Link:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, April 2026 Impact Factor: 8.2

• AI systems used for infrastructure monitoring can be targeted
• Attackers use automation to scale attacks
DDoS Diagram Botnet Traffic → Target Server → Overload → Service Failure

E. AI-Generated Cyber Threats
Case Study 5: AI-Powered Phishing Attacks (2025)
AI tools are used to generate highly convincing phishing emails and messages.
Features:
• Personalized content
• Human-like language
• Automated targeting
Example:
• Fake bank emails requesting login details
• Deepfake voice scams
Impact:
• Increased success rate of scams
• Difficult to detect attacks

Real-World Case Studies
This section presents real-world case studies demonstrating how AI-related risks affect different sectors. These examples highlight practical implications of security threats, privacy breaches, and misuse of AI technologies in modern systems.

A. Financial Sector Attacks
The financial sector is one of the most targeted domains due to the high value of sensitive data such as banking credentials, transaction records, and personal identification details.

1. Data Breaches
Financial institutions have experienced large-scale data breaches where attackers gain unauthorized access to customer information.
Causes:
• Weak authentication mechanisms
• API vulnerabilities
• Insider threats
Impact:
• Exposure of customer data (KYC, account details)
• Identity theft and fraud
• Loss of customer trust

2. Cloud Misconfiguration
Modern banking systems rely on cloud platforms for data storage and processing. Misconfigured cloud resources (e.g., unsecured storage buckets) can lead to data exposure.
Example Scenario:
• Publicly accessible cloud storage without proper security settings
• Sensitive financial data leaked online
Impact:
• Data leakage
• Financial loss
• Regulatory penalties

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2

B. Android Malware Attacks

Mobile devices have become a primary target for cybercriminals due to widespread usage and access to financial applications.

1. Fake APK Attacks

Attackers distribute malicious applications disguised as legitimate apps (e.g., government or banking apps).

Attack Flow:

1. User receives phishing link
2. Downloads fake APK file

VIII. THREAT MODEL DIAGRAM

The threat model represents the potential risks and attack vectors associated with an AI system. It helps in identifying different types of attackers, their methods, and the possible impact on the system. This structured representation is essential for designing secure and reliable AI systems.

A. Overview of Threat Model

An AI system can be targeted by multiple entities, including external attackers, insiders, and adversaries. Each entity exploits different vulnerabilities to compromise the system.

A. Threat Actors

1. Adversary

An adversary is an external attacker who intentionally manipulates the AI system to produce incorrect or harmful results.

Attack Methods:

- Data poisoning
- Adversarial input attacks

Impact:

- Incorrect predictions
- System manipulation

2. Insider

An insider refers to authorized individuals (employees, developers) who misuse their access privileges.

Attack Methods:

- Data leakage
- Unauthorized data sharing

Impact:

- Confidential data exposure
- Loss of trust

3. External Attacker

External attackers target the AI system without having direct access.

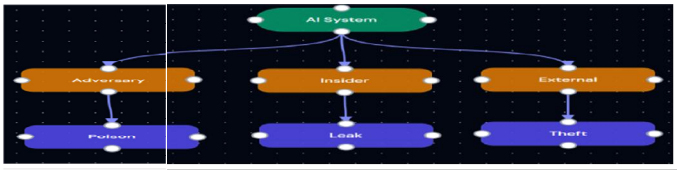
Attack Methods:

- API exploitation
- Model theft
- Network attacks

Impact:

- Intellectual property loss
- System compromise

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2



C. Types of Threats

1. Data Poisoning (Poison)

• Injection of malicious data into training datasets
• Leads to incorrect model behavior

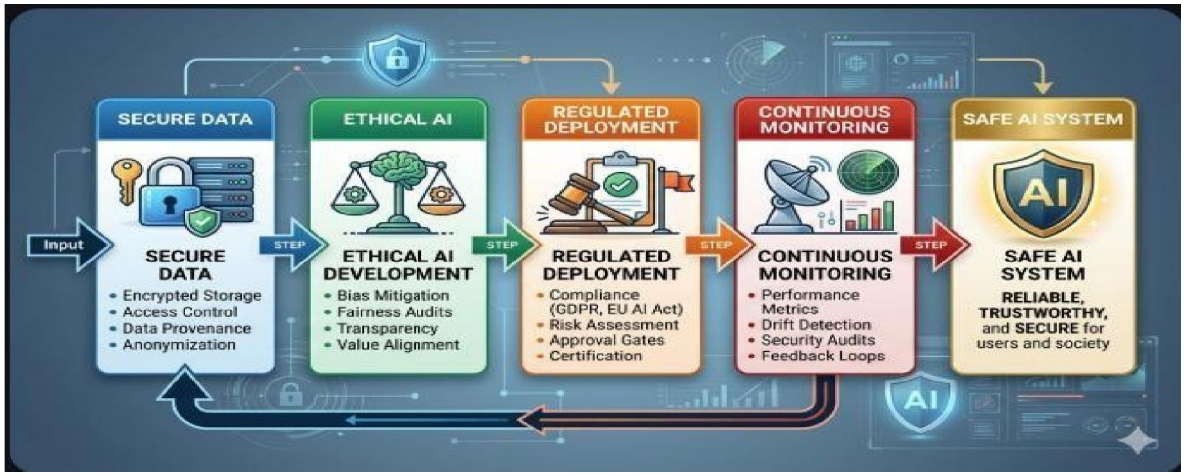
2. Data Leakage (Leak)

• Unauthorized exposure of sensitive data
• Occurs due to weak security controls

3. Model Theft (Theft)

• Extraction or duplication of AI models
• Exploiting APIs or system vulnerabilities

Future Safe AI Architecture



F. How AI Will Become Safer in the Future

1. Strong Governance

• Governments will enforce strict AI regulations
• Mandatory compliance and audits

2. Built-in Security

• Security integrated into AI development lifecycle
• Protection against adversarial attacks

3. Continuous Monitoring

• Real-time performance tracking
• Early detection of failures and threats

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, April 2026
Impact Factor: 8.2

4. Human Oversight

- Human-in-the-loop systems
- Critical decisions reviewed by experts

5. Responsible Innovation

- Balance between innovation and risk control
- Ethical guidelines for developers

IX. CONCLUSION

Artificial Intelligence (AI) is a transformative technology that enhances efficiency, automation, and decision- making across sectors like healthcare, finance, and cybersecurity. Its ability to process large data and generate intelligent insights makes it essential for modern digital systems. However, AI also introduces significant risks, including security threats (adversarial attacks, data breaches), privacy concerns, reliability issues (model drift), and ethical challenges such as bias and lack of transparency. Therefore, the safe adoption of AI requires strong security measures, data protection, continuous monitoring, and explainable models, along with
effective government regulations. In conclusion, a balanced approach combining innovation with ethical and regulatory frameworks is necessary to ensure that AI remains both powerful and safe for society.

ACKNOWLEDGMENT

The author expresses sincere gratitude to the faculty members and the institution for their continuous guidance, support, and encouragement throughout the course of this research work.
Their valuable insights, constructive feedback, and academic mentorship have played a crucial role in the successful completion of this study.
The author also acknowledges the contributions of researchers, scholars, and organizations whose published work and reports provided a strong foundation for this research. Access to academic resources, journals, and technical materials greatly supported the analysis and understanding of the subject.
Finally, the author extends appreciation to all individuals who directly or indirectly contributed to this work, making this research both meaningful and comprehensive.

REFERENCES

[1]. Uddin et al., Generative AI in Cybersecurity, Springer, 2025.
[2]. Okdem & Okdem, AI in Cybersecurity, MDPI, 2024.
[3]. Wen et al., AI Security Assurance Review, Springer, 2024–25.
[4]. Swetha et al., AI for Cybersecurity, Springer, 2025.
[5]. Akhtar & Rawol, AI-Powered Security, 2024.
[6]. Ofusoria et al., AI in Cybersecurity Review, 2024.
[7]. Merlano, AI & ML in Cybersecurity, 2024.
[8]. CERT-In Report, Government of India, 2025.
[9]. NASSCOM AI Report, India, 2025.
[10]. Okdem et al., 2024
[11]. Wen et al., 2025
[12]. CERT-In Report, 2025
[13]. NASSCOM Report, 2025