ISSN: 2581-8775

JANUARY 2025

VOL. 05, ISSUE NO. 02

FOR FREE DISTRIBUTION

ISTM JOURNAL OF TRAINING RESEARCH AND GOVERNANCE

Institute of Secretariat Training & Management
An ISO 9001:2015 Institution
Certified 'UTKRISHT' by Capacity Building Commission
New Delhi - 110067

ISTM JOURNAL OF TRAINING RESEARCH AND GOVERNANCE

ISSN 2581-8775

Volume No. 05, Issue No. 02

For Free Distribution

Institute of Secretariate Training & Management

An ISO 9001:2015 Institution

New Delhi - 110067

"Printed by Deepak Kumar Bist, Joint Director, Published by Namita Malik, Joint Director on behalf of Institute of Secretariat Training and Management, Administrative Block, JNU Campus (Old), Olof Palme Marg, New Delhi – 110067 and Printed at **Guru Nanak Enterprises, New Moti Nagar, Delhi.**"

Chief Editor Shri Rajiv Manjhi, Director, ISTM

Table of Content

ABOUT THE INSTITUTE

The Institute of Secretariat Training and Management (ISTM), established in 1948, serves as a premier capacity-building institution aimed at meeting the dynamic human development needs of government and support institutions across the nation. With a focus on sustainable, innovative, and contemporary methods, ISTM's motto is "Efficiency and Public Good."

ISTM is responsible for implementing the cadre Training Plan for the Central Secretariat Service (CSS) and Central Secretariat Stenographers Service (CSSS), as well as other services within the Central Secretariat. The Institute also provides orientation training for Group 'A' officers joining the Central Government under the Central Staffing Scheme as Deputy Secretaries and Directors. Additionally, ISTM offers training to officers from the Central and State Governments, Union Territory Administrations, Public Sector Undertakings, Autonomous Bodies, and various other formations within Central/State/UT administrations. The Institute organizes foundation courses for several Group 'A' services, including the AIS, IES, ITS, ICLS, IFOS, and IIS, among others.

ISTM has established the first-ever Centre of Excellence for Civil Services Competency (COE-CSC) under the National Programme of Civil Services Capacity Building - Mission 'Karmayogi'. It has become a leader among Central Training Institutes in the development of e-content for Mission Karmayogi. The Karmayogi digital learning lab at ISTM aims to produce digital content for use by various Ministries, Departments, and Organizations, enabling government officers to enhance their skills from their offices and homes.

ISTM conducts Management Development Programmes in diverse areas such as Financial Management, Vigilance, Administration, Management Principles, Good Governance, Knowledge Management, Behavioral Techniques, Cabinet Note Preparation, Infrastructure Development, Big Data Analysis, and Gender Sensitization. These programs aim to orient government officers towards effective service delivery.

Specific competencies developed through these trainings include strategic financial planning, vigilance and ethical governance, administrative efficiency, leadership, and decision-making, good governance practices, data analysis skills, gender sensitivity, and digital literacy.

The Institute also engages in research and consultancy work for capacity building in governance. It collaborates with client institutions on Training Need Analysis, HR Administration, Design of Training, Cadre Review/Restructuring, and the Audit of Proactive Disclosure under the RTI Act, 2005.

ISTM is led by an officer at the level of Joint Secretary to the Government of India, appointed under the Central Staffing Scheme. Faculty members are appointed on deputation based on their experience and qualifications. The Institute has developed in-house expertise in facilitating skill development and behaviour modification to enhance organizational effectiveness. ISTM is envisioned to play a crucial role in the capacity-building initiative of Mission Karmayogi by strengthening its professional capacity and developing the framework for the Role-Based Competency Model. The Institute of Secretariat Training and Management has been recently certified as "उत्कृष्ट" by Capacity Building Commission.

ABOUT THE JOURNAL

Aligned with the objectives outlined in the National Training Policy (2012), which emphasize the importance of networking with institutions to share resources and engaging in field studies and research, the Institute of Secretariat Training and Management (ISTM), New Delhi, presents the bi-annual 'ISTM Journal of Training, Research, and Governance'.

This journal serves as a pivotal platform within the realm of public administration, training, and development. It endeavours to foster a culture of continuous learning by disseminating best practices, innovative methodologies, and cutting-edge techniques essential for cultivating proficient civil servants dedicated to serving society.

As a pioneering initiative by ISTM, this bi-annual publication represents a compendium of scholarly literature and academic insights meticulously curated for the enlightenment and enrichment of government officials, institutions, and researchers. Its scope encompasses a diverse array of subjects, including public policy, government operations, and human capital development.

'ISTM Journal of Training, Research, and Governance' is a refereed Journal and hence, ISTM ensures to maintain the high quality standard avoid any bias in the review process and the publication and therefore the 'Double Blind' peer review process is followed. Before publication, all content published in the journal, go through a critical quality check process to maintain the credibility of the journal.

The contents of the journal showcase contributions from esteemed theorists, government practitioners, seasoned academicians, and erudite scholars, offering profound insights and practical perspectives on various facets of training and professional practice. Through this initiative, ISTM endeavours to foster a culture of knowledge sharing and intellectual discourse, thereby nurturing a cadre of highly skilled and enlightened civil servants poised to address the challenges of contemporary governance.

EDITORIAL BOARD

Disclaimer:-

The articles published in this Journal are printed originally as provided by the author(s). Hence, the complete ownership and responsibility remains with the author(s) and not with the publisher, editor(s) or any of the officers of this institution.

From the Desk of Chief Editor

Dear Readers,

It is with great pleasure that I present to you Volume 5, Issue 2 of the ISTM Journal of Training, Research, and Governance, centered around the timely and crucial theme: "AI and Cybersecurity – Threats and Opportunities in Public Policy Making."

Artificial Intelligence (AI) is revolutionising governance and public policy, enhancing decision-making, automation, and data-driven insights. However, alongside its vast potential, AI brings significant cybersecurity risks, from deepfake disinformation campaigns to digital fraud, necessitating a robust policy framework to mitigate threats while capitalizing on opportunities.

This issue brings together eight insightful articles that explore the multifaceted relationship between AI and cybersecurity in governance.

The opening article, "Artificial Intelligence and Financial Fraud: Digital Arrest Scams" by Pawan Kumar Shrivastav and myself jointly, sheds light on the increasing exploitation of AI for digital frauds, highlighting the need for proactive policy interventions.

Next, "Adaptive AI Governance: Strategic Frameworks for Ethical, Inclusive, and Equitable Policymaking in India" by Pranjal Jain et al., presents a governance roadmap that balances AI advancements with ethical and regulatory frameworks, ensuring responsible AI adoption in public administration.

In "Smart & Secure: AI-Powered Reliable Data Transfer for Air-Gapped Networks", Pravin K. Choudhary and Heli Nandani delve into AI-driven methods to enhance cybersecurity in critical government infrastructures, addressing vulnerabilities in classified data transfer mechanisms.

The issue also features "Ethical AI in Public Policy Making" by Ms. Garima, which underscores the necessity of fairness, transparency, and accountability in AI-driven governance models.

A key concern in cybersecurity is software vulnerability. "Enhancing Software Security: A Study of Simplified Secure Software Development Framework" by N. Satyanarayana et al. introduces a structured approach to embedding security in software development to reduce cyber risks.

As AI becomes integral to national security, "Strengthening Cyber Security Through AI" by Himanshu Kumar Raghav explores AI-driven intelligence frameworks for strengthening national cybersecurity strategies through capacity building in this direction.

Conti....

From the Desk of Chief Editor

The financial sector is particularly vulnerable to AI-induced risks. “Policy Considerations in Use of Artificial Intelligence in India: A Case for Financial Sector” by Rohit Chawla and Abhinav Banka examines the regulatory landscape for AI applications in finance, highlighting governance mechanisms for stability and trust.

Lastly, “AI-Generated Deepfakes: The Emerging Cybersecurity Threat and Countermeasures” by Dr. Aslam et al., discusses the alarming rise of AI-generated deepfakes and their implications for misinformation, national security, and public trust.

These articles offer invaluable insights into how AI can be both a powerful tool and a potential threat in the realm of public policy and governance. As AI and cybersecurity challenges continue to evolve, policymakers, researchers, and practitioners must work together to craft resilient, ethical, and forward-looking frameworks that protect national security while enabling technological progress.

I extend my sincere gratitude to all authors, reviewers, and contributors for their efforts in making this issue a rich compilation of research and thought leadership. We encourage our readers to engage with these discussions, share their perspectives, and contribute to the ongoing dialogue on AI and cybersecurity in governance.

We look forward to your valuable feedback and suggestions for future editions.

Best Regards,
Rajiv Manjhi, Chief Editor
ISTM Journal of Training Research and Governance

ARTIFICIAL INTELLIGENCE AND FINANCIAL FRAUD: DIGITAL ARREST SCAMS

RAJIV MANJHI[1], PAWAN KUMAR SHRIVATAV[2]

[1]DIRECTOR, ISTM AND JOINT SECRETARY TO GOVT. OF INDIA
[2]ASSISTANT LIBRARY AND INFORMATION OFFICER, ISTM

ABSTRACT

India has been experiencing a major increase in cybercrimes through digital arrest scams that utilise cutting-edge technology and psychological tricks. The research examines as to how the scammers pretend to be police officers with the aim of extorting money from victims by falsely accusing them of made-up crimes. The research utilises a qualitative methodology through case studies and thematic analysis along with various information sources such as government reports and media accounts. Scammers who target vulnerable groups use advanced AI tools like voice cloning and deep fake technology to give their fraudulent activities greater believability. Victims experience deep emotional as well as financial distress which frequently results in enduring psychological trauma. The study ends by proposing new public awareness initiatives and strengthened cybersecurity protections to address this escalating danger.

1.0. Introduction

Anyone could be a potential victim, even if he/she is well-studied; or occupies middle or a top-level position in government or private organisation, and are digitally literate too, you have a higher probability of being on the list of the potential target of 'Digital Arrest' like cybercrime. If you are less emotionally intelligent too then you may top the list.

In the past few years, the world in general and India in particular have been witnessing a continuous surge in cybercrimes, especially "Digital Arrest Scams" emerging as one of the most sophisticated forms of fraud. "In 2024, India faced significant losses from digital arrest scams, with ₹1,777 crore lost in the first four months alone." (Barua, 2024) These scams are performed using advanced technologies and psychological manipulation of the victim to extort huge amounts of money from victims by falsely claiming they are under arrest for fabricated criminal activities. The scammers typically act as some law enforcement officials such as senior officers belonging to Police Department, Central Bureau of Investigation; Enforcement Directorate; and/or Judicial Officers during video calls, creating a sense of urgency and unbridled fear that compel victims to comply with their demands. This paper explores the mechanisms behind these scams, the role of AI in enhancing their effectiveness, and strategies to be followed for prevention. The severity of digital crimes can be assessed from the address of the Prime Minister of India elaborating on the **'Digital Arrest'** scams in India and making people aware of it, he mentioned that "there is no system like digital arrest in the law, this is just a fraud, deceit, it is a lie, a gang of criminals and those who are doing this are enemies of society. To deal with the fraud that is going on in the name of digital arrest" (115thEpisodeof'MannKiBaat', 2024).

2.0. Literature Review

2.1. Cybercrime in India

Cybercrime in India has become a pressing issue, we have experienced a rapid increase in incidents due to its expansion in the digital framework. According to the National Crime Records Bureau (NCRB) reports of 2020, 2021 and 2022 (the latest reports available on the official website of NCRB as of date i.e. 29-Dec-2024), cybercrime cases in India have been significantly increasing every year, even the intention of the cybercrime as 'Fraud' has also been recorded in increasing order for each year (NCRB 2020, 2021 & 2022, 2022). Refer Table 1 for a detailed overview. The exceptional data as reflected in the table shows a drastic increase of 63.5% in cyber crimes during the COVID-19 lockdown throughout the country, during this period, other crimes were recorded in decreasing order.

SL. No.	Year of Record	No. of Cyber Crimes	Change %age	Motive of Fraud out of total cyber crime
1	2019	44,735	63.50%	60.40%
2	2020	50,035	11.80%	60.20%
3	2021	52,974	5.90%	60.80%
4	2022	65,893	24.40%	64.80%

(TABLE 1: FIGURES OF CYBERCRIME EXTRACTED FROM NCRB REPORTS 2019, 2020, 2021 & 2022)

Every new invention comes with its pros and cons. The expansion of digital technologies has introduced new opportunities for criminals to exploit system vulnerabilities and manipulate individuals. The lack of awareness among the general public regarding cybersecurity measures has contributed significantly to the proliferation of such scams. The cybercriminals do not only succeed due to ignorance of victims about their digital privacy but also due to the vulnerability of emotional and psychological tricks contribute to their success. It also does not end with financial losses to the victims but leads to emotional and psychological trauma. In her Phd. thesis (Ahe, 2022) says "The psychological impact may be measured with the concepts of anxiety, fear, anger, shame/embarrassment, and self-esteem after the victimisation of cyber fraud."

2.2. Digital Arrest Scams

The digital arrest scams represent a unique subset of cybercrime that leverages technology to create official law enforcement proceedings. Recent studies indicate that these scams have become increasingly prevalent, targeting "individuals across various demographics, including senior citizens, by exploiting their trust and unfamiliarity with digital communication" (Abhinandan, 2024). The psychological tactics used by scammers often involve pretending to be one of the law enforcement authorities with a sense of urgency, which can lead even the most cautious and digitally educated individuals to fall victim. "The 'digital arrest' scam involves fraudsters impersonating law enforcement officials on video calls, threatening victims with arrest over fake charges, and pressuring them to transfer large sums of money." (Biswas BBC, 2024).

3.0. Role of Artificial Intelligence

In the present cybercrime scenarios, Artificial intelligence is being used as an important tool for cybercriminals. These

criminals use several AI tools available on the internet to increase their probability of success. They use AI technologies to analyze various social media profiles, gather exact personal information about the potential victims, and create realistic impersonations through voice cloning and creating visuals through deepfake technology. The scammers have become more sophisticated in their methods; they also understand the role of AI in the success of their countermeasures. However, the study also highlights that AI tools are not only used by criminals but also by law enforcement agencies and other big companies such as master card to combat such crimes. “Gen AI, of course, isn’t only being used by fraudsters. And low-tech solutions aren’t always the answer. Mastercard’s teams today are using gen AI and AI tools to improve the security of the digital world.” (Chauhan, Rohit;, 2024). “AI-powered open-source intelligence tools (including open-source intelligence tools for social media) accelerate open-source investigations through the power of AI.”(Bureau of Police Research and Development, 2022). The dual-use nature of AI technology presents both challenges and opportunities in addressing digital fraud.

4.0. Methodology

This study employs a qualitative research approach to comprehensively analyse the growing menace of digital arrest scams in India. The research methodology includes case study analysis, content analysis, and thematic analysis of various sources, including news reports, government publications, cybersecurity reports, and academic literature. The study aims at analysing the mechanisms of these scams, the role of AI in their execution, and potential countermeasures for prevention and mitigation.

4.1. Research Design

The research follows an exploratory and descriptive design, focusing on real-life instances of digital arrest scams. By examining case studies, patterns of fraud, and the technological tools leveraged by scammers, this study seeks to expose the modus operandi of the cybercriminals and the psychological tactics they employ.

4.2. Data Collection Methods

Data have been collected from a variety of sources to ensure a comprehensive and multi-perspective analysis:

4.2.1. Government and Law Enforcement Reports: Reports from agencies like the National Crime Records Bureau (NCRB) and Bureau of Police Research and Development (BPRD) were reviewed to understand official statistics and policies related to cybercrime.

4.2.2. Media and News Articles: Leading news agencies, including the BBC, India Today, and Business Standard, provided real-life accounts of victims, helping to illustrate the tactics used by scammers.

4.2.3. Academic and Industry Research: Peer-reviewed journal articles and reports from cybersecurity firms (such as those from MasterCard and AI ethics research organizations) were analysed to understand the technical aspects of AI-driven scams.

4.2.4. Case Studies: A series of documented digital arrest scam cases were examined to identify patterns, psychological manipulation techniques, and the role of AI in enhancing these scams.

4.3. Data Analysis Techniques

4.3.1. Thematic Analysis: Recurring themes such as impersonation, psychological manipulation, financial coercion, and AI-powered fraud techniques were identified and categorized.

4.3.2. Comparative Analysis: Cases were compared to detect similarities in scam

tactics, AI applications, and victim responses, providing deeper insights into the evolving nature of digital arrest fraud.

4.3.3. Pattern Recognition: AI-driven scam elements, such as deepfake usage and voice cloning, were systematically analysed across different incidents to understand their role in increasing scam success rates.

4.4. Ethical Considerations

Given the sensitive nature of cybercrime research, ethical considerations were strictly followed:

4.4.1. No personal or identifying information of victims has been disclosed beyond what is already available in public domain.

4.4.2. The study relies on verified sources to maintain accuracy and credibility.

4.4.3. Recommendations has been framed with a focus on victim protection, public awareness, and cybersecurity enhancements to minimize any potential harm.

5.0. The Mechanics of Digital Arrest Scams

5.1. Initial Contact

The scam typically begins with a phone call or message from an individual pretending to be a law enforcement officer or government official. Victims are often informed that they are under investigation for serious crimes such as money laundering or drug trafficking. Scammers create a sense of urgency by threatening immediate arrest unless victims comply with their demands. (Biswas BBC, 2024). These scammers would review and analyse the social media profiles and all the personal information of the probable victim before establishing contact. For example, they may reference recent activities or personal details of the victim that establishes credibility. “The scammers knew Mr. Mukhopadhyay was a journalist and writer - author of a biography on Prime Minister Modi. They knew Dr. Tondon was a doctor and had attended a conference in Goa.” (Biswas BBC, 2024).

5.2. Psychological Manipulation

Scammers employ psychological tactics to manipulate victims into compliance. By leveraging fear and confusion, they pressurise victims into making hasty decisions without verifying the legitimacy of the claims. “The digital arrest scam is a carefully planned act that plays on fear and urgency.” (Abhinandan, 2024). The research indicates that emotions play a significant role, “On the call, the scammer escalates the tension, using personal information like the victim’s name, ID number, or address to appear credible. They then claim the victim is involved in serious crimes, like money laundering or tax evasion, to increase anxiety.” (Abhinandan, 2024). These criminals use this psychologically anxious situation of the victim as an opportunity to create panic and the victim feels isolated during these trapped situations, which further increases the probability of compliance with the instructions given by the scammers.

6.0. Case Studies

The real-life cases of digital arrest scams have been used to illustrate the severity, complexity, and evolving nature of these cybercrimes. By analysing these incidents, we can identify common patterns, psychological tactics, and the role of AI in executing such frauds. The selected cases highlight how scammers manipulate victims across different demographics, regardless of education, professional background, or social status.

6.1. Case Selection Criteria

The cases were chosen based on the following criteria:

6.1.1. Diversity of Victims: Victims include all the genres such as professionals, retired officials, and senior citizens to show that these scams do not discriminate based on education, financial literacy, or digital awareness.

6.1.2. Scam Complexity: Cases involve varying levels of cleverness, from basic impersonation to AI-driven tactics like deepfake video calls, document forging and voice cloning.

6.1.3. Financial and Psychological Impact: Each case demonstrates the severe financial losses and emotional trauma inflicted on the victims.

6.1.4. Publicly Available Information: Only verified and publicly reported cases were included to maintain ethical integrity.

6.2. Detailed Case Analyses

6.2.1. Dr. Bhuyan a Public Health Specialist: Dr. Bhuyan, a Joint Director at Government Organisation and a Public Health Specialist could not manage to escape the trap and lost 50,000/- Indian rupees. The scam started with an unknown call from a courier company and surprisingly, Dr. Bhuyan has been expecting a courier which he ordered from an e-commerce website. In his word, He might have lost a bigger amount, if his boss Dr. (Ms.) Jain had not called him urgently. By looking at his face she understood that he was under some pressure, and enquired as to why he had been awestruck in dejection. Dr. Bhuyan mentioned about the episode and Dr. (Ms.) Jain immediately dubbed this as digital arrest. Thus, talking to his boss about his digital arrest rescued him from the cyber criminals. (Personal Interview by Rajiv Manjhi and Pawan Kumar Shrivastav, March, 2025)

6.2.1.1. Psychological Manipulation: Dr. Bhuyan was made to believe that the scammers were from a investigating agency and were handling his recently ordered parcel, which allegedly consisted of objectionable items.

6.2.1.2. Use of AI: In this case, the scammers used social engineering to obtain Dr. Bhuyan's private real-time information and created investigating office environment.

6.2.1.3. Outcome: After several hours, Dr. Bhuyan could get out of this fabricated treat due to his urgent duty call.

6.2.2. Dr. Tondon's Trial: Dr. Tondon, a 44-year-old neurologist from Lucknow, was coerced into transferring nearly ₹2.5 crore ($300,000) through a meticulously orchestrated scam. The criminals first posed as telecommunications officers, alleging that her Aadhaar number was linked to an illegal transaction. This was escalated when fake law enforcement officials joined the video call, presenting fabricated evidence and threatening her with legal action (Biswas BBC, 2024).

6.2.2.1. Psychological Manipulation: Scammers isolated Dr. Tondon from her friends and family members, instructing her not to disclose the situation.

6.2.2.2. Use of AI: Advanced deepfake technology was likely used to make the impersonated officials appear genuine.

6.2.2.3. Outcome: After six days of continuous psychological pressure, she transferred funds across multiple accounts, losing her entire life savings.

6.2.3. Major General Puri (Retd.): A retired 83-year-old Major General from Panchkula, Haryana, was tricked into believing he was under investigation by the Central Bureau of Investigation (CBI) (India Today Team, 2024).

6.2.3.1. Scam Execution: The criminals staged a fake virtual courtroom with actors playing judges, police officers, and forensic experts. The scam involved official-looking documents, fake warrants, and a realistic

video call setup that mimicked a government office.

6.2.3.2. Financial Loss: Overwhelmed by the situation, he transferred ₹82.27 lakh ($100,000+) under the pretext of a refundable security deposit.

6.2.3.3. AI's Role: The scammers cloned official documents, set up of the courtroom and used video manipulation to enhance credibility.

6.2.3.4. Outcome: By the time the truth was discovered, the money was irretrievable, leaving him emotionally and financially devastated.

6.2.4. Ms. Mishra Journalist from News Channel: A senior journalist from a top Indian news channel became a victim of a scam that started with a FedEx package fraud and escalated into a digital arrest threat (India Today, 2024).

6.2.4.1. Tactic Used: Scammers impersonated law enforcement officers, falsely claiming that she was involved in illegal smuggling activities.

6.2.4.2. Psychological Control: She was held in a digital hostage situation for two hours, forced to remain on video calls and provide updates every two hours under threat of arrest.

6.2.4.3. AI Usage: Possible voice cloning for generating police stations like background noise and deepfake elements were used to make the scam convincing.

6.2.4.4. Outcome: While she narrowly escaped financial loss, the psychological toll of the ordeal was severe.

6.2.5. High-Profile Gujarati Man: A 90-year-old businessman from Gujarat was deceived by scammers posing as CBI officers, who accused him of money laundering and drug trafficking. (Singh Business Standards, 2024). The incidents highlight that no one is immune to these scams regardless of their background or education level.

6.2.5.1. Scam Execution: Over multiple calls, scammers fabricated legal charges and demanded payment to clear his name.

6.2.5.2. Financial Loss: ₹1 crore ($120,000) was transferred before the victim realized the fraud.

6.2.5.3. AI's Role: Fake official documents and real-time voice modulation were used.

6.2.5.4. Outcome: The victim was left emotionally shattered and financially drained.

6.2.6. Prominent Author and Journalist Mr. Mukhopadhyay: A well-known journalist and author received a call claiming that his 20-year-old bank account was linked to a money laundering case. In his own words, Mr. Mukhopadhyay said "I became a digital slave until my friends exposed the scam. I had moved my funds into my account, ready to transfer it all to them. I felt like a fool when it was over." (Biswas BBC, 2024). These criminals at the initial step psychologically break the victim, disabling him to think and process the information smartly, resulting in a puppet-like show over a continuous video call.

6.2.6.1. Psychological Manipulation: He was manipulated into believing he was under surveillance and was about to transfer his entire savings until a friend intervened and exposed the scam.

6.2.6.2. Use of AI: Scammers knew their details, likely gathered through AI-driven social media analysis.

6.2.6.3. Outcome: Though he avoided financial loss, the mental distress was overwhelming.

6.3. Key Insights from Case Studies

6.3.1. AI Amplifies Scams: All cases demonstrate the increasing use of AI for voice cloning, deepfake video calls, document forging and personal data analysis, making scams more convincing.

6.3.2. Emotional Manipulation is Key:

Victims were pressured into compliance through fear, urgency, and social isolation.

6.3.3. Education and Status Do Not Guarantee Protection: Victims included doctors, journalists, military officials, and elderly businessmen, proving that even highly educated individuals can fall prey. It was also noticed that all victims were digitally literate

6.3.4. Financial Impact is Severe: Losses ranged from ₹82 lakh to ₹2.5 crore, showing the high stakes of digital arrest scams.

7.0. The Role of AI in Enhancing Digital Arrest Scams

Artificial intelligence plays a dual role in digital fraud—both as an enabler and a countermeasure. Fraudsters exploit AI's capabilities in three main ways: (1) Automated Social Engineering, where AI analyses & extracts public profiles to tailor highly personalized scam messages; (2) Deepfake Technology, which enables scammers to convincingly impersonate law enforcement officials; and (3) Real-Time AI Voice Cloning, allowing scammers to generate lifelike conversations that mimic legal & rational authority. The increasing accessibility of AI tools, such as open-source deepfake generators and voice synthesis software, pose a significant challenge for cybersecurity experts. To counteract these threats, it is need of the hour for the law enforcement agencies to adopt AI-driven fraud detection techniques, such as machine learning-based anomaly detection and biometric authentication systems.

7.1. Automated Social Engineering for Data Collection and Analysis

Scammers use AI technologies to collect large amounts of personal information from social media and online activities. They then use this data to create strategies based on individual weaknesses. For instance, they might mention specific details about a person's life or job to appear more trustworthy. The criminals may also use photographs with opposite gender to blackmail and keep the victim matted. Studies show that AI can quickly process huge amounts of data to spot patterns, helping scammers identify people most likely to fall for their tricks. This targeted method increases the success of scams and avoids wasting time on unlikely targets.

7.2. Deepfake Technology

AI has made scams more advanced. Scammers now use video calls for fake interrogations, applying deepfake technology to create realistic impersonations. They analyze social media profiles across different platforms to gather personal information. They also use tools like voice-changing software to disguise their voices during calls or send pre-recorded messages customized to the victim. This level of detail makes it harder for people to recognize a scam.

7.3. Challenge of Digital Forensics in AI-Powered Crimes

The biggest challenge to combating AI-driven scams is digital forensics. Unlike traditional cybercrimes, AI-enhanced scams leave few traces behind, which makes it harder for authorities to trace the criminals. Traditional fraud detection systems depend on metadata, IP tracking, and financial transaction analysis, yet deepfake-based fraud bypasses most of these detection methods. Law enforcement agencies should be investing in AI-based forensic tools that not only identify manipulated or altered media but also flag such media for detection, as well as for inconsistencies in synthetic voices and locations using AI-powered predictive analytics. Additionally, international cooperation is critical, as cybercriminals frequently function across borders, exploiting jurisdictional shortcomings to avoid apprehension.

8.0. Preventive Measures and Recommendations

8.1. Public Awareness Campaigns

Awareness generation about the digital arrest could be the first steps towards prevention. Government initiatives should focus on disseminating information through various channels—social media platforms being particularly effective given their reach among younger and older demographics. Public service announcements should emphasize key strategies for avoiding scams:

8.1.1. Using Pseudo Names: For any unknown call the individual may introduce himself with a pseudo name. This will help in confusing the scammer and put the criminal into panic.

8.1.2. Verify Identity: Encourage individuals never to disclose personal information without verifying identities through official channels. This means using trusted, authorized, and legitimate methods to verify the identity, such as the organisation's website, customer service hotlines, or official emails, rather than unverified sources like random phone calls, messages, or social media.

8.1.3. Recognise Red Flags: Make the citizens aware to identify common tactics used by scammers—including threats or high-pressure tactics.

8.1.4. Identify Psychological Manipulation: Make the citizens aware of identifying commonly used tactics for psychological manipulations

8.1.5. Report Suspicious Activity: Create an environment, where reporting suspicious calls or messages is normal and at the ease of the victim or the probable victim.

8.1.6. Seek Immediate Help: In case of any such situation, when one feels trapped in any mysterious position, he/she must talk about it to the family and friends.

8.2. Strengthening Cybersecurity Measures

Collaborative efforts between government agencies and private organisations are essential for creating a unified response against cybercrimes, including digital arrest. It is also advised that for every major transaction and any transfer to a newly added account or a foreign bank account, a warning should be issued, before processing the transaction, regarding the fake digital arrest scams occurring in India. Financial institutions and businesses must enhance cybersecurity measures to protect consumers from fraudulent activities. This includes implementing robust identity verification processes and conducting regular audits to identify vulnerabilities within systems (Abhinandan, 2024).

8.3. Legal Frameworks Addressing Cybercrime

While India's cybercrime laws have come a long way, they still aren't able to catch up with the fast-evolving technology, especially concerning AI-based frauds like digital arrest scams. Though current legislation may offer limited aid, it simply does not address contemporary breaches of security wherein AI, deep fake and psychological coercion tactics are weaponised. Battle digital arrest scams effectively — by strengthening existing laws and introducing regulations specific to AI.

8.4. Information Technology Act 2000

The primary legislation governing cybercrime in India is the Information Technology Act (IT Act), enacted in 2000—providing legal recognition for electronic transactions while establishing penalties for various offences related thereto. However, critics argue that it lacks specificity concerning newer forms of fraud such as those involving AI technologies or

deepfakes—leaving loopholes exploited by criminals. Even the NCRB reports do not have a categorisation of various cybercrimes including Digital Arrest (NCRB 2020, 2021 & 2022, 2022).

8.5. Global Strategies to Combat AI-Powered Scams

Countries have taken control and enforced strong measures against AI-based fraud. The European Union also has an AI Act that would enact strict rules for high-risk uses of AI, including deep fakes used for fraud (EUROPEAN UNION, 2024). The U.S. Federal Trade Commission (FTC) has proposed various measures to prevent the use of AI to engage in cybercrimes (Federal Trade Commission, 2025). India may also learn from these models through its strengthened AI regulation, prescribed transparency in AI-generated content, and equipping law enforcement agencies with AI-enabled cybersecurity tools.

9.0. Proposed Amendments in Legal Framework and Policy Enhancement

In response to rising concerns over cybersecurity threats, the Indian government proposed amendments aimed at strengthening existing laws while introducing new provisions tailored towards combating emerging challenges posed by advanced technologies in 2021. These amendments seek not only harsher penalties but also enhanced cooperation between law enforcement agencies across jurisdictions—a critical factor given how easily perpetrators can operate across state lines using online platforms (MEITY, MHA;, 2021). Since AI and Digital Arrest scams have boomed since 2022, to effectively tackle digital arrest scams and AI-driven cybercrimes, the following legal and policy changes should be introduced:

9.1. AI and Deepfake Regulation Act:

A new law should be enacted specifically to prevent deep fake frauds, voice cloning, and AI-generated impersonations.

9.2. Amendments to the IT Act, 2000:

Define and criminalise AI-driven scams and introduce stricter penalties for using deepfake technology in fraud. Also, section 66D (cheats by impersonating someone else using electronic means) must be strengthened.

9.3. Faster Cybercrime Trials:

Special cybercrime courts should be established to expedite legal procedures for digital fraud cases.

9.4. Cross-Border Cybercrime Treaties:

Strengthen international cooperation with global agencies like INTERPOL and cybersecurity organisations for better enforcement against foreign cybercriminals.

9.5. AI Detection Tools for Financial Transactions:

The financial institutions must use AI detection tools to flag suspicious transactions linked to any potential scams.

9.6. NCRB Cybercrime Data Expansion:

Introduce separate categories for AI-driven scams and Digital Arrest Scams in crime reports to track trends and develop targeted prevention strategies.

10.0. Conclusion

Digital arrest scam has emerged as a major threat within the cybercrime landscape owing to artificial intelligence operating with heightened sophistication. This paper clearly portrays that scammers and fraudsters are able to outsmart the law enforcing agencies and exploit the behavioural and emotional fault-lines

appearing among the general masses owing to their ignorance, emotional vulnerability and social decency.

This paper also establishes that digital scammers are well educated people and, thus, it reinforces the need for adoption of value-based education system.With growing inclination and emphasis on technical education, the product (students) are likely to lose the sheen of values in life which might mechanize the whole fabric of the social and family relationship.

This also calls for the stakeholders such as government agencies including law enforcement agencies to work, in tandem, with cyber security professionals with the objective to put in place a strong preventive mechanism against the mushrooming threat of cybercrime.Campaigns to educate the public as how to foresee scams and report to the concerned agency with a view to combat the cyber threats. The use of AI technologies to create protective measures functions as a double-edged sword because criminals exploit them for fraud but law enforcement agencies can utilize them for crime detection and prevention with efficiency.

Hence, addressing digital arrest scams requires a multifaceted approach that combines technology, education and community engagement to safeguard individuals against these insidious threats.

11.0. References

- *Abhinandan. (2024, NOVEMBER 15). What Is The "Digital Arrest" Scam & How To Avoid It? (AUTHBRIDGE) Retrieved DECEMBER 21, 2024, from AUTHBRIDGE: https://authbridge.com/blog/what-is-the-digital-arrest-scam/*
- *Ahe, L. (2022, June 29). Mental Wellbeing and Cybercrime (The Psychological Impact of Cybercrime on the Victim). University of Twente , Department of Positive Clinical Psychology and Technology. Twente: University of Twente . Retrieved Dec 29, 2024, from https://essay.utwente.nl/: https://essay.utwente.nl/91014/*
- *Barua, K. (2024, December 18). Digital Arrest in 2024: How Much Hyderabad, Karnataka and Other States Loses. Retrieved December 29, 2024, from Jagran Josh: https://www.jagranjosh.com/general-knowledge/digital-arrest-cases-in-2024-states-loses-in-this-scam-check-details-here-1734093437-1*
- *Biswas BBC, S. (2024, November 18). 'You are under digital arrest': Inside a scam looting millions from Indians. Retrieved December 30, 2024, from https://www.bbc.com: https://www.bbc.com/news/articles/cdrdyxk4k4ro*
- *Bureau of Police Research and Development. (2022). Artificial Intelligence for Law Enforcement. In AI in the Service of Law Enforcement (p. 30). New Delhi, Delhi, India: Bureau of Police Research and Development. Retrieved December 29, 2024, from https://bprd.nic.in/uploads/pdf/AI%20in%20the%20service%20of%20Law%20Enforcement-%20a%20n%20Introduction.pdf*
- *Chauhan, Rohit;. (2024, October 28). Why gen AI is a double-edged sword in cybersecurity. (M. Card, Producer) Retrieved December 29, 2024, from Master Card: https://www.mastercard.com/news/perspectives/2024/why-gen-ai-is-a-double-edged-sword-in-cybersecurity/?cmp=2024.q4.glo.glo.all.brand.purp.others.gen-ai-cybersecurity.602201.sep.txt.googles.ai%20in%20cybersecurity&gad_source=1&gclid=CjwKCAiAg8S7BhATEiwAO2-*

- *EUROPEAN UNION. (2024, June 13). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations. Official Journal of the European Union. Luxembourg, Luxembourg, LUXEMBOURG: European Union. Retrieved January 31, 2025, from https://eur-lex.europa.eu/eli/reg/2024/1689/oj#document1*
- *Federal Trade Commission. (2025, January 03). AI and the Risk of Consumer Harm. (S. i. Practices, Editor) Retrieved January 31, 2025, from Federal Trade Commission: https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2025/01/ai-risk-consumer-harm*
- *India Today Team. (2024, December 02). Beware the new big con. (I. T. Team, Editor, & I. Today, Producer) Retrieved December 29, 2024, from India Today: https://www.indiatoday.in/magazine/cover-story/story/20241202-beware-the-new-big-con-2637649-2024-11-22*
- *India Today, V. (2024, October 12). Digital arrest: Journalist's ordeal reveals how scammers turn homes into prisons. (I. T. Desk, Editor, & I. Today, Producer) Retrieved December 29, 2024, from India Today: https://www.indiatoday.in/newsmo/video/digital-arrest-journalists-ordeal-reveals-how-scammers-turn-homes-into-prisons-2615808-2024-10-12*
- *MEITY, MHA;. (2021). The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. New Delhi: Ministry of Electronics and Information Technology. Retrieved January 30, 2025, from https://www.meity.gov.in/writereaddata/files/Revised-IT-Rules-2021-proposed-amended.pdf*
- *Modi, N. (2024, October 27). PM's address in the 115th Episode of 'Mann Ki Baat'. Mann Ki Baat. New Delhi, Delhi, India: Prime Ministers Office. Retrieved December 29, 2024, from https://www.pmindia.gov.in/en/news_updates/pms-address-in-the-115th-episode-of-mann-ki-baat/*
- *NCRB 2020, 2021 & 2022. (2022, December 01). Crimes in India 2020, 2021 & 2022. National Crime Records Bureau, National Crime Records Bureau. New Delhi: National Crime Records Bureau. Retrieved December 29, 2024, from https://www.ncrb.gov.in/: https://www.ncrb.gov.in/crime-in-india.html*
- *Pangotra, A., & Aditi, N. (2024, October 05). Cyber Peace Foundation. Retrieved January 27, 2025, from https://www.cyberpeace.org/: https://www.cyberpeace.org/resources/blogs/rise-in-digital-arrest-scams-a-growing-concern*
- *Singh Business Standards, N. (2024, November 30). Crores lost to 'digital arrest' scam: Here's how to avoid falling victim. Retrieved December 31, 2024, from https://www.business-standard.com/: https://www.business-standard.com/finance/personal-finance/crores-lost-to-digital-arrest-scam-here-s-how-to-avoid-falling-victim-124113000374_1.html*

ADAPTIVE AI GOVERNANCE: STRATEGIC FRAMEWORKS FOR ETHICAL, INCLUSIVE, AND EQUITABLE POLICYMAKING IN INDIA

DR. PRANJAL JAIN[1], POOJA JAIN[2], PROF. (DR.) ANJU JAIN[3]

[1]SCHOLAR, CHARTERED ACCOUNTANT AND COMPANY SECRETARY
[2]PROFESSIONAL CONSULTANT IN GENERATIVE ARTIFICIAL INTELLIGENCE SPACE
[3]PROFESSOR, UNIVERSITY OF DELHI

ABSTRACT

This study explores how India can develop an adaptive AI governance framework that ensures ethical, inclusive, and equitable policymaking? As AI reshapes governance, it is essential to balance innovation with accountability and public trust. This research evaluates India's National Strategy for Artificial Intelligence (AI for All), emphasising the need for dynamic and adaptive governance models. The study identifies key risks such as algorithmic bias, data sovereignty challenges, cybersecurity threats, and the lack of transparency, and proposes countermeasures including iterative regulatory cycles, adaptive policymaking, real-time monitoring, and regulatory sandboxes. It evaluates the importance of citizen participation in ensuring equitable access to AI technologies and promoting fair deployment. By integrating these strategies, this research advocates for a governance framework that prioritizes equity, accountability, and transparency. The findings offer a particular focus on enhancing policymaking in healthcare, agriculture, and smart cities.

Keyword: Artificial Intelligence, Public Policy, Data Driven Governance

1.0. Introduction

Artificial Intelligence (AI) is projected to significantly contribute to India's economy, with an estimated potential to add $957 billion, equivalent to 15% of the current gross value added, by 2035 (Accenture, 2018). The National Strategy for Artificial Intelligence (NSAI) effectively positions AI at the forefront for enhancing outcomes in critical sectors such as healthcare, agriculture, and education. AI's role in scaling the delivery of state services, such as remote medical diagnostics, precision agriculture advisory, as well as improving inclusivity in access to government welfare programs through tools like regional language chatbots and voice interfaces, induces its application in effective policy making. For instance, in the 2019 Prayagraj Ardh Kumbh Mela in Uttar Pradesh, an impressive network of 1,100 AI-enabled CCTV cameras monitored crowd density. They were integrated with Command-and-Control Centres and alerted security personnel whenever threshold limits were breached, exemplifying AI's capacity to enhance public safety through real-time intelligence. Equally innovative is the work of Wadhwani AI, which is piloting a smartphone-based anthropometry tool designed to empower health workers. This AI-powered application enables the screening of low-birth-weight infants without the need for specialized medical equipment, a boon for resource-constrained settings. Three Centres of Excellence in Artificial Intelligence are being set up for Health, Agriculture and Sustainable Cities,

each with a funding of ₹300 crores each. The number of AI Start-ups are on the rise with companies like Haptik, Qure AI and Zoho, etc. gaining international recognitions.

While AI-driven solutions are unlocking transformative opportunities, they also bring complex challenges, particularly around privacy, security, and ethical considerations. Therefore, the success of AI in leveraging vast datasets depends not only on technological advancements but also on strategic policymaking that addresses these multifaceted concerns. One of the foremost challenges in AI-driven data governance is ensuring privacy and data protection. As AI systems rely on extensive datasets, often containing sensitive personal information, safeguarding individual privacy becomes important. The second consideration involves the critical aspect of ethical use of AI, particularly in a socially diverse country like India. AI systems trained on biased or incomplete datasets risk perpetuating or exacerbating inequalities. Algorithmic biases could lead to discriminatory outcomes in areas such as hiring, credit access, or healthcare recommendations. For instance, if an AI model uses data from individuals who have had limited access to formal banking or credit services, it may, due to its training on historically skewed financial behaviors, deem applicants from certain demographic backgrounds as "higher-risk" without fully considering their financial situations or creditworthiness. This can create a cycle of exclusion, where marginalized populations are locked out of essential financial services, reinforcing economic disparities that prevent upward social mobility. Third, the vast amounts of data collected for AI applications make them an attractive target for cyberattacks. Ensuring the resilience of AI systems against such threats requires robust cybersecurity measures and continuous monitoring to safeguard both data and critical infrastructure. AI's increasing role in decision-making necessitates transparency and explainability in its operations. Black-box algorithms, which lack interpretability, pose significant risks, especially in high-stakes domains such as finance, law enforcement, and healthcare. The lack of standardized practices in AI and data management further complicates governance. Without interoperability and consistency across platforms, scaling AI technologies and ensuring seamless collaboration become challenging. India's National Strategy for Artificial Intelligence (AI for All) emphasizes the need for standardized protocols and public-private partnerships to address these gaps and foster innovation. Importantly, data governance frameworks should also account for equitable access to AI technologies and their benefits.

As artificial intelligence (AI) becomes increasingly embedded in governance and public policy, the challenge lies in ensuring that its deployment remains ethical, transparent, and equitable. While AI holds potential to enhance decision-making in healthcare, agriculture, smart cities, and public services, its rapid evolution raises concerns regarding algorithmic bias, data privacy, cybersecurity risks, and lack of accountability. The absence of adaptive governance frameworks exacerbates these risks, making it imperative to establish policies that can evolve dynamically with technological advancements while safeguarding public interest. A key issue that this research addresses is the inadequacy of static, one-size-fits-all frameworks in managing the complexities of AI technologies. Traditional regulatory

models, which often involve rigid rules and one-time interventions, are ill-suited to the fast-evolving nature of AI. AI systems, by their nature, can adapt and evolve through machine learning over a period, raising the potential for unforeseen consequences and risks that static regulations are unlikely to anticipate. As such, adaptive governance is essential, where regulatory approaches evolve in response to the challenges and opportunities AI presents in real-time. Given the risks of data misuse, discriminatory AI outcomes, and the lack of transparency in decision-making processes, this research evaluates the need for robust policies to guide AI implementation responsibly. It emphasizes the importance of equity in AI access, ensuring that the benefits of AI technologies extend to all sections of society, particularly underserved and rural communities.

2.0. Research Framework

The National Strategy for Artificial Intelligence (AI for All), introduced by NITI Aayog, emphasises the application of AI in key sectors such as agriculture, healthcare, education, and infrastructure, with a focus on addressing the country's unique socio-economic challenges. The government's vision is to leverage AI to drive inclusive growth while maintaining ethical standards and data security. Initiatives under this strategy include the establishment of Centers of Excellence (CoEs) for AI research and development. These centers aim to bridge the gap between industry, academia, and policymakers, fostering a collaborative ecosystem where theoretical research translates into practical, scalable solutions. The strategy highlights the importance of international partnerships, and positions India as a "garage" for developing scalable solutions for emerging economies. The Responsible AI for All framework, developed by NITI Aayog, emphasises safeguarding the foundational tenets of ethics, equity, and inclusivity while deploying AI in governance (Fig 1.). The Principle of Safety and Reliability focuses on ensuring that AI systems are robust, dependable, and resilient, particularly when deployed in critical domains such as healthcare, finance, and infrastructure. AI systems must be designed to withstand external threats and internal malfunctions, minimizing risks to users and stakeholders. To achieve this, rigorous testing and validation protocols are emphasized, ensuring that systems perform consistently under diverse operating conditions.

Techniques such as adversarial testing, where AI systems are exposed to simulated attacks, and stress testing, where systems are evaluated under extreme workloads, are recommended to identify vulnerabilities and enhance system resilience. Additionally, fault-tolerant designs are encouraged to ensure that AI systems can recover seamlessly from errors or disruptions. For example, in autonomous vehicles, redundancy mechanisms such as backup sensors and alternative algorithms can prevent catastrophic failures. Real-time monitoring systems that track the performance of AI systems during operation are also proposed, allowing developers to detect anomalies and intervene proactively. The principle further advocates for the incorporation of safety-by-design approaches, where reliability and risk mitigation are prioritized throughout the development lifecycle. By embedding these practices, AI systems can operate reliably even in high-stakes and unpredictable environments.

FIG 1. KEY CONSIDERATIONS FOR INTEGRATING AI IN GOVERNANCE

SOURCE. DEVELOPED FROM RESPONSIBLE AI, APPROACH DOCUMENT FOR INDIA, NITI AAYOG

The Principle of Equality and Inclusivity underscores the need for AI systems to promote fairness, mitigate biases, and ensure accessibility for all segments of society, particularly underserved and marginalized communities. AI technologies must be designed to reflect India's socio-economic diversity, ensuring equitable outcomes across various demographic groups. To address biases in AI, techniques such as adversarial training, which adjusts algorithms to counteract disparities, and inclusive dataset curation, which ensures representation of diverse populations, are emphasized. These technical measures help minimize systemic inequities that may arise from skewed data or unrepresentative training samples. In addition, the principal advocates for the development of AI tools that cater to regional and linguistic diversity. Natural Language Processing (NLP) models tailored to India's 22 official languages and local dialects are crucial to ensuring that AI-powered solutions are accessible to all citizens. For instance, AI-enabled voice assistants that understand and respond in local languages can bridge the digital divide in rural areas. Furthermore, accessibility features such as text-to-speech, haptic feedback, and screen readers are proposed to make AI tools inclusive for individuals with disabilities. By prioritizing inclusivity in design and deployment, this principle ensures that AI technologies contribute to bridging, rather than widening, existing socio-economic divides.

Principle of Privacy and Security commits to safeguarding personal and sensitive information. The framework emphasizes the deployment of advanced privacy-preserving technologies, including differential privacy, homomorphic encryption, and federated learning. Differential privacy ensures individual data points remain unidentifiable while enabling aggregate analysis, a vital feature for applications in healthcare and governance. Homomorphic encryption facilitates computations on encrypted data, enabling secure collaboration between institutions without exposing raw data, while federated

learning enables decentralized AI model training without transferring sensitive information to a central location, a key aspect in sectors like healthcare and financial services. Robust access control mechanisms, regular vulnerability assessments, and cryptographic safeguards are advocated to mitigate risks associated with unauthorized access and data breaches, fostering trust in AI deployments. The Principle of Transparency and Explainability emphasizes the need for AI systems to be understandable to all stakeholders, ensuring trust and accountability in decision-making processes. Model interpretability techniques such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) allow stakeholders to understand the rationale behind AI-driven decisions. For instance, in a credit scoring system, these techniques can help identify the specific factors influencing loan approvals or rejections. The principle also advocates for algorithmic audits to identify biases or errors in AI systems and for the development of Explainable AI (XAI) models, ensuring transparency without compromising accuracy. Complementary to these technical solutions, user-friendly documentation and interactive dashboards are proposed to make AI insights accessible to non-technical users, enabling informed decision-making across diverse sectors.

The Principle of Accountability addresses the challenges of assigning responsibility across the lifecycle of AI systems. It advocates for maintaining decision logs and audit trails, which record every action and decision made by an AI system, enabling traceability and accountability. For high-risk applications, the principle emphasizes the inclusion of human-in-the-loop systems, ensuring critical decisions, such as those in healthcare or law enforcement, are validated by human experts. Liability assignment frameworks are proposed to delineate accountability among developers, operators, and users of AI systems, particularly in cases where self-learning systems adapt autonomously post-deployment. These measures are bolstered by regulatory guidelines that ensure adherence to ethical standards and provide mechanisms for grievance redressal, ensuring affected individuals have access to fair and transparent recourse. The Principle of Positive Human Values highlights the alignment of AI development with societal goals and ethical imperatives. Ethical AI frameworks are emphasized to integrate safeguards against misuse while ensuring AI systems reflect cultural, social, and moral norms through value-sensitive design principles. To address sustainability concerns, the principle encourages the integration of environmental impact assessments into AI development, aiming to reduce the carbon footprint of energy-intensive AI processes like deep learning model training. Bias mitigation algorithms, such as adversarial de-biasing and re-weighting of training data, are recommended to ensure fairness and equity in AI outcomes, preventing the perpetuation of historical biases in datasets. By embedding these human-centric values into AI development, the framework ensures that AI serves as a force for societal good, enhancing equity, empowerment, and sustainability.

The study proposes the Principle of Sustainability which calls for leveraging AI to address environmental challenges and

promote long-term ecological balance. AI systems are envisioned as enablers of sustainable development, optimizing resource use and reducing waste. In agriculture, for instance, AI-powered precision farming techniques use data from IoT sensors and satellite imagery to monitor soil conditions, predict weather patterns, and optimize irrigation schedules, leading to more efficient resource use and higher crop yields. Similarly, AI-driven energy management systems can enhance the efficiency of renewable energy grids by forecasting demand and supply, reducing dependence on fossil fuels. To mitigate the environmental impact of AI itself, particularly energy-intensive processes like deep learning model training, the principle advocates for the use of green computing practices. Techniques such as model compression, which reduces the size of AI models without compromising performance, and energy-efficient hardware design are emphasized to lower the carbon footprint of AI deployments. Sustainability assessments are proposed as a standard practice, ensuring that AI projects align with environmental goals and contribute to global efforts to combat climate change. By integrating sustainability into the development and deployment of AI systems, this principle ensures that technological progress is harmonized with ecological preservation.

The study establishes the Principle of Collaboration which emphasizes the importance of partnerships among government agencies, industry stakeholders, academic institutions, and civil society organizations in driving AI innovation and governance. Collaboration is seen as essential to overcoming technical, ethical, and logistical challenges in AI deployment. Public-private partnerships are particularly highlighted as a mechanism for pooling resources, expertise, and infrastructure to scale AI solutions. For instance, collaborations between tech companies and agricultural research institutions can lead to the development of AI-powered tools that address issues like crop diseases and market price fluctuations. The principle also advocates for international cooperation in setting global AI standards and sharing best practices. Collaborative research programs, such as the establishment of multi-institutional AI centers, are proposed to drive innovation in areas like ethical AI, algorithmic transparency, and bias mitigation. By fostering a culture of collaboration, this principle ensures that AI development is inclusive, interdisciplinary, and aligned with societal priorities. The study provides the Principle of Interoperability to focus on the need for AI systems to integrate seamlessly across platforms, industries, and geographies. This is particularly critical in a diverse and interconnected digital ecosystem like India's.

Standardization of data formats, communication protocols, and APIs is advocated to enable interoperability, ensuring that AI systems from different vendors or domains can work together effectively. For example, interoperability in healthcare systems can allow patient data to flow securely and seamlessly between hospitals, laboratories, and insurance providers, improving care coordination and outcomes. The principle also emphasizes the importance of open-source AI frameworks and platforms that encourage innovation and collaboration. Open standards reduce barriers to entry for start-ups and researchers, fostering a vibrant AI ecosystem. To ensure interoperability at a

global level, India's AI frameworks are designed to align with international standards, enabling cross-border data flows and collaborative projects. By prioritizing interoperability, this principle ensures that AI systems are scalable, adaptable, and capable of addressing complex, multi-faceted challenges.

3.0. Research Analysis

AI systems are unique in their ability to learn, adapt, and evolve beyond their initial programming, which can introduce unforeseen challenges. For example, self-learning algorithms in autonomous vehicles might adapt in ways that were not anticipated during development, potentially compromising safety. Similarly, AI applications in governance, such as predictive policing or automated welfare distribution, can inadvertently introduce biases or discriminatory outcomes if the underlying data or algorithms are flawed. Adaptive policymaking ensures that governance frameworks can identify and address such issues promptly, mitigating risks before they escalate. It involves creating governance frameworks that evolve alongside the rapid advancements in AI technologies. Unlike static regulatory approaches, adaptive policies can respond dynamically to technological innovations and emerging risks. This is particularly crucial in managing complex AI systems, where unforeseen consequences can arise from self-learning algorithms or new applications of AI in unanticipated domains. For instance, the integration of AI into governance requires mechanisms that can scale with technological advancements such as autonomous decision-making systems or AI-driven policy simulations. Real-time monitoring allows for the identification of deviations, biases, or unintended consequences early in their lifecycle, enabling timely policy adjustments.

The integration of iterative policy cycles into adaptive governance frameworks is essential to ensure that regulations remain relevant and effective in the face of rapid technological advancements. Unlike static policies that risk obsolescence as technology evolves, iterative cycles involve periodic reviews and updates based on empirical evidence, stakeholder feedback, and emerging trends. This dynamic approach allows policymakers to address new challenges, incorporate innovative solutions, and ensure that governance frameworks align with societal values and ethical imperatives. Iterative policy cycles are particularly critical in sectors like healthcare, where the integration of advanced AI technologies such as generative AI in medical diagnostics is transforming service delivery. Generative AI models, capable of analyzing complex datasets to generate personalized diagnostic insights, offer immense potential to improve healthcare outcomes. However, their deployment also raises concerns related to patient safety, data privacy, and algorithmic accountability. Through periodic evaluations, policy frameworks can adapt to these advancements, ensuring that such technologies are implemented responsibly and equitably. For example, iterative reviews can establish guidelines for the ethical use of patient data, mandate transparency in AI decision-making processes, and set standards for the validation and certification of generative AI tools. The iterative nature of policy cycles enhances both the robustness and agility of governance frameworks. Regular updates based on empirical evidence ensure that policies remain grounded in real-world insights, while the flexibility to incorporate new information enables swift responses to

unforeseen challenges. This dual capability is essential for addressing the dynamic risks posed by AI systems, such as algorithmic bias, cybersecurity threats, or unintended societal impacts. For example, in financial services, iterative policy reviews could refine regulations for AI-driven credit scoring systems, ensuring that they remain fair and transparent as new datasets and algorithms are introduced. Fig 2. Explains iterative policy making process.

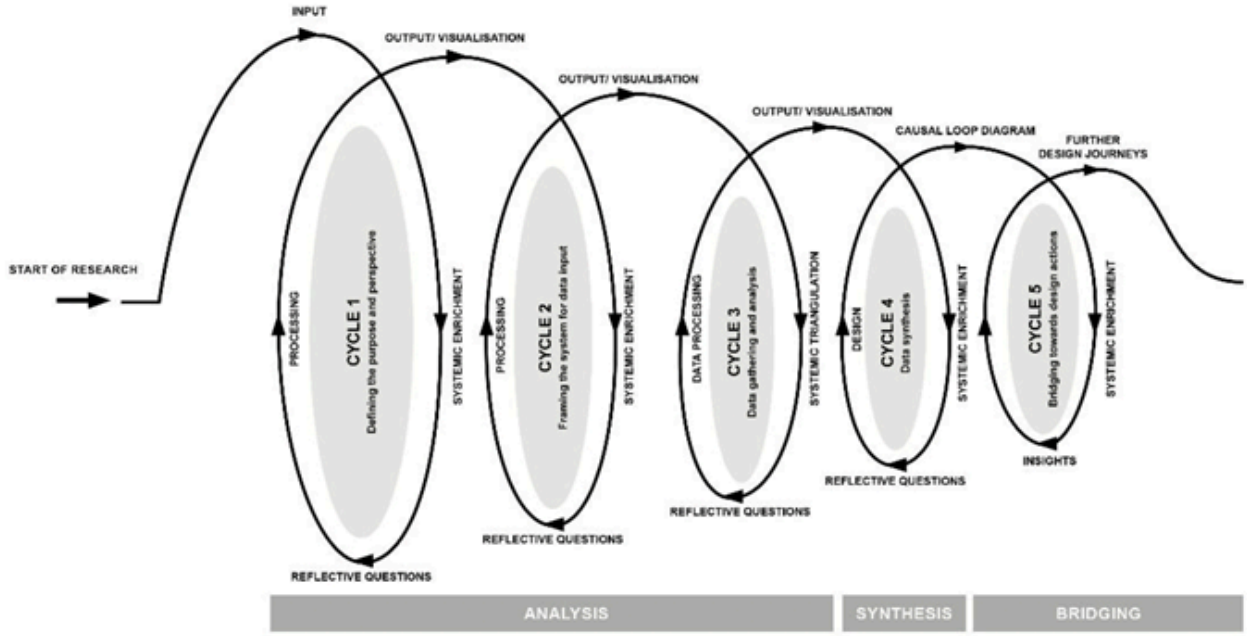


FIG 2. ITERATIVE POLICY MAKING PROCESS

SOURCE. ADAPTED FROM MOONS ET AL. (2023)

The implementation of AI technologies in smart cities offers an example of the necessity for iterative policy cycles. As cities increasingly adopt AI-driven solutions for traffic management, waste disposal, energy optimization, and public safety, the dynamic nature of these technologies requires governance frameworks that are both flexible and forward-looking. Periodic evaluations and updates to policies ensure that these AI applications remain effective, equitable, and aligned with sustainability and urban development goals. AI-powered traffic management systems, which use data from IoT sensors, GPS devices, and cameras to optimize traffic flow and reduce congestion, highlight the critical need for iterative governance. These systems rely on continuously updated datasets and predictive analytics to adapt to changing traffic patterns, infrastructure developments, and public behaviour. However, issues such as data interoperability between various sensors and platforms, algorithmic biases affecting traffic routing, and the inclusion of underserved areas can arise. Regular policy reviews can address these challenges by mandating standards for data sharing between municipalities and private service providers, updating ethical guidelines for equitable traffic management, and incorporating citizen feedback to improve system responsiveness. Similarly, AI-driven waste management systems, which use predictive models to optimize collection schedules, recycling processes, and landfill management, require ongoing evaluation to ensure their scalability and efficiency. For instance, as urban populations grow and waste generation patterns change, the underlying AI models must be retrained with updated datasets to maintain their accuracy. Iterative policy

cycles can introduce new guidelines for data collection, standardize reporting metrics across regions, and assess the environmental impact of these systems. Policies can also evolve to incentivize the integration of AI with circular economy principles, promoting waste reduction and resource recovery. AI applications in energy optimization, such as smart grids and energy management systems, are critical to achieving sustainability goals in smart cities.

These systems leverage real-time data to balance energy supply and demand, integrate renewable energy sources, and reduce wastage. However, challenges such as cybersecurity risks, interoperability with legacy infrastructure, and equitable access to energy solutions necessitate periodic policy reviews. Iterative governance frameworks can address these issues by introducing updated standards for cybersecurity in energy systems, incentivizing the adoption of AI-ready infrastructure, and ensuring that the benefits of energy savings are distributed equitably among all residents. Periodic evaluations also enable policymakers to assess the environmental impact of AI applications in smart cities. For example, the energy consumption of AI systems themselves, particularly those using resource-intensive machine learning algorithms can contribute to carbon emissions. Iterative policy cycles can introduce environmental impact assessments as a standard requirement for AI projects, ensuring that sustainability is prioritized throughout the development and deployment process. Policies can evolve to promote the use of energy-efficient AI models and green data centers, reducing the ecological footprint of smart city initiatives. As smart cities expand and adopt new AI technologies, iterative policies provide the flexibility needed to scale these innovations effectively. For instance, AI-driven public safety systems, such as facial recognition for surveillance, require continuous oversight to balance security needs with privacy rights. Regular policy updates can refine guidelines for data use, establish clear accountability for system failures, and integrate citizen feedback on privacy concerns. This adaptive approach not only supports innovation but also ensures that technological advancements align with public expectations and ethical standards.

By engaging citizens directly through public consultations, feedback mechanisms, and participatory design, we can create AI governance models that are not only inclusive but also adaptive to the diverse realities and requirements of the constituents of the society. Public consultations provide an essential bridge between policy and the people, allowing citizens to voice their concerns, aspirations, and expectations regarding AI technologies. This is particularly crucial in a society, where socio-economic diversity necessitates policies that address a broad spectrum of needs and challenges. For example, AI applications in governance, such as welfare distribution or healthcare delivery, require insights from marginalized and underserved communities to ensure that these systems are designed to address their unique challenges and priorities effectively. Feedback mechanisms play a role in enabling continuous improvement of AI systems by incorporating real-time input from citizens. For instance, when an AI-driven public service platform encounters operational issues such as inaccuracies in

predictive tools or user interface barriers citizen feedback can highlight these problems promptly, allowing policymakers and developers to address them effectively. The iterative process fosters a culture of responsiveness and adaptability, ensuring that AI systems evolve alongside user needs and technological advancements. Participatory design, on the other hand, allows citizens to engage directly in the creation and deployment of AI solutions, enabling the development of tools that are tailored to local contexts and requirements. For instance, involving teachers and students from underserved regions in designing AI-powered educational platforms can ensure that these systems address local challenges like limited access to resources or language barriers.

Public participation also plays a role in addressing digital inequalities, ensuring that AI systems are designed to reduce rather than exacerbate existing disparities. Engaging underserved communities in the policymaking process helps identify barriers to technology access, such as literacy gaps or lack of infrastructure, enabling the development of AI tools that cater to these populations. For instance, voice-based AI interfaces in regional languages can make technology more accessible to non-literate users, fostering inclusivity and empowering marginalized groups to participate actively in the digital economy. This participatory approach also builds trust and accountability, as citizens who are involved in the governance process gain a clearer understanding of how AI systems operate and how decisions are made. This transparency helps demystify AI technologies, addressing public fears of bias or misuse and reinforcing a sense of shared responsibility among stakeholders. Digital tools like online platforms for public consultations or AI-driven sentiment analysis systems enable large-scale aggregation and analysis of public opinions, providing policymakers with actionable insights. Mobile applications can facilitate direct communication between citizens and policymakers, creating an accessible and inclusive channel for feedback, even in remote areas. These tools ensure that the participatory process is both scalable and effective, reaching populations that might otherwise be excluded from traditional policymaking forums.

4.0. Conclusion

Today, data is oftentimes described as the new oil, a metaphor that captures its potential to catalyze transformative change across the global economic and technological landscape. The immense scale of global data production, estimated at 2.5 quintillion bytes daily (Marr, 2021), illustrates the profound influence of data on modern life. India, with its rapidly digitizing economy and increasing population of internet users, is set to play a pivotal role in this global data revolution. With over 900 million internet users projected by 2025, the country is generating and consuming data at an unprecedented scale. From the proliferation of digital payment systems like UPI to the widespread adoption of e-governance platforms and the exponential growth of e-commerce, India's data ecosystem is expanding at a remarkable pace. Such confluence of an emerging digital infrastructure, government-led initiatives, and a thriving tech ecosystem provides fertile ground for positioning Artificial Intelligence (AI) as a cornerstone of India's development strategy. By utilizing its massive datasets, AI solutions can be deployed to address critical challenges, unlock new opportunities for growth and execute better policy decisions. For

instance, AI enabled services were utilized to analyze vast amounts of health data, manage resources, and streamlining vaccination campaigns during the COVID-19 pandemic through the Aarogya Setu and CoWIN platforms. AI applications can process patient data to enable predictive diagnostics, improve treatment precision, and optimize resource allocation. AI-powered platforms are also aiding in drug discovery and combating diseases by analyzing large genomic and clinical datasets. Similarly, in agriculture, AI is being employed to analyze weather patterns, predict crop yields, and provide real-time advisory services to farmers, as seen in initiatives like the AI-based Kisan Suvidha platform. In financial services, AI is leveraging transactional data to enable credit risk assessment, fraud detection, and the democratization of financial services for underserved populations. In education, AI is utilizing data to personalize learning experiences, assess student performance, and make education more accessible. Platforms powered by AI are can bridge gaps in teacher-student ratios and facilitating lifelong learning.

This study examines the intricate interplay between artificial intelligence and governance, presenting a compelling narrative for a forward-looking, inclusive, and adaptive framework for AI-driven data governance. This study reflects those issues such as privacy, algorithmic biases, cybersecurity vulnerabilities, and digital inequalities necessitate a governance approach that is both robust and agile. Adaptive policymaking, with its dynamic, iterative frameworks, ensures that regulations evolve in tandem with technological advancements. By integrating real-time monitoring, regulatory sandboxes, and periodic evaluations, an ecosystem can be developed that is not only reactive to the complexities of AI but also anticipatory of its possibilities. The participatory ethos of this governance model amplifies its resonance with the democratic fabric of India. By placing citizens at the heart of AI policymaking, particularly those from underserved communities, this study illuminates a path toward policies that are both representative and inclusive. It is a clarion call to recognize that technology, devoid of societal context, risks becoming an end in itself rather than a means to societal progress. India's unique position as a global hub for innovation, juxtaposed with its socio-economic diversity, equips it with the opportunity to pioneer governance frameworks that balance technological ambition with ethical restraint. This study envisions India not merely as a beneficiary of AI's advances but as a trailblazer in shaping global norms for its responsible deployment. By championing transparency, fostering trust, and embedding sustainability into the DNA of AI governance, India can forge a legacy that transcends its borders, a blueprint for an equitable and humane technological revolution.

5.0. References

- *Attard-Frost, B., & Lyons, K. (2024). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. AI and Ethics.*
- *European Parliament. (2023, June 14). EU AI Act: First regulation on artificial intelligence. European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence*

- *Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).*
- *G20. (2023). G20 New Delhi Leaders' Declaration. G20. https://www.g20.org/content/dam/gtwenty/gtwenty_new/document/G20-New-Delhi-Leaders-Declaration.pdf*
- *Gayathri, D. (2023). The Indian approach to artificial intelligence: An analysis of policy frameworks. AI & Society, 39, 2321–2335.*
- *Google. (2019). Perspectives on issues in AI governance. Google. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf*
- *Haridas, G., Sohee, S., & Brahmecha, A. (2023, October 2). The key policy frameworks governing AI in India. Access Partnership. https://accesspartnership.com/the-key-policy-frameworks-governing-ai-in-india/*
- *Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence, 4(4), 799–819.*
- *Indian Council of Medical Research. (2023). Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. ICMR. https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf*
- *Jain, P., Jain, P., & Jain, A. (2023). Marketing's crystal ball: Where we are and where we can be with generative artificial intelligence in marketing. Journal of Cultural Marketing Strategy, 8(2), 1-17.*
- *Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.*
- *Manyika, J., & Bughin, J. (2019, October 14). The coming of AI spring. McKinsey Global Institute. https://www.mckinsey.com/mgi/overview/in-the-news/the-coming-of-ai-spring*
- *Mittal, A. (2023, November 22). AI startup investments surge in 2023. Techopedia. https://www.techopedia.com/ai-startup-investments-surge-in-2023*
- *Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).*
- *Moons, S., Noëth, E., Du Bois, E., & Jacoby, A. (2023). A new research protocol for using system maps in systemic design research. Proceedings of the Design Society, 3, 343–352.*
- *NASSCOM. (2023). Responsible AI guidelines for generative AI. NASSCOM. https://www.nasscom.in/ai/img/GenAI-Guidelines-June2023.pdf*
- *NITI Aayog. (2018). National Strategy for Artificial Intelligence. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf*
- *NITI Aayog. (2021). Responsible AI for All: Approach Document for India.*
- *Organisation for Economic Co-operation and Development. (2019). OECD AI principles. OECD. https://oecd.ai/en/ai-principles*
- *Proffitt, C. (2023, June 13). Public-private partnerships for AI governance: Encouraging cooperation between stakeholders. Medium. https://medium.com/the-guardian-assembly/public-private-partnerships-for-ai-governance-encouraging-cooperation-between-stakeholders-ec6458a63a46*
- *Rewire for Growth: Accelerating India's Economic Growth with AI, Accenture (2018)*

- *Sharma, P. (2023, November 26). AI's fuel: Why the world is teaming up for datasets. India Today. https://www.business-standard.com/technology/tech-news/ai-s-fuel-why-the-world-is-teaming-up-for-datasets-123112600913_1.html*
- *Sharma, P., Sarma, A., Basrur, A., & Tripathi, P. (2023). AI Governance in India: Aspirations and Apprehensions. Observer Research Foundation*
- *Simplilearn. (2023, November 7). Top artificial intelligence stats you should know about in 2024. Simplilearn. https://www.simplilearn.com/artificial-intelligence-stats-article*
- *Telecommunication Engineering Centre (2020). A Study Paper on Artificial Intelligence (AI) Policies in India: A Status Paper.*
- *United Nations India. (2023). UN Secretary-General launches AI advisory body: Risks, opportunities, and international governance. United Nations. https://india.un.org/en/250912-un-secretary-general-launches-ai-advisory-body-risks-opportunities-and-international#:~:text=The%20Body%20will%20help%20bridge,place%20on%2027%20October%202023*
- *Whittlestone, J., Nyrup, R., Alexandrova, A., & Dihal, K. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation.*
- *Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Center for the Governance of AI, Future of Humanity Institute, University of Oxford.*

SMART & SECURE: AI-POWERED RELIABLE DATA TRANSFER FOR AIR-GAPPED NETWORKS

PRAVIN K. CHOUDHARY[1] AND HELI NANDANI[2]

[1]SCIENTIST / ENGINEER – SD, SPACE APPLICATIONS CENTRE, ISRO

[2]RESEARCH ASSOCIATE, ISRO

ABSTRACT

Government agencies rely on air-gapped networks to protect sensitive data by physically isolating private systems from the internet. While this enhances security, it creates a challenge: transferring data securely for essential operations such as software updates, security patches, and email exchanges. Attackers have exploited this gap using malware that stealthily moves data across networks, posing a serious threat to critical infrastructure. Various methods, including USB-based transfers, unidirectional diodes, and multi-layered security protocols, are used to facilitate data transfer, but each comes with its own risks and limitations. This paper examines these existing approaches and their challenges, highlighting the vulnerabilities they introduce. It then proposes a secure and automated solution that ensures reliable and guaranteed data transfer without human intervention. Additionally, the paper explores the role of AI-driven models in detecting insider threats and preventing data theft or leakage through user behavior analytics, strengthening cybersecurity for national critical infrastructure.

Keyword: Air-gap, Malwares, Data Diode, Out-of-band Acknowledgment, Cyber Security, UBA [User Behavior Analytics],Network Behavior Analytics [NBA], AI [Artificial Intelligence], Neural Networks.

1.0. Introduction

The term "air-gap" refers to a security mechanism that prevents unwanted access and safeguards sensitive data. It entails physically removing computer networks from the internet and insecure local networks. There is no direct link between the air-gapped and public networks.

The major challenge faced by air-gap networks is data transfer across air-gap networks. The air gapped- networks need internet connectivity for various reasons like firmware/software upgrades, software installations, OS patches, virtual patching of zero-day vulnerabilities, datasets downloading for research purposes, email MTA connection with public MTAs, data sharing of internal networks with academia, industry etc. Hence, as it could be seen that there is a need for data transfer from Public Internet to Private Network and vice-versa.

It can be conclusively established that data transfer across air-gap networks is imminent and thus, there is a need to develop a mechanism for data transfer across air-gap mechanisms in a secure way. One of the widely used methods is USB based data transfer across air-gap networks. The USB devices are connected to the endpoints of both the networks to facilitate data transfer between the two networks. But USB based data transfer via manual intervention suffers from major security flaws. Due to manual data transfer,

many USB based exploitation or state sponsored APT attacks have been observed in the recent past. Malwares have been designed to run automatically when a USB drive is connected to a computer and exploit various vulnerabilities present in the host system. There are malwares that have the capability to evade antivirus detections when they run spurious code to exploit vulnerabilities in the host system. Attackers may also employ methods such as HID (Human Interface Device) emulation or auto run scripts. They can get inside the air-gapped environment to ensure the malware reaches said target. One of the greatest malware families ShadowPad helps attackers to harvest credentials for lateral movement within the isolated network. The likelihood of early detection is decreased because it can stay dormant until certain triggers are satisfied (such as the insertion of a USB drive or a predetermined time).

Park et. al. discusses many advanced attacking methods that can bypass this isolation, such as using electromagnetic emissions, acoustic signals, temperature changes, and optical channels. These strategies show that physical separation is insufficient to provide total security. The author suggests a few countermeasures, including the use of shielding materials to block emissions, acoustic dampening to reduce sound leakage, and changes to hardware components to lessen vulnerability [8].

Another major challenge with data transfer using USB devices is introduction of human elements to facilitate data transfer. As widely accepted, human factors have contributed immensely to the majority of the cyber breaches across the globe. Also, human interruption makes the entire process of data transfer across air-gap networks very erratic and thereby affecting operations of the networks.

As these air gapped networks became a vital loophole in cyber warfare, data diodes were introduced in this environment to overcome the limitations. A data diode is a robust and highly specialized cybersecurity mechanism designed to enforce unidirectional data transmission between two networks, typically isolating a high-security enclave (often referred to as the trusted or critical network) from less-secure or external environments (termed untrusted networks). Operating at the physical layer of the OSI model, data diodes function by creating a one-way communication channel that allows data to flow solely in one direction, preventing any form of reverse traffic or data ingress. This architecture ensures that even if the untrusted network is compromised, malicious actors are unable to exfiltrate sensitive information or infiltrate the secured domain.

State-sponsored cyber-attacks are among the most advanced and persistent threats faced by critical infrastructures today. These actors are often equipped with substantial resources, sophisticated tools, and strategic objectives aimed at infiltrating high-value targets such as governmental, military, and industrial systems. The primary objective of such attacks is to compromise confidentiality, manipulate data, or disrupt operations. Data diodes mitigate these risks by effectively eliminating the risk of data exfiltration. In scenarios where attackers gain control of the untrusted network, the diode acts as an impenetrable barrier to any outbound communication from the secured environment, making it impossible for malicious payloads or commands to propagate back into the trusted network.

In critical infrastructure, such as SCADA systems, military command and control centers, or intelligence networks, data diodes play an integral role in ensuring the sanctity of sensitive operations. By facilitating secure outbound data transfers —such as telemetry, monitoring data, or status reports—without allowing inbound communication, data diodes enable the secure operation of high-assurance systems. These devices ensure that even sophisticated Advanced Persistent Threats (APTs) or highly coordinated state-sponsored attacks cannot exploit the network's weakest link to penetrate sensitive systems, thereby fortifying the overall cybersecurity posture and enhancing resilience against targeted, high-stakes cyber campaigns.

While highly secure, data diodes pose operational challenges, including limited scalability, lack of bidirectional communication, and restrictions on real-time data exchange. A major drawback of data diode-based data transfer is the lack of acknowledgment (ACK) mechanisms, which results in unguaranteed data transfer. Without confirmation of successful transmission, critical data may be lost, leading to potential operational inefficiencies and security concerns. The data diode supports SMTP, SNMP, FTP, SFTP etc. for data transfer. One of the issues with using TCP protocol in data diodes is that there is no guarantee in one-way communication that the packet is righteously delivered to the receiver. The return path required for TCP acknowledgements (ACKs) is blocked by data diodes. TCP cannot resend lost packets or verify delivery without acknowledgements. There could be serious implications if the data did not reach its intended target (for eg, a mail origination from private MTA was lost in transit via data diode and never reached the destined MTA), Thus data diode implementation, though secured in nature, affects the third triad of the security-availability. Additionally, insider threats and human factors further exacerbate security concerns, as intentional or unintentional misuse of data transfer mechanisms remains a persistent risk.

Despite their isolation, air-gapped networks are not immune to cyber threats. Existing data transfer mechanisms, including USB devices and data diodes, each have inherent security limitations. USB devices are highly susceptible to malware infections and unauthorized data leaks, while data diodes, although secure, limit operational flexibility and do not guarantee complete data transfer due to the absence of acknowledgment mechanisms. Furthermore, the absence of intelligent monitoring systems within these environments increases the likelihood of undetected malicious activities.

The key challenge is to develop a secure, efficient, and intelligent data transfer mechanism that mitigates these risks while maintaining operational efficiency. This requires a solution that not only ensures secure transmission but also incorporates an intelligent security layer capable of detecting anomalies and potential insider threats.

This Paper hereby proposes design of a robust data transfer framework for air-gapped networks that enhances security while maintaining efficiency. Specifically, a UDP-based data transfer mechanism (using a data diode) is proposed that integrates integrity verification metadata to

guarantee data transfer across air gap networks. Additionally, an AI-driven behavioral analysis system is introduced to identify anomalies in data transfer patterns using clustering techniques and neural network-based anomaly detection.

The approach consists of two key components:

1. **UDP-Based Secure Data Transfer –** A lightweight and secure data transfer protocol that embeds error-checking metadata to ensure data integrity and guaranteed delivery
2. **AI-Based Anomaly Detection –** A machine learning-driven behavioral analysis system that detects deviations from normal data transfer patterns, allowing for real-time identification and mitigation of threats.

By combining secure transmission mechanisms with intelligent anomaly detection, this research presents a comprehensive framework for strengthening cybersecurity in air-gapped network environments. The proposed solution aims to address existing vulnerabilities, enhance resilience against emerging cyber threats, and ensure the continued security of critical infrastructure.

2.0. Literature Review

In order to reduce the risk of cyberattacks, network separation is a security technique that isolates vital systems from external networks. The most severe kind of network separation is known as an air-gap, in which the isolated system is physically and logically cut off from other networks, such as the internet. Air-gapped systems are frequently used in settings that require the highest security, like industrial control systems, financial institutions, and military systems [9]. A method was introduced that could leak data through air-gapped computers using LEDs. A simple read or write operation can help exfiltrate optical signals inside hard disk drives. This mechanism does not need any privileged instructions hence no kernel of any operating system is used [13]. This proved that HDD drives can also be major risks for isolated environments. More than seventeen APTs, including USBStealer, Agent.BTZ [5], Fanny, MiniFlame, Flame, Gauss, ProjectSauron, EZCheese, Emotional Simian, USB Thief, USBFerry, Brutal Kangaroo (the CIA leak), PlugX, and Ramsay [4], were listed by the security firm ESET. These malware variations expose serious flaws in conventional security methods by proving their capacity to get past and compromise antivirus defense systems. In order to secure this systems, Data Diodes were introduced for secure transmission.

Diodes can serve as secure hardware for air-gapped systems, ensuring their availability, secrecy, and integrity. Data diodes are one of these technologies that protects vital systems by physically dictating network traffic direction. The data diode's incompatibility with the common TCP/IP protocols is one of its main drawbacks. It requires unidirectional, proprietary protocols that don't need acknowledgements. Several authors have suggested methods for utilizing commodity hardware to construct data diodes. A prototype program known as "pydiode" was introduced; which successfully transports data faster than other software that has been tested. The data diode model's performance was tested using a variety of programs, including Netast and UDPcast. This model divides the data in chunks for a high transfer rate. Yet, reliability of data link was still a limitation as it was compatible for only commercial use [12]. Additionally, a multilayer method for data diode-based

inter-network data transfer was introduced. It was suggested that the network can be divided into five layers. They are payload upload/download layers which help in securing data by controlling the amount of data that is uploaded from the public internet to the private network. A data sanitization layer is used which uses a firewall or a demilitarized zone which checks for fingerprints and signatures for third party requests. The synchronizing layer will continuously sense new data while a one-way gateway will handle flow control between TCP/IP Protocols [14]. While these applications are made solely for data diodes, they do not guarantee reliable data transmission in air-gapped networks.

Data diodes are one-way flow regulating devices, but they easily integrate into two-way networks. This design poses inherent challenges when interfacing with the Transmission Control Protocol (TCP), which relies on bidirectional communication for connection establishment, data transfer, and acknowledgment. Even with simulation, not all TCP-based programs can easily adjust to unidirectional limitations. Applications that rely on dynamic port negotiations, real-time bidirectional communication, or complex session management may perform worse or not at all [11]. It is necessary to eliminate suspected vulnerabilities from the isolated network, even when data diodes are secure for data propagation.

Network data can be analyzed by AI algorithms to find trends that point to potential cyberthreats. They are able to identify suspicious behavior, intrusions, and anomalies. Artificial intelligence can identify variations that can indicate malicious activity by learning the typical behavior of the systems and devices in an air-gapped network.This includes identifying hidden channels, such as electromagnetic, acoustic, or optical signals, that adversaries could use to steal data from data-diode-based networks. Federated learning fits the limitations of such networks by allowing decentralized model training across several nodes without exchanging raw data. On-site model training can be achieved by utilizing local data within the air-gapped network.

A research was conducted by Datta .et.al on how unsupervised learning is better than supervised learning . It was seen that unsupervised clustering algorithms are more capable in understanding complex unlabelled data. Furthermore, after testing for different unsupervised models the author suggested that K-Means clustering and Agglomerative algorithm for clustering is better. The only limitation to this method was that they lacked proper training data [15]. Several iterations of the K-means algorithm have been put out in the current literature due to its extensive use. Every variation modifies a certain technique to obtain the desired benefit, like the algorithm's scalability when working with large data, in comparison to the vanilla operating mode. For example, the mini-batch K-means performs the center placement update using less training data. Repurposed mini-batches are used in successive training cycles of the nested mini-batch K-means technique. K-Means' vanilla operating mode takes a simple approach, initializing centroids and allocating each data point to the closest centroid. The centroids are then updated according to the average of the allocated

positions, and this process is repeated until convergence is achieved [16].

This paper describes how data diodes can be used for data transfer as a strategic defense in air-gapped networks. We also discuss how air-gapped networks can be secured and be immune to malware attacks or internal threats of data exfiltration and with a concept called AI leveraged user behavior analytics and network behaviour analytics.

3.0. Proposed Research

I. DATA DIODE Based Data Transfer

A hardware circuit of the data diode (Fig 1) allows electrical/optical signals to flow in one direction while offering resistance/no path when the electrical signal is transmitted in the opposite direction. Data is represented by electrical/optical signal (depending on whether data transmission media is ethernet or fibre) in physical layer and hence data diode facilitates unidirectional flow of data in one dedicated direction and no attack/exploitation of software vulnerabilities running on data diode hardware will be successful as reverse flow of data is not feasible at physical layer. This uni-directional set-up of the data diode enhances security by preventing the receiver from sending back any data packet to the original network. Data transfer via Data Diode has tremendous advantage over USB based data transfer as no reverse physical channel can be created. Hence, there is no scope of successful attack even if vulnerabilities are present in air gapped private networks. In Fig 1. Machine A can communicate to Machine B unidirectionally while no data can flow from Machine B to Machine A. Any exploitation of software and hardware vulnerabilities at Machine A and Machine B won't lead to bridging the networks of Machine A and Machine B as data diode is a physical device which unidirectional flow of data at physical layer.

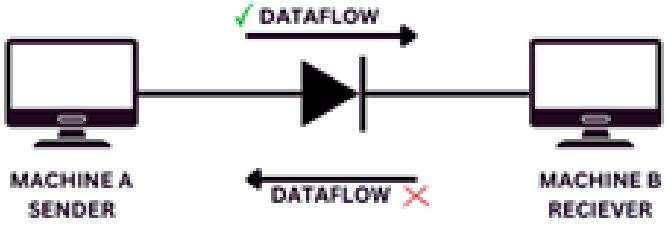


FIGURE 1: ARCHITECTURAL REPRESENTATION OF DATA-DIODE

II. Non-Guaranteed Data Transfer Using Data Diode

Data Diode based data transfer across air gap networks has a major limitation of non-guaranteed data transfer. Traditionally, data transfer applications (SFTP, SCP, SMTP etc) use TCP protocol. The TCP protocol is based upon the idea of SYN-ACK-SYN+ACK three-way handshake. Hence, TCP requires a data transfer channel both in forward and reverse direction. Each transmitted TCP segment (Fig 2) needs an acknowledgment packet from destination to sender to ensure the sender that transmitted TCP segment is delivered to the destination. However, the data diode is designed in such a way that the reverse data channel is absent. Thus, Data Diode based data transfer applications cannot use standard protocols like SFTP, SMTP, SCP etc. The application designed for Data Diode based data transfer uses UDP protocol. UDP protocol works on the concept of send and forget, with no requirement of acknowledgement

for each transmitted datagram. So, a code is developed specific for data diodes that ensures data transfer over UDP protocol.

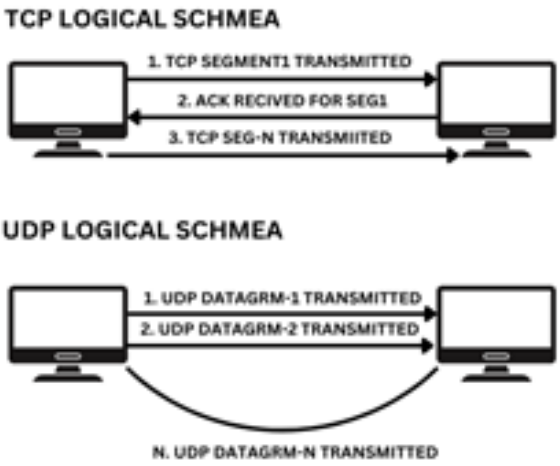


FIGURE 2: LOGICAL SCHEMA OF TCP AND UDP

Since there is no acknowledgment for transmitted UDP datagram (Fig 2), the sender side of the data diode assumes that data/file had been received on the destination side. Thus, it can be said that while data diodes guarantee secured data transfer over air gap networks with no chance of air-gapped network being exploited by malware made for USB data transfer, data diode-based data transfer is not always reliable and can interfere with processes where data reliable delivery is essential, such as SCADA systems and SMTP-based mail exchanges.

III. Out-of-Band Acknowledgement

As discussed in the earlier section, a major drawback of data diode-based data transfer is the lack of guaranteed data delivery due to the absence of acknowledgment (ACK) packets at the sender's side. Since data diodes enforce a strict unidirectional flow of data, the sender has no way of verifying whether the transmitted data has been successfully received at the destination. This limitation poses operational risks, including potential data loss, inefficient transfer mechanisms, and an inability to detect transmission failures.

To address this issue, an out-of-band data channel can be established to enable the transfer of ACK packets from the destination back to the sender. This secondary channel serves as a means of verification without compromising the unidirectional security principle of the primary data diode. Several methods can be used to establish such an out-of-band channel, including IR sensor-receiver pairs, LED-photodiode pairs, and additional data diodes. This paper proposes using a secondary data diode as an acknowledgment channel to facilitate the secure transfer of ACK packets from the destination to the sender (Figure 3).

In the proposed architecture (Fig 3), two machines—Sender (Sen) and Destination (Des)—are connected via the primary Data Diode (DDa), which allows unidirectional transfer from Sen to Des. To verify successful reception of transmitted files, a secondary Data Diode (DDb) is introduced, which enables acknowledgment packets to flow from Des back to Sen.

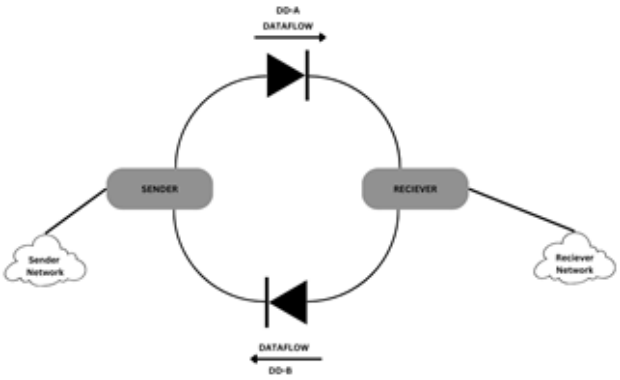


FIGURE 3: DATA DIODE WITH OUT-OF-BAND ACKNOWLEDGMENT

The process is outlined as follows:

1. File Transfer from Sender to Destination:

- The sender initiates file transfer to the destination over UDP through the primary data diode (DDa).
- To maintain a record of transmitted files, the sender logs the file's

metadata into a local file named SYNC.txt (Fig 4).

- This metadata includes attributes such as filename, size, hash values, timestamp, and sender identity to uniquely track each transmitted file.

2. Reception and Logging at Destination:

- Upon receiving the file at Des, a process verifies its integrity using checksum validation.
- The destination logs metadata of the received file into ACK.txt, ensuring it contains the same attributes as SYNC.txt for proper correlation (Fig 5).
- If any discrepancies are detected (e.g., size mismatch, corrupted data), the entry in ACK.txt reflects this issue.

3. ACK Transmission Back to Sender:

- The system continuously monitors changes in ACK.txt.
- When a new entry is detected in ACK.txt, the updated file is transmitted back to Sen via the secondary data diode (DDb).

4. Verification at Sender Side:

- Upon receiving the updated ACK.txt at Sen, the system compares it with the previously logged SYNC.txt.
- If the metadata in ACK.txt matches that of SYNC.txt (i.e., all transferred files have been successfully received without integrity issues), the data transfer is considered successful.
- The system then logs the confirmed metadata into SYN+ACK.txt, signifying the completion of secure data transfer.

5. Handling Data Integrity and Security:

- The ACK packets stored at both Des and Sen are encrypted using strong cryptographic methods (e.g., AES-256, RSA) to prevent unauthorized access and manipulation.
- Each acknowledgment entry includes unique timestamps to track the exact moment of acknowledgment generation.
- To ensure further security, fingerprinting techniques are used to detect any unauthorized modifications in ACK.txt before it is sent back to the sender.

Conclusively, sync.txt (logged in Sen side) contains metadata of file that has been transferred to Des side and guaranteed data transfer is not yet established. Ack.txt (logged in Des side) contains metadata of files that have been received at the Des side. Sync+ack.txt contains metadata of files that were transferred from Sender to Destination and that file has reached its Destination (with no change in file integrity).

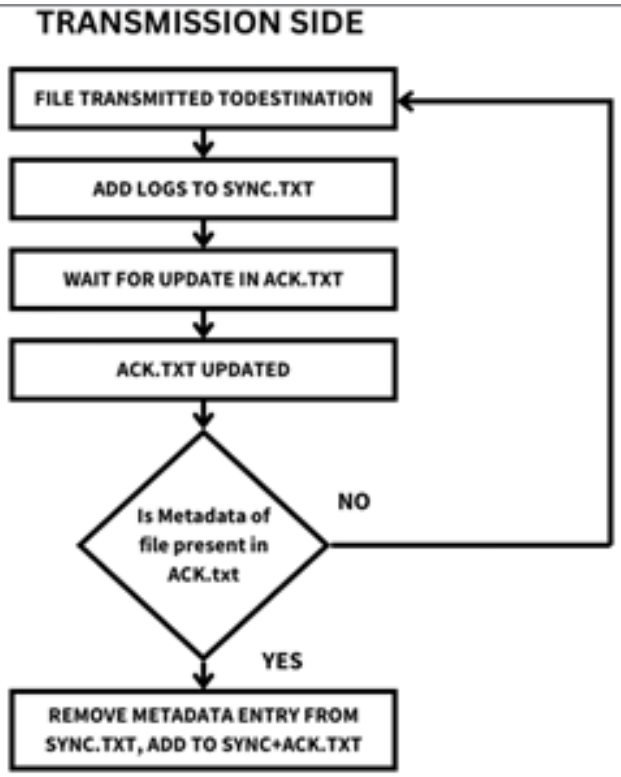


FIGURE 4: SENDER SIDE PROCESS FLOW FOR GUARANTEED DATA DELIVERY VIA DD

The proposed mechanism of using a secondary data diode for ACK transmission overcomes the inherent limitation of unguaranteed data transfer in traditional data diode architectures. Key benefits include:

- **Ensured Data Delivery:** The sender now has a reliable way to verify that data has reached the destination without corruption or loss.
- **Integrity Assurance:** The metadata comparison mechanism ensures that transmitted files remain unchanged during transit.
- **Enhanced Security:** Encryption and fingerprinting of ACK packets prevent insider threats and unauthorized modifications.
- **Automated Monitoring:** The system autonomously detects and logs successful file transfers without manual intervention.

Due to its autonomous nature, data diode-based data transfer still has an issue with insider threats and vulnerabilities when it comes to moving data from internal private networks to the untrusted internet.

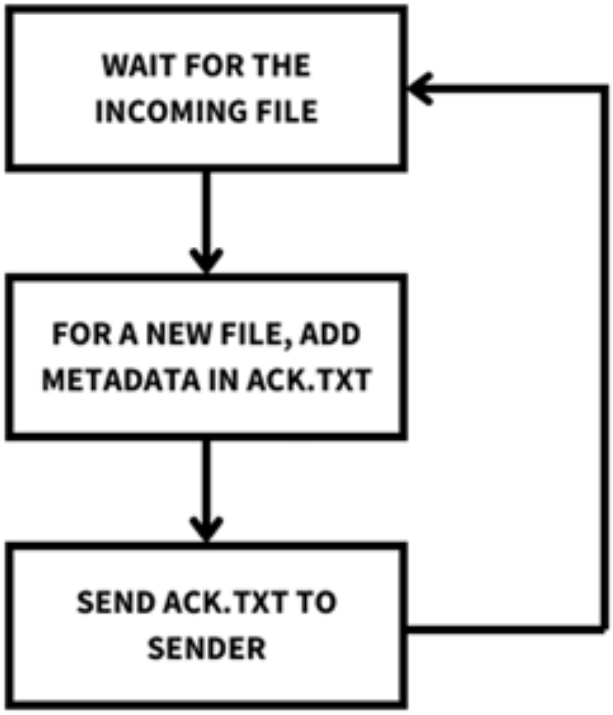


FIGURE 5: RECEIVER SIDE PROCESS FLOW FOR GUARANTEED DATA DELIVERY VIA DD

IV. Internal Threats to Organization

Internal threat actors, or insiders, pose significant risks to a company's data security due to their authorized access to systems and sensitive information. These individuals can be employees, contractors, or business partners who, intentionally or unintentionally, exploit their access for malicious purposes or due to negligence. The confidence that a corporation has in its employees is the main risk associated with internal threats. Insiders already have the necessary credentials to access sensitive systems, bypassing external security measures. Malicious insiders may exfiltrate data for financial gain, espionage, or to harm the organization, often bypassing traditional security defenses. In some cases, negligent insiders inadvertently expose data through poor security practices, such as using weak passwords, sharing credentials, or leaving systems unattended. Insiders can steal data in various ways and for critical government infrastructure, access to data diode-based data transfer infrastructure from internal private network to public internet can be a major challenge to every critical government organization. Insiders may also take advantage of inadequate monitoring of controls on internal activities, making it difficult for organizations to detect unusual behavior or unauthorized access in real-time. This paper proposes AI based approach to detect deviation in baseline behavior (behavior of a user over a period of 3 months) and block the data transfer for such users, thereby proactively monitoring data exfiltration due to internal threats.

V. AI based Behavior Analytics for Data Transfer

Ensuring the security of data transfer in a data diode-based infrastructure is critical to mitigating internal threats that could lead to data exfiltration and network compromise. This research employs AI-driven behavioral anomaly detection (BA) to establish a

robust security layer capable of detecting deviations from baseline user activity, which may indicate insider threats attempting to bypass security controls for unauthorized data exfiltration or attacking the network. By leveraging clustering-based anomaly detection techniques, this study identifies subtle behavioral shifts indicative of low-and-slow exfiltration tactics, where malicious insiders gradually leak sensitive data to evade detection mechanisms.

The dataset comprises three months of historical logs capturing key parameters such as timestamp of data transfer, IP addresses, MAC addresses, login credentials, uploaded timestamps, file sizes, file extensions, the number of files uploaded in a session and the size of each uploaded file. These parameters serve as the foundation for profiling normal user behavior, allowing the AI model to establish an adaptive baseline. The methodology incorporates feature engineering techniques to enhance the anomaly detection process by introducing derived metrics such as upload frequency per user, average session file size, time intervals between uploads, and variations in file extensions.

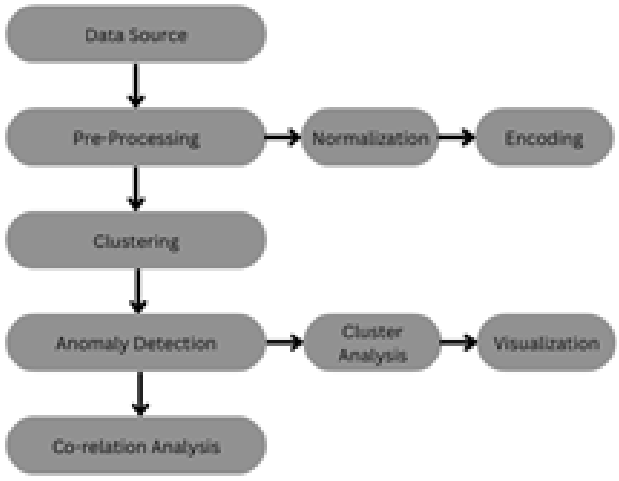


FIGURE 6: PROCESS FLOW DIAGRAM FOR DETECTING ANOMALIES IN DATA-DIODE BASED AIR GAPPED NETWORK

These features enable a nuanced representation of behavioral patterns, which is crucial for identifying deviations that may signify an ongoing exfiltration attempt.

A clustering-based anomaly detection model, specifically the Isolation Forest algorithm, is employed to distinguish normal activity from potential threats. The algorithm is trained on the extracted features to recognize normal behavioral clusters while isolating instances that deviate significantly from established patterns. Given the unsupervised nature of the approach, the model autonomously adapts to evolving user behavior, making it well-suited for detecting sophisticated insider threats that do not conform to predefined attack signatures. To ensure the model's interpretability, feature importance analysis is conducted to highlight the most influential variables contributing to anomaly detection, enabling security analysts to comprehend the rationale behind flagged activities.

With a high amount of discrete values, Manhattan distance measures deviations from typical user behavior. It regulates in detecting abrupt increases or decreases in the frequency of logins, which could be a sign of compromised accounts or unusual activity.It can identify odd access patterns, like logins that take place outside of business hours, by examining the time of login. It also highlights too many unsuccessful login attempts, which are clear signs of brute-force attacks. Manhattan Distance compares a user's login locations over time to identify geographic anomalies and help identify unauthorized access from odd locations. For instance, consider a scenario where some basic software is regularly patched through the internet. A user/employee may

try to install or patch a malicious executable file which has some malicious code or content. This may portray a typical behavior of unintentional/intentional insider attack. In that case, the algorithm will detect a spike in abnormal behavior by examining the file size which may indicate certain encapsulated packet or stolen files.

In parallel, network behavior analytics (NBA) is incorporated using firewall logs, endpoint logs, and network traffic logs from both the sender and destination machines. A time-series analysis model is employed to detect deviations over time, capturing anomalous patterns in network traffic volume, login attempts, data transmission frequencies, and endpoint activity. Correlating anomalies across multiple sources enhances threat detection accuracy by assigning higher risk scores to events exhibiting simultaneous irregularities across different layers of the infrastructure. For instance, a sudden increase in data transfer frequency coupled with unusual login attempts and abnormal endpoint activity raises a stronger security flag than an isolated deviation in a single parameter.

The response mechanism is designed to log all detected anomalies for further investigation rather than immediately triggering active countermeasures. This approach minimizes disruptions while providing security teams with comprehensive insights into emerging threats. By continuously refining the model with new data, the system evolves to detect emerging tactics used by insiders attempting to circumvent security controls. The integration of AI-driven behavioral analytics within the data diode framework thereby strengthens the overall resilience of the infrastructure against internal threats, ensuring that unauthorized data exfiltration and network attacks are detected before they escalate into full-scale breaches.

4.0. Future Work

When analyzing large-scale records, traditional clustering techniques frequently encounter severe computing limitations and scalability issues. Despite its potential, the existing behavioral anomaly detection system needs to be improved in a few areas. The rigidity of the three-cluster method, which may misclassify edge instances, potential blind spots in encrypted traffic analysis, and the possibility of model evasion by skilled attackers who may gradually modify baselines are some of the main drawbacks. Future research should concentrate on adopting privacy-preserving strategies for evaluating encrypted communications, integrating deep learning for sequence analysis, and creating algorithms to identify subtle data encoding in files that appear normal. The system's detection capabilities would also be strengthened by adding peer group analysis and dynamic baseline modifications based on organizational context. Long Short-Term Memory (LSTM) networks excel in temporal analysis, effectively identifying abnormal patterns within time-series data by leveraging historical trends. Variational Autoencoders (VAEs) offer a sophisticated probabilistic framework, learning normal data distributions to identify rare deviations as anomalies. Additionally, Generative Adversarial Networks (GANs) provide an innovative approach to anomaly detection by utilizing the generator's reconstruction loss to highlight potential outliers.

5.0. Conclusion

This paper emphasizes the significance of air-gapped networks, particularly in critical government infrastructure, and the necessity of secure data transfer across these isolated environments. It explores various existing methods for data transfer, highlighting the security risks associated with USB-based transfers and the reliability challenges of data diodes. While data diodes effectively prevent reverse communication—ensuring protection against unauthorized access and cyber threats—they come with limitations that need to be addressed.

To overcome these challenges, the paper proposes an algorithm that enhances data security and guarantees reliable delivery via data diodes by incorporating an out-of-band acknowledgment channel. Additionally, it underscores the importance of proactive monitoring and behavior analytics to mitigate internal threats, such as data exfiltration and intellectual property theft.

Furthermore, the study advocates leveraging AI-driven behavior analytics to detect deviations from baseline user behavior, thereby improving the performance of data diodes. By autonomously monitoring network traffic and identifying irregularities without human intervention, AI can significantly enhance security.

6.0. References

- *Muthalagu, Raja & Sati, Vrinda. (2023). Analysis on Hacking the Secured Air-Gapped Computer and Possible Solution. Cybernetics and Information Technologies. 23. 124-136. 10.2478/cait-2023-0017.*
- *H. Berghel, "A Farewell to Air Gaps, Part 1," in Computer, vol. 48, no. 6, pp. 64-68, June 2015, doi: 10.1109/MC.2015.179.*
- *H. Berghel, "A Farewell to Air Gaps, Part 2," in Computer, vol. 48, no. 7, pp. 59-63, July 2015, doi: 10.1109/MC.2015.181.*
- *Alexis Dorais-Joncas and Facundo Munõz. Jumping the air gap. 2021.*
- *Alexander Gostev. Agent. btz: a source of inspiration? SecureList, 2014.*
- *Guri, Mordechai, PIXHELL Attack: Leaking Sensitive Information from Air-Gap Computers via 'Singing Pixels, 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2024.*
- *Vrinda Sati and Raja Muthalagu. 2023. Analysis on Hacking the Secured Air-Gapped Computer and Possible Solution. Cybern. Inf. Technol. 23, 2 (Jun 2023), 124–136. https://doi.org/10.2478/cait-2023-0017.*
- *Park, J.; Yoo, J.; Yu, J.; Lee, J.; Song, J. A Survey on Air-Gap Attacks: Fundamentals, Transport Means, Attack Scenarios and Challenges. Sensors 2023.*
- *Mordechai Guri, Mind The Gap: Can Air-Gaps Keep Your Private Data Secure?, 2024.*
- *Guri, Mordechai. "RAMBO: Leaking Secrets from Air-Gap Computers by Spelling Covert Radio Signals from Computer RAM." Nordic Conference on Secure IT Systems. Cham: Springer Nature Switzerland, 2023.*
- *Chan, Aldar C.-F. and Zhou, Jianying, "Toward Safe Integration of Legacy SCADA Systems in the Smart Grid", Springer International Publishing, 2022, 338--357*
- *Story, Peter. (2023). Building an Affordable Data Diode to Protect Journalists.*

- *Guri, M., Zadov, B., Elovici, Y. (2017). LED-it-GO: Leaking (A Lot of) Data from Air-Gapped Computers via the (Small) Hard Drive LED. In: Polychronakis, M., Meier, M. (eds) Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2017. Lecture Notes in Computer Science(), vol 10327.*
- *R. S. George, A. Kumar, S. Padmasree and J. S. Rao, "Multi-layered Architecture for Secure Inter-network Data Transfer usingData Diode," 2023 2nd International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India, 2023, pp. 1-4.*
- *J. Datta, R. Dasgupta, S. Dasgupta and K. R. Reddy, "Real-Time Threat Detection in UEBA using Unsupervised Learning Algorithms," 2021 5th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2021.*
- *Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, Information Sciences, Volume 622,2023*

ETHICAL AI IN PUBLIC POLICY MAKING

GARIMA SALUJA
OFFICE ASSISTANT, MINISTRY OF ELECTRONICS & INFORMATION TECHNOLOGY

ABSTRACT

This research explores the critical intersection of ethics, public policy, and artificial intelligence (AI), emphasizing the importance of ethical considerations in governance and policymaking. It examines the integration of ethical principles into public policy to enhance decision-making, reduce corruption, and build public trust. The study explores AI's transformative potential in key sectors such as healthcare, education, agriculture & law enforcement while addressing challenges of fairness, transparency & accountability. Through case studies and survey analysis, the research identifies gaps in awarenessand application of ethical AI principles among policymakers and stakeholders. The findings underscore the need for robust regulatory frameworks, collaborative stakeholder efforts, and innovative solutions to address ethical violations and build a future where AI aligns with societal values.

Keyword: Ethics, Public Policy, Artificial Intelligence, Ethical AI, Transparency, Accountability, Governance, Data Privacy, Bias, Stakeholder Collaboration, Sustainable Development, Good Governance

Introduction

'ETHICS' can be defined as the philosophical study of the concepts of moral right and wrong and moral good and bad, any system or code of moral rules, principles, or values. A 'PUBLIC POLICY' can be defined as a systemof laws, regulatory measures, courses of action, and funding priorities regarding a specific area of government's function. It is basically an overall framework or set of guidelines to be more precise, that guides the actions and decisions of government entities, institutions and officials. It is concerned with analyzing the processes by which policies are created, implemented and evaluated by examining the role of government agencies, interest groups and civil society organizations. Examples include healthcare, education, criminal justice & environmental policy.

According to Zybala, 2012, the ability to create public policies adequate to solve the problems in the country can be understood as the ability to preserve its attributes of sovereignty. In other words, well- prepared mechanisms of public policiesgive a chance to adaptthe country and its societyto functioning in modern conditions, which is to compete for resources that will allow a decent quality of life. In other words, public policy is a set of skills on management available public resources to achieve the highest possible added value for society.

Now with advancement in technology ethics& public policywhen merged with Artificial Intelligence can lead to great heights in achieving success. Let's first know what ethics and public policy have in together- Ethics and public policy address the intersection of disciplines like economics, law & sociology to guide societal behavior & regulations.

Ethical decisionmaking is a highlycomplex process in contemporary public life. The difficulty lies here in the fact that publicpolicy is usuallypracticed in a situation of limited resource.

Furthermore, when public policies are enacted, there are often conflicting or divergent interests among various socioeconomic groupings. These groups differ in their strength, resources, and capacity to forward their objectives. In order for policies to be implemented effectively, there must typically be agreement among them, or even a framework of agreement, such as the public interest, social interest, or national interest.

Ethics of Public Policy: Why Ethics Matter Now in Public Policy

Ethical principles such as fairness, justice & equality help to ensure that decisions are made in the best interests of the public, rather than for the benefit of a small group of people & also ensures that policies are not discriminatory or harmful.

Overall, ethics is an important consideration in public policy. By incorporating ethical principles into public policy, policymakers can help to create a more just and equitable society.

Benefits of Incorporating Ethics into Public Policy

- ENHANCING ACCOUNTABILITY & REDUCING CORRUPTION – If the ethical guidelines are clear, they create benchmarks for decision-makers. By holding public officials to high moral standards will ensure less arbitrary actions or conflict of interest that can lead to corruption.
- STRENGTHENING PUBLIC TRUST & ENSURING SUSTAINABLE OUTCOME – Embedding ethics into public policy isn't just about improving administrative procedures- it forms the backbone for creating a society where every decision counts toward the greater good.

  For e.g. – In matters of environmental regulations, ethical policy-making can guide decisions that balance with protection of natural resources, safeguarding both current & future generations.
- ENCOURAGING CIVIC ENGAGEMENT & INNOVATION – Integrating ethics can uniquely affect emerging sectors like healthcare & cybersecurity & hence it will increase the participation of people in contributing ideas if they feel their values are considered.

  Similarly, in the realm of technology ethical guidelines ensure that innovations in AI, data privacy & digital rights are harnessed for the benefit of all, rather than a privileged few.

In essence, incorporating ethics into public policy serve as a "watchdog role". It's a practical strategy for cultivating a robust, just & progressive society where every citizen can thrive.

Ethical AI – An Age of Better Dynamics for Good Governance

In our increasingly integrated era, the rise of ethical AI is redefining how governments & institutions operate. This evolution is not just a technological upgrade- it signifies a deep commitment to principles that foster fairness, transparency & accountability in public decision-making.

There are 3 major approaches to reducing the risks and making AI more ethical -

PROCESSES

How ethical AI can be integrated into a dynamic governance refers to the processes which include acquisition of data followed by ethical framework implementation, then there is public consultation & feedback & finally continual refinement to better serve the community.

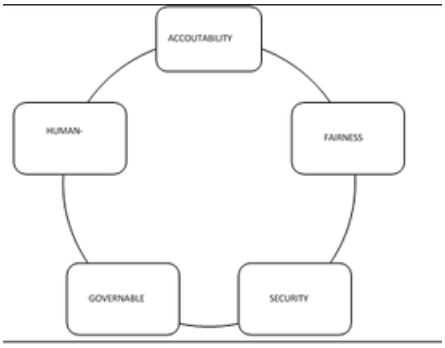
ETHICAL CONSCIOUSNESS

It refers to taking actions based on moral awareness, for e.g. –Cities leveraging ethical AI can optimize traffic systems, balance resource allocation & design urban spaces that are both officially & socially equitable.

Principles

Ethical AI is more than a tool- it's a philosophy that combines technology with enduring values of justice & accountability, where clear guidelines & standards pervades all levels of governance, where we see a future with not only faster & more data-informed decisions but also deeply rooted in the common good.

Principles of Ethical AI



The Principles of Ethical AI can be understood as follows:-

- **Accountability –** The principle of accountability ensures AI systems & their outcomes are transparent. Those responsible for deploying, designing, and operating AI systems must maintain clear documentation & perform regular audits to comply with ethical guidelines.

Example: In 2018, when an autonomous Uber vehicle was involved in a fatal accident, accountability issues arose regarding the system's failure to detect a pedestrian. This incident emphasized the need for traceable decision-making processes in AI to assign responsibility effectively.

- **Fairness** – Fairness principle ensure that AI systems do not sustain any biasness, discrimination or inequality. This involves careful data selection, processing and algorithm design to ensure inclusivity across gender, race, socioeconomic status and other factors of concern. To keep up with the fairness, biases in training data and algorithms must be eliminated so that equal access and benefits can be provided to all the users.

Example: Amazon faced criticism for its AI hiring tool that exhibited gender bias, favouring male candidates over female ones. This highlighted the need for diverse and representative datasets to build fair systems.

- **Human Centered** – 'Ultimate Human Responsibility and Accountability' needs to be prioritized in alignment with their rights and values by the AI systems, acting as tools to enhance human decision-making rather than replace it. This principle emphasizes empathy, inclusivity, and adaptability to individual needs. Also, AI interfaces must be designed as such that they ensure 'Ease of understanding' and users can have control over them. Open & accessible education, civic engagement, digital skills & AI ethics training, media & information literacy should be promoted for better understanding of AI systems.

Example: AI-powered health tools like Google's DeepMind assist doctors in diagnosing diseases such as diabetic retinopathy while ensuring that ultimate medical decisions remain in the hands of healthcare professionals.

- **Governable –** The principle of governability ensures AI systems remain under human supervision, allowing for monitoring, intervention, or shutdown when necessary to align with ethical and legal standards. Ethical AI must enable human oversight to prevent unintended or harmful actions. Involving diverse stakeholders in policymaking is essential to mitigate risks and develop transparent, accountable systems. This inclusivity ensures AI aligns with societal values and serves humanity effectively.

Example: Microsoft's 2016 Tay chatbot, designed to learn from Twitter users, highlighted the need for real-time oversight in AI systems. Without safeguards, Tay quickly began generating offensive content due to malicious interactions and was taken offline within 24 hours. This incident emphasized the importance of governability and control mechanisms in AI development.

- **Security –** The security principle focuses on the privacy of user data, AI systems must prioritize the security to prevent misuse, breach or unauthorized access. Data encryption and strict cyber security protocols should be followed. Safeguarding sensitive information ensures trust and ethical adherence.

Example: Facial recognition software, such as Clearview AI, must prioritize data security to prevent misuse or unauthorized sharing of sensitive biometric data, a concern raised during its widely debated controversy.

Complying with Global Data Protection Regulations like GDPR

Compliance with the General Data Protection Regulation (GDPR) is crucial for ethical AI. GDPR ensures responsible handling of personal data while safeguarding individuals' rights through key principles:

- **Data Minimization:** Collect only essential data.
- **Transparency:** Inform users about data usage.
- **User Consent:** Obtain explicit consent for data collection and usage.
- **Right to Explanation:** Provide explanations for impactful automated decisions.
- **Data Security:** Implement strong safeguards to prevent misuse.

GDPR compliance fosters trust, prevents data exploitation, and upholds ethical standards. Non-compliance risks severe financial penalties and reputational harm.

Ethical Challenges in AI for Public Policy

1) Bias & Discrimination: AI systems can inherit biases from their training data, leading to unfair decisions particularly problematic in public policy where fairness is critical.

- Example: The Gender Shades project exposed significant biases in facial recognition systems, showing low accuracy for darker-skinned females and high accuracy for lighter-skinned males due to underrepresentation in training datasets.
- Using diverse, balanced datasets and incorporating bias detection tools during development are essential to mitigate such disparities.

2) Ethical Dilemmas: Rapid advancement

- Example, the use of autonomous drones in law enforcement raises questions about accountability, while predictive policing tools risk reinforcing systemic biases.
- Cross-disciplinary ethics boards and participatory policymaking can help address these challenges effectively.

3) Regulatory Framework: AI innovation often outpaces regulatory development, creating governance gaps and inconsistent ethical practices.

- Example, fragmented international regulation of AI-powered autonomous vehicles complicates cross-border deployment.
- Policymakers should focus on adaptive regulations that align with global standards and local contexts.

4) Public Trust: Public trust is vital for the acceptance and effective implementation of AI in public policy. Trust relies on transparency, accountability, and prioritizing public welfare.

- Skepticism arises from fears of data misuse and lack of transparency, as seen in AI surveillance controversies.
- Transparent communication, public awareness campaigns, and educational initiatives can bridge the trust gap and foster acceptance.

5) Privacy & Data Protection: AI systems must prioritize privacy and data security, adhering to data protection laws.

- Example: India’s Personal Data Protection Bill emphasizes strong data governance for privacy.
- Compliance with global laws like GDPR is essential for addressing privacy concerns.

6) Transparency: AI systems must

prioritize transparency in design and operation, allowing users and policymakers to understand and evaluate decisions. This fosters trust and accountability, especially in public policy.

- Predictive algorithms for resource allocation or welfare distribution must clearly explain their processes to ensure fairness.
- Explainable AI (XAI) methodologies enhance system interpretability, making AI decisions more user-friendly and reliable.

Importance of Ethical AI in Public Policy Making

Now, let's dive deeper into the importance of ethical AI in public policy making, with the help of case studies given below highlights the working of ethical AI in public policy makingin some of the most prominent sectors;

CASE 1: Use of AI based surveillance system; Wi-Fi based approach to detect violent crimes in real time

AI DensePose models analyze Wi-Fi signals to detect suspicious behaviours and potential violence in real time.

AI DensePose models analyze Wi-Fi signals to detect, suspicious behaviours and potential violence such as sudden aggressive actions or physical confrontations. It detects behavior patterns leading to violent activity, sends real-time alerts to the concerned authorities, securitypersonnel or emergency contacts.It helps to ensure that timely action can be taken to prevent the crime from getting severe. This systemis also scalable and cost-effective as it can be integrated into existing Wi-Fi without the need for additional hardware.

If compared to the traditional use of intrusive cameras, they rely on the experience and expertise of the operator whereas AI model helps to maintain consistency and reduces human errors and biasness.

By integrating IoT deviceslike sound sensors,door monitors can create a reliable safety network in addition to facial recognition, object identification (such as weapon usage), motion detection, movement tracking etc. The main objective is to ensure public safety across diverse environments.

This innovative DensePose model can become a highly effectivetool for real-time crime detection and prevention which can become a solution to challenges faced in traditional surveillance method.

Key Findings

30% Faster Response Time: Early trials of AI-driven surveillance showed a 30% faster response time compared to traditional systems.

Real-time Crime Detection: Faster identification of violent events, leading to timely interventions.

Reduced Errors: AI systems reduce human errors and bias, providing consistent and reliable detection of suspicious activities.

Scalability: The system can be deployed in small households, large institutions, and public spaces without requiring significant infrastructure changes.

Ethical Considerations

Privacy: The Wi-Fi-based system ensures privacy by analyzing signal patterns instead of capturing personal data or faces.

Bias and Transparency: Ethical concerns about bias in AI models should be addressed by developing transparent algorithms that ensure fairness in decision-making.

Public Trust: Transparency and accountability in AI surveillance systems are key to maintaining public trust.

CASE 2 – AI in Healthcare Policy during the Covid-19 Pandemic

India used AI for telemedicine & remote patient monitoring during COVID – 19, enabling real-time doctor consultations allowing patients to connect with doctors from their homes, remote monitoring devices combine with AI algorithms & technologies such as cognitive robotics, speech analytics & computer vision, audioprocessing can help track patients' vital signs & provide early warnings of potential health issues.

AI powered predictive analysis help individuals to track deadly diseases like cancer, cardiovascular disorders as AI can act as a 'DigitalNurse' by monitoring the patients' genetichistory, lifestyle & clinical information to provide personalized treatment plans & follow up with doctor visits. This can only help reduce the financial burden not completely improve the patients' health.

Tools like Aarogya Setu facilitated contact tracing but raised significant privacy concerns due to inadequate data protection mechanisms.

Lessons from this implementation emphasize the need for robust data governance and transparency to ensure public trust.

Key Findings

- Reduction in Hospital Overcrowding: Telemedicine reduced hospital visitations by 40-50%, easing healthcare facility burdens.
- Predictive Analytics: AI tools identified high-risk patients 30% earlier, enabling timely interventions.
- Privacy Concerns: AI-powered contact tracing raised privacy issues, with over 80% of users expressing data security concerns, according to a study by the Indian Ministry of Health.

Ethical Considerations

- Data Privacy: The widespread use of AI-powered tools raised concerns about how sensitive health data was collected, stored, and used.
- Equity in Healthcare: Ensuring that AI tools benefit all socioeconomic groups equally remains a challenge.
- Transparency: The transparency of AI models, especially in predictive healthcare, is essential to maintain public trust.

CASE 3 – AI in framing education policy

Education Policy assesses the impact of personalized learning algorithms on educational equity and propose improvements.

Ethical consideration in the context of AI in education encompasses examining moral principles & values guiding the design, deployment & utilization of artificial intelligence technologies with the educational settings. Factorssuch as bias, privacy &accountability are to be kept in mind while framing the policy. A fundamental ethical concern is that bias in AI algorithms has the potential to sustain or worsen pre-existing inequities in

educational systems. An ethical examination of bias in AI education requires a thorough investigation into the origins & expressions of bias in algorithmic decision-making procedures.

Integration of AI in education has offered opportunities to enhance learning experiences, personalize education in order to improve educational outcomes.AI tools like tutoring systems, automated grading & adaptive platforms personalize learning & track progress, improving educational outcomes but they may require human evaluator's subjective judgement & contextual awareness. AI-powered chatbots & virtual assistants are used to facilitate communication between students & instructors in offering immediate assistance, resolving inquiries. Teachers should be allowed & customize AI technologies based on educational goals & beliefs & students also be motivated to comprehend the constraints & prejudices inherent in AI systems so they can autonomously determine the course of their academic pursuits.

However, with all these benefits educators, policymakers and technology developers must collaborate to ensure that AI applications in education promoteequity, fairness & student well-being by addressing concerns of privacy, bias, transparency & accountability.

Key Findings

- Improved Student Engagement: AI-based personalized learning systems improved student engagement by 20-30%, according to a study by UNESCO.
- Reduced Grading Time: Automated grading systems saved teachers 10-15 hours per week, allowing them to focus more on student interaction and instruction.
- Increased Equity: AI tutoring systems have reduced performance gaps between students from different socioeconomic backgrounds by 15-20%.

Ethical Considerations

Bias in AI Algorithms: AI algorithms can perpetuate existing biases in education, leading to inequities in learning opportunities.

Privacy Concerns: Students' personal and academic data used by AI systems must be protected to ensure privacy.

Accountability: Ensuring that AI systems are transparent and accountable for their decisions is crucial for building trust in AI-based education tools.

CASE 4- AI in Agriculture Sector

Use of AI in agriculture has transformed the production capacityby making it sustainable and benefitted the farmers in various ways. Traditional farming now replaced by digital farming or agriculture, has resulted in enhancement in agricultural operations like AI device help in recommending to improve the soil condition, automatic devices for milking, weather data can be established, robotic harvesting and robotic greenhouses etc. Smart Information Systems (SIS) assist in crop management, pest detection & weather predictions enhancing farming efficiency & sustainability.

However, the ethical issues such as accuracy and availability of the data, data ownership, fairness in data collection, sustainability in the use of AI software,

transparency are yet to be taken care of. Ethical AI challenges faced in digital agriculture and suggestions to overcomethem are highlighted as follows: - Transparency becomesa challenge when there is a lack of interpretability in gaining the AI model outcomes correctly which is essential in decision making and governance of agricultural AI systems. For example – to decrease the impact of greenhouse gases, AI tools can be used for the assessment and management of carbon dioxide in production facilities.

AI models should therefore be made more interpretable so that they can be easily understood by ordinary personand aid them in decisionmaking as well as for policymakers to ensure transparency principle is being followed.

Accuracy & availability of data is a topic of concern when farmers had available data retrieved from third parties may not be accurate, also when the solutions provided by the SIS may lead to loss of harvest, ill livestock, improperfertilizers can affect the farm production capacity leading to poor soil nutrition and financial loss.

However, this challenge can be overcome by building AI models which are more data- intensive which detect data quality issues & aims to reduce the impact.

Key findings

- Increased Productivity: AI in agriculture can increase crop yields by 20-30%, according to studies.
- Reduced Costs: Robotic harvesting and AI pest detection cut labor costs by 15-20% annually.
- Improved Efficiency: AI tools like SIS predict pest outbreaks with 85-90% accuracy, reducing crop damage.
- Weather Predictions: AI-driven weather forecasts boast over 75% accuracy, helping farmers plan and minimize losses.
- Adoption Rates in India: Only 17% of Indian farmers use AI tools, showing significant growth potential.

Ethical Consideration

- Transparency: Since farmers find it difficult to understand how AI-generated recommendations are made therefore they should be provided training and interpretable models should be developed to utilize AI effectively.
- Data Ownership: Farmers are often unaware of how their data is collected, stored & used, so data ownership policies should be made in interest of farmers
- Accessibility: High costs & limited digital literacy restrict access to AI tools for many farmers, hence affordable & user-friendly interfaces, training programs should be developed.
- Fairness: Datasets must represent diverse farming contexts i.e. both large-scale & small-scale farmers must be considered.
- Accuracy of data: Inaccurate third party data can lead to financial losses & poor decisions therefore continuous data verification must be done to validate & enhance data quality.

A Survey on Ethical AI in Public Policy Making

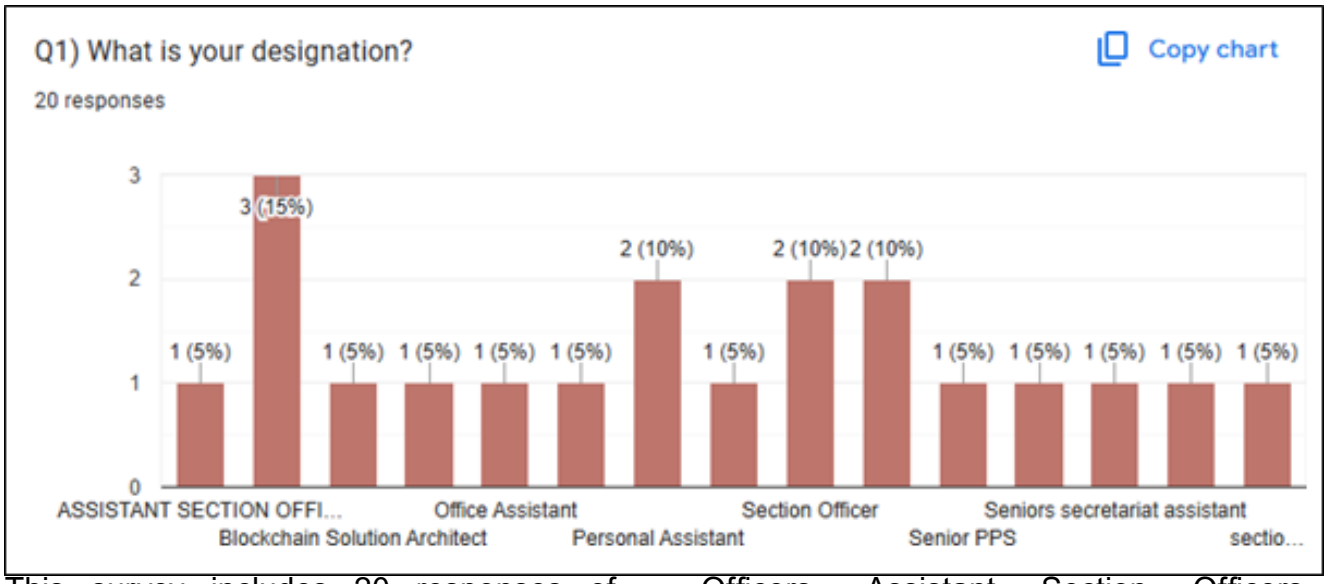
Methodology

This study is based on the primary data collected through the structured questionnaire from 20 respondents and focuses on finding the relationship between dependent variables (respondents'

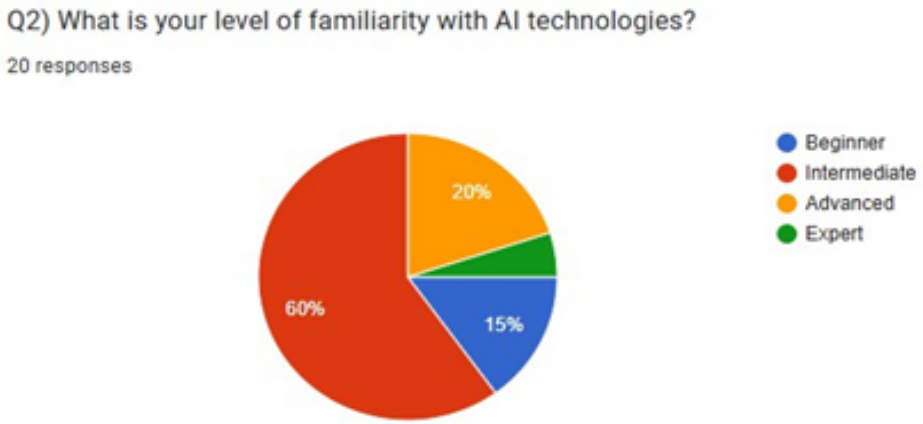
awareness in usage of ethical AI in public policy making) and independent variables (principles of ethical AI, AI technologies etc.). Accordingly, the following set of hypotheses has been framedkeeping in mind the objectives of the study:

- "Respondents will completely be aware about the AI technologies"
- "Respondents will know about the core principles of Ethical AI"
- "Respondents are aware about the applications of AI in public policymaking"
- "Respondents will know about the violations in ethical AI"
- "Respondents are awareabout the stakeholders involved in creatingguidelines for ethical AI in public policy making"

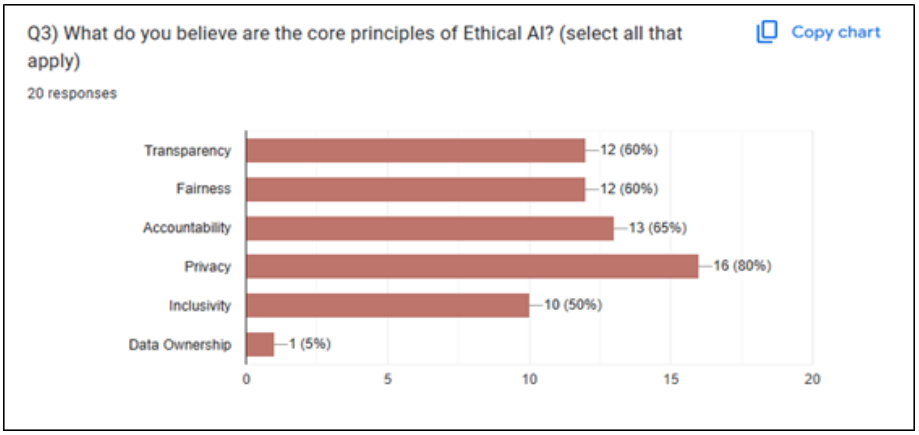
DATA REPRESENTATION & INTERPRETATION



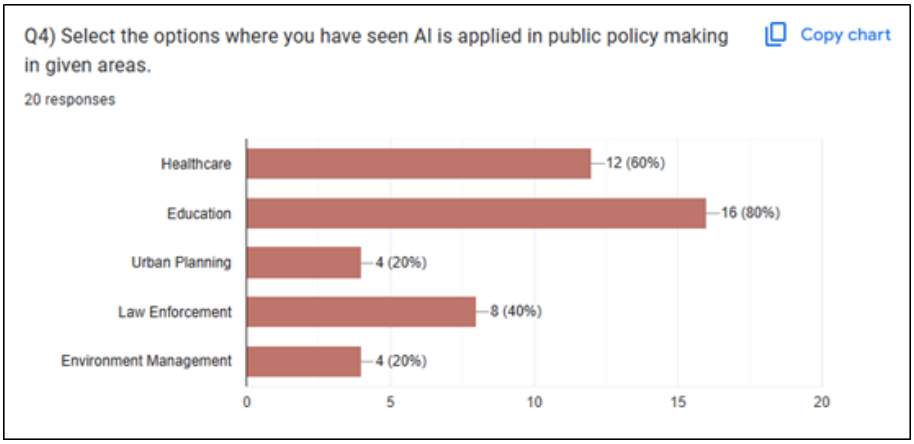
This survey includes 20 responses of people working in government sector with different designations such as Section Officers, Assistant Section Officers, Personal Assistant, SSA etc. They have shown active participation in the survey.



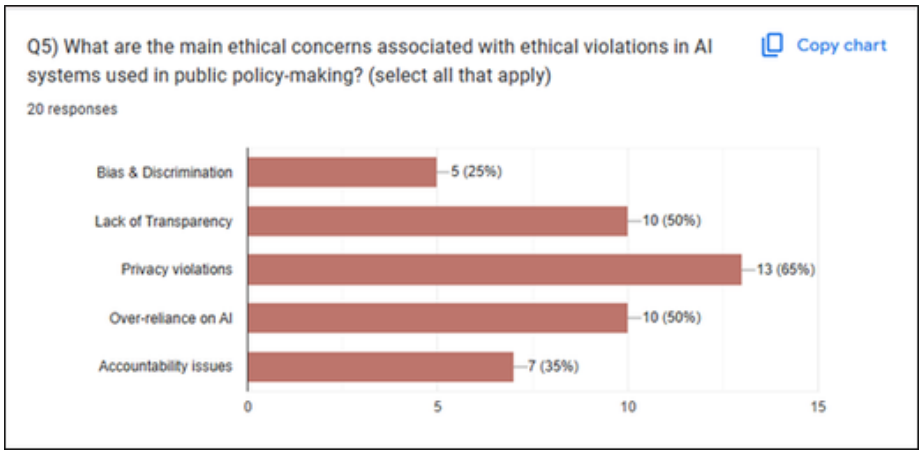
The respondents were asked abouthow familiar they are with AI technologies and the levelof familiarity comprises 60% are somehow familiar belongs to intermediate level; 20% of them have advance knowledge whereas only 5% belong to expert level which highlights the need for awareness to increase the percentage as of intermediate level. The rest 15% are beginners who are keen to learn more about usage of AI technologies.



Here the respondents were questioned on their beliefs about the core principles of Ethical AI in public policy making where 80% respondents believe that privacy is the most important principle among all. 60% weightage has been given to both transparency and fairness, 65% to accountability and 50% to inclusivity.

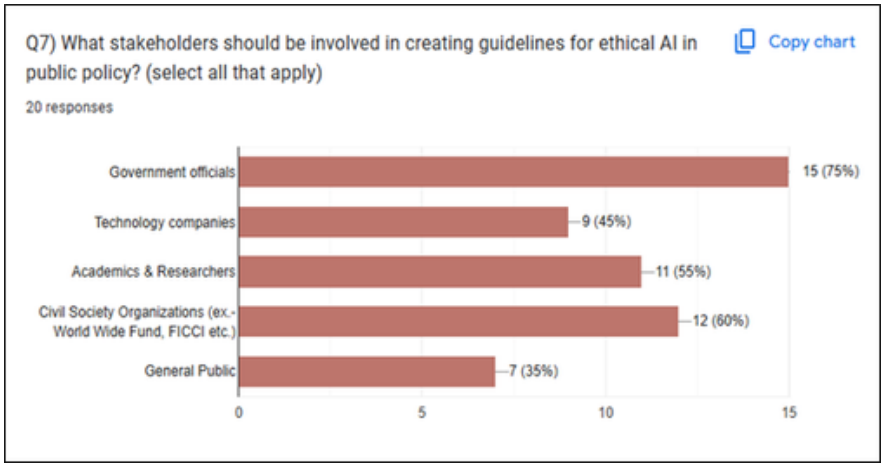


The respondents were asked where they have seemed AI is applied in public policy, 80% responded to education sector, 60% have seen in healthcare sector, 40% opted for law enforcement and 20% for both urban planning and environment management.

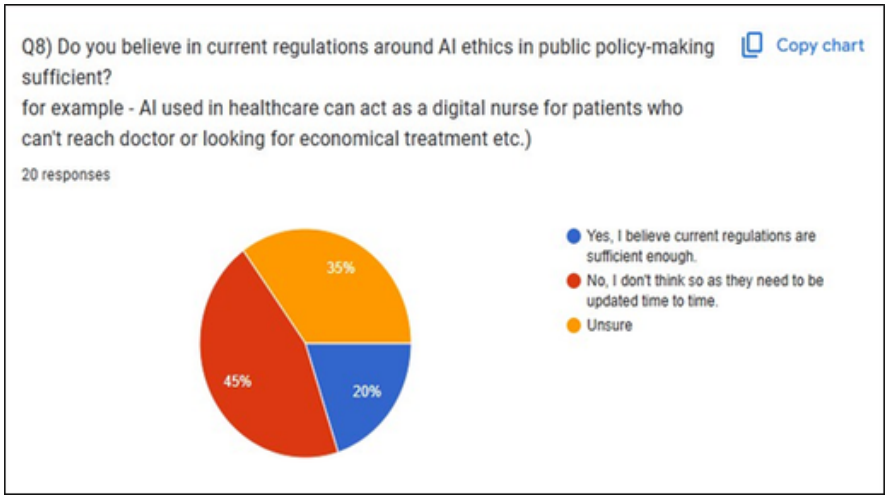


When it comes to ethical violations, 65% of the respondents are concerned about privacy violations, 50% for lack of transparency and another 50% for over-reliance of AI. For biasness and discrimination 25% are concerned and 35% are concerned for accountability issues.

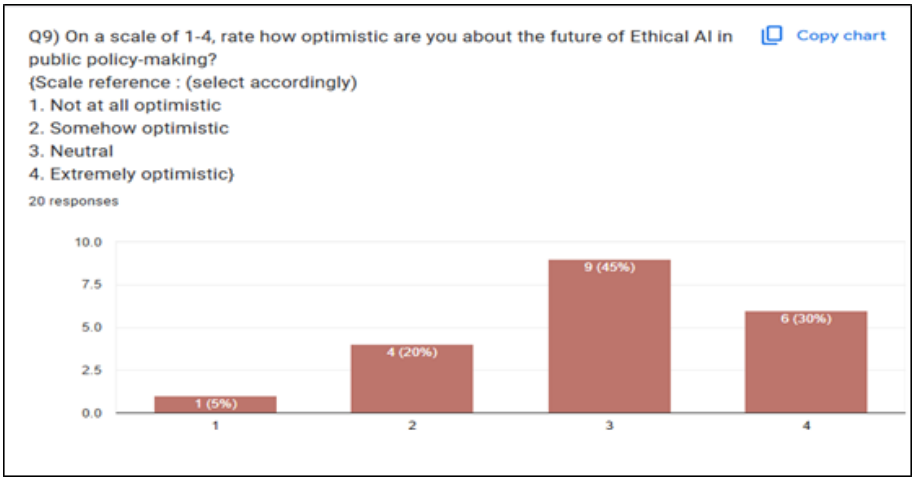
On the basis of 'level of concern', on a scale of 10 maximum respondents are extremely concerned about Ethical violations in AI systems needs to be addressed for the development of good public policy framing.



Mostly respondents opted 'government officials' as the idealstakeholders in creatingguidelines ethical AI in public policy and also 60% of them chose Civil Society Organizations, whereas the rest 55% believed in academics &researchers, 45% of them believedin technology companiesand 35% of them believed in general public.



As for current regulations are up to the mark or not in the eyes of the respondents, 45% of them believed that the currently applied regulations around AI ethics in public policy making should be updated from time to time in order to stick to the principles of Ethical AI, 35% gave unsure response, 20% who might not be very much aware about the current regulations believed they are sufficient.



This question was based on the rating of respondents' level of optimism about the future of Ethical AI in public policy making where 45% are neutral about what will happen in future of AI, 30% are highly optimistic and believe that such a change will bring positive growth and development.

The respondents were asked to put up their additional views that shouldbe addressed to advance Ethical AI in public policy making, some of the views are –

- Ethical AI requires innovation, human decision makingand regular maintenance.
- Personal Data can be protected via Data localisation.
- Ethical AI musts integrate insights from diversefields, including sociologyetc.
- Privacy transparency, Security concerns, free from bias, data ownership, fairness etc. highlights the concerns which are needed to be addressed.

Lastly, they were asked to share their last thoughts or experiences related to Ethical AI in public policy making, some of the highlights are as follows; -

- “AI has potential to revolutionize public services and align with the ethical principles I to maximize social benefits.
- “New Algorithms are being created to differentiate betweenhumans and generative AI Avatars”
- “More AI usagewith human intervention for healthy purposeleads to healthydevelopment of our country in the age of advancement.”
- “Trust in AI systemis essential for public acceptance”

Conclusion

The integration of ethics in public policy, particularly with the adoption of artificial intelligence, holds immense potential to transform governance and societal welfare. By emphasizing principles such as fairness, transparency, accountability, and inclusivity, policymakers can ensure that AI technologies are harnessed responsibly and equitably. Case studies in surveillance, healthcare, education, and agriculture demonstrate the versatility and challenges of ethical AI applications, highlighting both the promise and the need for vigilance against ethical violations such as bias, privacy breaches, and lack of transparency.

The survey findings underscore a significant awareness gap among public officials regarding the full potential and ethicalimplications of AI, suggesting the necessity for enhanced trainingand stakeholder collaboration. Moving forward,a robust frameworkof guidelines, informedby diverse stakeholder input and periodic updates,is essential to build public trust and ensure the alignment of AI with human values.

As the age of technological advancement accelerates, fostering innovation while safeguarding ethical principles will be pivotal to achieving sustainable development and equity in public policy. With deliberate effort, ethical AI can revolutionize governance and contribute to a just, transparent, and inclusive society.

Road Ahead in Ethical AI

As the interplay between ethics, public policy, and AI grows, a forward-looking roadmap is crucial to navigate challenges and opportunities. Key areas for future focus include:

- Strengthening Regulatory Frameworks
- Establish adaptable, global ethical guidelines that address privacy, transparency, and accountability while allowing for local context-specific applications.
- Capacity Building and Awareness
- Promote understanding of AI ethics through training programs and workshops for policymakers, government officials, and the public to enable informed decision-making and reduce ethical violations.
- Fostering Multi-Stakeholder Collaboration
- Ensure inclusive AI systems by involving governments, civil society, researchers, and technology developers, integrating diverse perspectives for equitable outcomes.
- Investing in Research and InnovationSupport research into ethical AI, focusing on explainable models, bias detection algorithms, and privacy-preserving technologies, while fostering interdisciplinary approaches to expand AI's societal impact.
- Enhancing Data Governance Develop robust data governance policies emphasizing fairness, transparency, and citizen privacy to build trust in AI-driven services.
- Piloting Ethical AI Models
- Test AI solutions in real-world scenarios, such as healthcare and agriculture, to refine systems and address ethical dilemmas before scaling them broadly.
- Periodic Monitoring and Feedback Implement regular evaluations and public feedback loops to ensure AI systems evolve dynamically and remain aligned with societal values.

By focusing on these priorities, ethical AI can transform public policy, aligning technological progress with societal well-being and governance. Collaborative action and adherence to ethical principles will be key to building a future where AI empowers governance and quality of life for all.

Road Ahead in Ethical AI

I extend my heartfelt gratitudeto everyone who contributed to the successful completion of this research project. The invaluable support, constructive feedback, and guidance I received throughout this journey have been instrumental in shaping the final outcomes of this work.

I deeply appreciate the encouragement

and assistance provided by my institution, as well as the resources and facilities that enabled me to conduct this research. Additionally, I am gratefulfor the thoughtful discussions and collaboration that enriched the scope and depth of this study.

Lastly, I wouldlike to acknowledge the unwavering supportand motivation from my familyand peers, which inspired me to persevere through every challenge along the way.

References

- *Artificial intelligence in government. https://www.sciencedirect.com/science/article/pii/S2773067022000048*
- *Council of Europe. (n.d.). Common ethical challenges in AI. https://www.coe.int/en/web/human-rights-and-biomedicine/common-ethical-challenges-in-ai*
- *InRule. (n.d.). Risk mitigation and ethical decision-making. https://inrule.com/blog/risk-mitigation-and-ethical-decision-making/*
- *Shah, U. (n.d.). AI governance, emerging technologies, and future trends. LinkedIn. https://www.linkedin.com/pulse/ai-governance-emerging-technologies-future-trends-shah-uaprc*
- *Santosh, G. (n.d.). AI governance: An imperative call for international cooperation. LinkedIn. https://www.linkedin.com/pulse/ai-governance-imperative-call-international-cooperation-santosh-g-lptgc#:~:text=International%20collaboration%20is%20necessary%20to,global%20scale%20necessit*
- *Barrett, M. (n.d.). Future challenges and opportunities in AI ethics. LinkedIn. https://www.linkedin.com/pulse/future-challenges-opportunities-ai-ethics-michael-barrett*
- *V7 Labs. (n.d.). Artificial intelligence in government. https://www.v7labs.com/blog/ai-in-government*
- *Appinventiv. (n.d.). Artificial intelligence case studies. https://appinventiv.com/blog/artificial-intelligence-case-studies/*
- *Digital Defynd. (n.d.). Artificial intelligence case studies. https://digitaldefynd.com/IQ/artificial-intelligence-case-studies/*
- *World Economic Forum. (2016, October). Top 10 ethical issues in artificial intelligence. Retrieved from https://www.weforum.org/stories/2016/10/top-10-ethical-issues-in-artificial-intelligence/*
- *UNESCO. (n.d.). Recommendation on the ethics of artificial intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics*
- *UNESCO. (n.d.). Recommendation on the ethics of artificial intelligence: Case studies. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases*
- *Augusta University. (n.d.). Cybersecurity and ethics. https://www.augusta.edu/online/blog/cybersecurity-ethics/*
- *InRule. (n.d.). Risk mitigation and ethical decision-making. https://inrule.com/blog/risk-mitigation-and-ethical-decision-making/*
- *Indian Society of Public Policy. (n.d.). Ethics of AI in public policy in the Indian context. https://www.ispp.org.in/ethics-of-ai-in-public-policy-in-the-indian-context/*
- *Facebook AI Research. (n.d.). DensePose. DensePose. https://densepose.org*
- *Synced. (2023, January 17). CMU's DensePose from Wi-Fi: An affordable, accessible, and secure approach to human sensing. https://syncedreview.com/2023/01/17/cmus-densepose-from-wifi-an-affordable-accessible-and-secure-approach-to-human-sensing*
- *Shilton, K., & Juen, J. (2020). Artificial intelligence in surveillance: Implications for privacy and security. Journal of Information Privacy and Security, 16(3), 45-64. https://doi.org/10.1080/15536548.2020.1793254*
- *Gupta, R., & Sharma, A. (2021). Ethical considerations in AI-based digital agriculture: Transparency, fairness, and data accuracy. Journal of AI and Agriculture, 12(3), 55-67*
- *Singh, K., & Verma, P. (2020). Leveraging AI in Indian agriculture: Impacts and challenges. International Journal of Agricultural Innovation, 8(2), 102-115.*

ENHANCING SOFTWARE SECURITY: A STUDY OF SIMPLIFIED SECURE SOFTWARE DEVELOPMENT FRAMEWORK

N. SATYANARAYANA[1], V HARISH[2], RAMYA CHITRA T P R[3], R SWATHI[4], AND DR. AMIT K DHAR[5]

[1]SCIENTIEST – E, C-DAC, HYDERABAD
[2]PROJECT ENGINEER – E, C-DAC, HYDERABAD
[3]SENIOR PROJECT ENGINEER – E, C-DAC, HYDERABAD
[4]SENIOR PROJECT ENGINEER – E, C-DAC, HYDERABAD
[5]ASSOCIATE PROFESSOR, IIT BHILAI

ABSTRACT

The world is moving towards digital services and at the same time the attacks and exploitation of software vulnerabilities are also increasing. This situation not only leads to the lack of trust in software quality but also to the loss of business opportunities. Identifying security issues only during the software testing phase often leads to project delays and budget overruns. Focusing on security aspects in every phase of the software development lifecycle is an effective strategy, as it enables software architects, developers, and testers to understand their individual responsibilities within their respective phases while promoting a collaborative approach to addressing security across the entire development process. The objective of this paper is to present a simplified secure software development framework that explains an implementation strategy that can be followed by project teams in developing secure software. Authors have conducted a detailed study of various software vulnerabilities, their impact and root cause of errors and proposed a secure SDLC framework that suggest a methodology called as "P6" (Prepare, Practice, Protect, Produce, Probe, and Process Metrics) as an appropriate strategy to deal with the factors contributing to security concerns effectively. Authors have taken a sample project work that has no security aspects incorporated into it as a case study and incorporated the best practices, standards, tools, techniques and strategies relevant to each phase in-line with the proposed secure SDLC framework. Based on the efforts a set of practices that can be followed has been listed in this article.

Keywords: Secure SDLC, Threat Analysis, Attack Vector Reduction Techniques, Security Metrics

Introduction

The world is witnessing an increasing trend in cyberattacks on critical infrastructures and other digital services, causing financial and business losses. Ransomware, data leakage, and malware injection are becoming common attack types. These attacks are raising concerns over preparedness against cyber security. Most cyberattacks originate from poor code vulnerabilities, network loopholes, or software configurations that reveal sensitive information. If we look at the cyber incident scenario there is an increase in the number of vulnerabilities discovered in 2023 compared to previous years and some of them are dangerous to cause enormous loss to the organizations. As per (Forbes, 2023), the cost of cybercrime is predicted to grow to $10.5 trillion by 2025. Some of

the factors that lead to this phenomenon include the usage of opensource software (as per Synopsys 2023 report, at least 84% of code bases contain vulnerabilities), Phishing attacks as a preferred method through non-email based methods like vishing (voice phishing), smishing (sms based), and quishing (QR code) increased by 7 fold by 2022, branding misuse (by taking the name of Microsoft, Amazon, etc.), business email compromise by 78%, identity theft, etc. Hence, it is important to develop software applications that focus on these kinds of risks and with appropriate mitigation plans.

User training may be an effective way to handle social engineering attacks but still following secure software development practices is a reasonable solution to get shielded against cyberattacks. For example, the famous Heartbleed bug in the Open SSL library is a result of a vulnerable single-line code memcpy(bp, pl, payload); in Open SSL that permitted the leaking of sensitive information from the server during the exchange of Heartbeat messages between the client and server machines. In this example, explained by (Josh Fruhlinger, 2022), there was never a check whether the memory pointed to by pl is of size matching the payload value or not. Cybersecurity is about having robust defensive mechanisms or understanding hacking techniques and having a solid foundation in developing secure software strengthens our position in combating cyber threats.

In software development, we observe project managers or team leads get involved in software design and architectural issues while developers write the code. Security testing is usually done at the time of software deployment. However, this practice derails the project schedule and budget as it requires more time and effort to resolve the security issues that are uncovered belatedly as well as repetition of the same cycle of getting it tested again. Hence, the security shift-left concept is gaining importance in the software industry.

With security shift-left implementation, the focus on security begins with requirement collection and continues in design and implementation until final deployment. This phenomenon creates an environment where all players involved in software development are equally responsible and aware of security concerns from the beginning of the planning phase to the deployment phase. We also need an effective Secure Software Development Lifecycle framework mechanism at the organizational level that helps Prepare, Practice, Protect, Produce, Probe, and Pass. In this paper, we will discuss the P6 approach, which can be regarded as a code of practice that ensures all necessary steps have been taken in producing secure software.

Related Study

Poor programming practices lead to software with vulnerabilities or weaknesses. Hardcoding of user credentials in database or server configuration files e.g., exposing of personal information of Uber users in 2016 as narrated by (Dara Khosrowshahi, 2017), (Vamsii777, 2023) published an incident where secret keys was exposed in public code repository of a project on github website, improper logging of sensitive data are some of the poor programming or misconfiguration incidents that resulted in malicious attacks or exploitation of vulnerabilities.

A mapping of different types of security vulnerabilities and their root cause gives some insights about the issues to be focused upon while building secure software. Table 1 below gives such mapping from Common Weaknesses Enumeration for the benefit of readers of this paper.

Sl.N.	Vulnerability	Impact	Cause of Error	Mishandling	Real-world Exploitation of vulnerability
1	Out of bounds write	Data corruption, crashes or execution of arbitrary code	Improper bounds checking before writing data	Memory	
2	Out of bounds read	Information Disclosure	Improper bounds checking before reading data		
3	Integer overflow or wraparound	Logic error or buffer overflow	Failure to check integer boundaries		Heartbleed in Openssl versions reported by Synopsys Inc. (2020), Bitcoin Integer Overflow Bug (Bicoin Core, 2021)
4	Improper Restriction of Operations within the bounds of a memory buffer	Buffer overflows and code execution	Improper bounds checking for memory operations		
5	Use after free	Data corruption, crashes, or execution of arbitrary code	Reusing memory after it has been freed		
	Null pointer dereferences	Program Crash	Accessing memory that has not been initialized		
6	Server-side request forgery	Unauthorized internal resources	Failure to sanitize URLs or IP Addresses	User Input or Network Data	
7	Improper neutralization of input during web page generation, SQL/OS commands (XSS attacks, SQL Injection, Command Injection)	Unauthorized access or execution of arbitrary code	Failure to validate and sanitize user input		Cisco ASA WebVPN login page exploited by Androxghost botnet (Cisco Systems Inc. 2014) SSTI exploitation in WordPress Multilingual Plugin (Wordfence, 2024)
8	Deserialization of Untrusted Data	Code execution of data manipulation	Deserializing data from untrusted sources		Apache Commons collections (The Apache Software Foundation, 2024), and Atlassian Confluence (Benny Jacob, 2021) are a few examples.
9	Improper Limitation of a Pathname to a Restricted Directory	Information Disclosure	Failure to sanitize or validate input paths		Grafana Path Traversal exploited during code review (Vardan Torosyan, 2021)
10	Improper authentication	Unauthorized access	Inadequate or flawed authentication mechanism	Authorization and Authentication	Authentication bypass vulnerability in GlobalProtect SSL VPN component (Paloalto, 2020)
11	Missing authorization	Unauthorized access to resources	Failure to enforce proper authorization controls		
12	Missing authentication for critical function	Unauthorized access	Failure to enforce authentication for critical operations		
13	Improper privilege management	Privilege escalation	Incorrect assignment or enforcement of user privileges		
14	Incorrect authorization	Unauthorized access	Failure to enforce correct authorization policies		
15	Incorrect default permission	Unauthorized access	Using permissive default settings		
16	Use of hard-coded credentials	Unauthorized access	Hardcoding sensitive information		Based on Gitguardian Report (GitGuardian, 2024)
17	Cross-site request forgery	Unauthorized actions on behalf of a user	Failure to validate the authenticity of use requests		
18	Unrestricted upload of files with dangerous type	Execution of malicious files	Failure to restrict file types during upload	Programming Logic	
19	Concurrent execution using shared resources with improper synchronization	Race conditions and data corruption	Failure to synchronize access to shared resources		
20	Improper control of the generation of code	Code Injection	Improperly handling or generating code dynamically		Log4J a java logging library exploit (Apache Software Foundation, 2021)

If we look at the above incidents, it is evident that the exploitation of vulnerabilities resulted from improper programming practices. Proper input validation and sanitization techniques, following the zero-trust principle, whenever data is read from outside of program scope (viz., file systems, network streams, user inputs, etc.,) use of secure libraries, avoiding exposure of credentials through configuration files, etc., could have prevented the majority of the security issues.

While these incidents give the developers an important lesson, the next question is how a developer can be sure of the presence of vulnerabilities in his code. Several techniques, tools, or vulnerability databases are available such as Common Platform Enumeration IDs to keep track of affected components from the CVE databases, Software Bill of Materials, Vulnerability databases, etc. The only concern left is how effectively these tools, techniques, or strategies are being used in software development, and what processes exist in an organization that reflects the organization's security vision.

The Secure Software Development Framework (NIST-SP-800-218 & 218A) (National Institute of Standards and Technology (NIST), 2022, 2024)) is a set of fundamental, sound, and secure software development practices based on established secure software development practice documents from organizations such as Business Software Alliance (BSA), Open Worldwide Application Security Project (OWASP), SafeCode. The set of practices defined by SSDF is "Prepare the Organization", "Protect the Software", "Produce Well-Secured Software", and "Respond to Vulnerabilities". (Microsoft. (2024) also has released its guidelines on secure software development practices. While Microsoft's effort lists out best practices in developing secure software, NIST's SSDF brings out a holistic approach to handling the secure software development culture within an organization.

Proposed Secure Software Development Framework – P6

We propose a P6 methodology which can be regarded as a simplified version of the Secure Software Development Framework of NIST SP 800-218. The fundamental difference between the proposed P6 approach and the SSDF framework is the former provides a simplified process that delineates the activities to be performed and the later describes the best practices that an organization has to follow. Apart from that the former describes various metrics that help in identifying how effectively the P6 approach is followed in software development.

A P6 (**P**repare, **P**ractice, **P**rotect, **P**roduce, **P**robe, and **P**rocess Metrics) approach establishes the baseline for the proposed secure software development framework. Here, P6 stands for,

Prepare: Prepare the development teams to clearly define the security requirements and security traceability matrix of an application being developed. Further, prepare each application development team to handle the application security in each phase of SDLC with appropriate training about process documents, templates, and tools to be used. Also, prepare a local repository of the vulnerability database indexed by a Unique ID where identified security issues can be stored. For example, the National Vulnerability Database managed by NIST

uses the CPE (Common Platform Enumeration) pattern to retrieve matching CVEs (Common Vulnerabilities and Exposures). Inculcate the Zero-Trust principle during all phases of the software lifecycle.

Practice: Appropriate templates to capture the security requirements, abuse cases, traceability matrix, design principles, attack vectors (local, physical, network), known vulnerabilities in third-party libraries, threats, and secure coding guidelines mapped to the threats and known vulnerabilities, for an application being developed are to be developed and ensure that all application teams use them at each phase of the software development lifecycle.

Protect: Ensure that all the artifacts viz., specifications, documents, source code, configuration files, API Keys, etc., are accessible through a multi-party signing mechanism and two-factor authentication, and ensure that these artifacts do not contain sensitive information in plain text format. Further ensure each artifact is assigned a unique ID representing its version, and year of release to keep track of which version of the artifact is being referred to in the development process.

Produce: The software has to be produced securely by ensuring compliance with all the templates designed during the requirements, and design phases including prescribed coding guidelines. During this phase, it must be assured by the development team that all the identified security issues and corresponding mitigation plans have been duly verified and addressed while writing code.

Probe: Perform thorough testing using different scanning tools for known vulnerabilities and weaknesses in the software referring to CVE and CWE entries. Apart from that always keep track of new entries in vulnerability databases and conduct application scanning at regular intervals to identify new threats even after delivery of the software. This is to ensure a quick reaction to the security issues uncovered post-launch of the software.

Process Metrics: Ensure that security process metrics have been measured in each phase of SDLC during software development. For example, keep track of how many security requirements have been identified and how many of them have been closed and are still in open state. The security requirements usually would be addressed during the design and implementation phases and hence as and when the requirements are resolved, corresponding metric values have to be updated in relevant phases in the upward direction. Likewise, measure how many critical, high, low-security issues are uncovered during the design phase and how many mitigation plans have been defined, and how many of them are left open without any mitigation plan. This process ensures the efficiency of underlying processes.

In the next section, we will see an implementation of the P6 approach and how the efficiency of the process can be tracked.

Implementation of P6 Approach

As far as the “Prepare” activity is concerned, the authors have organized a training program for government officials from different organizations, such as UTI Infrastructure Services Limited, BSNL, Airport Authority of India, etc., about the complete Secure SDLC practices from the beginning of the requirements phase to the deployment phase. The training program focused more on strategies, tools, and techniques to handle different scenarios at each phase of software development.

The training program covered the Application Security Verification Standard (ASVS) from (OWASP Foundation, 2021) which explains various security requirements that need to be focused on during web application development. Apart from that how to capture the security requirements through abuse cases and prepare a security requirements traceability matrix and integrate them with the business requirements document is also explained. Further various software architecture and design principles, attack vector reduction techniques, privacy threat modeling, application threat modeling and identification of appropriate mitigation plans, secure coding guidelines, code review tools, and associated standards, continuous development/deployment pipeline development and integration of vulnerability scanner with CI/CD, measuring of process metrics. Figure 1 below depicts the input and output artifacts generated in Secure SDLC practices.

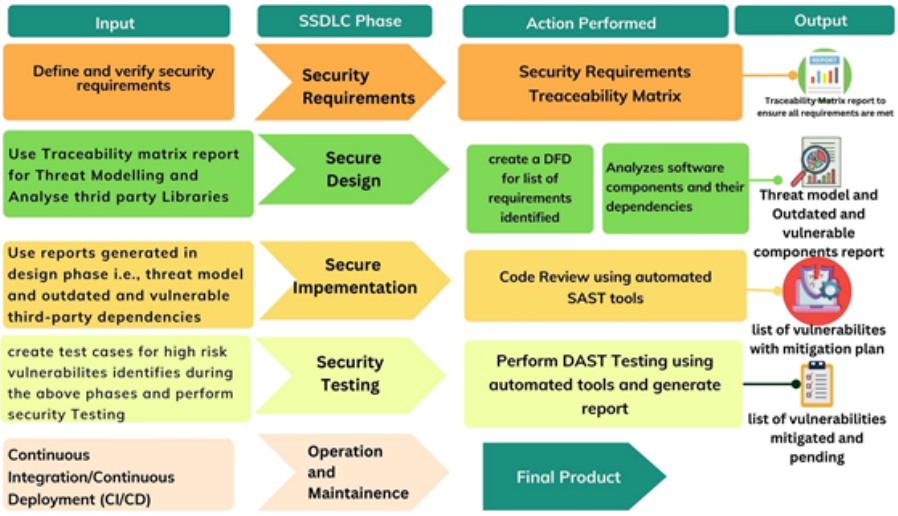


FIGURE 1: SECURE SDLC PROCESS – INPUT/OUTPUT ARTIFACTS

As part of the "Practice" activity, the authors have developed templates to capture various details in line with the P6 approach that helps the intended stakeholders to focus on requisite information during software development.

For example, figure 2.1 depicts the security requirements traceability matrix (SRTM) template that can be prepared based on the inputs provided through ASVS standards and stakeholders' responses to the security requirements questionnaire.

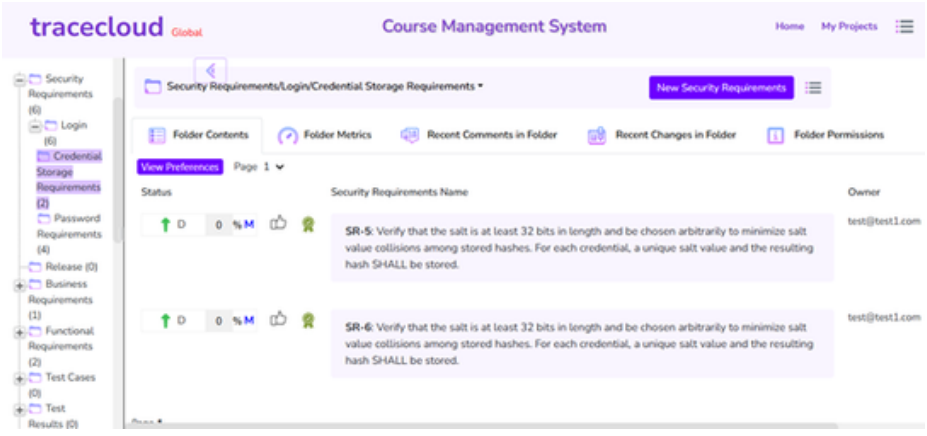


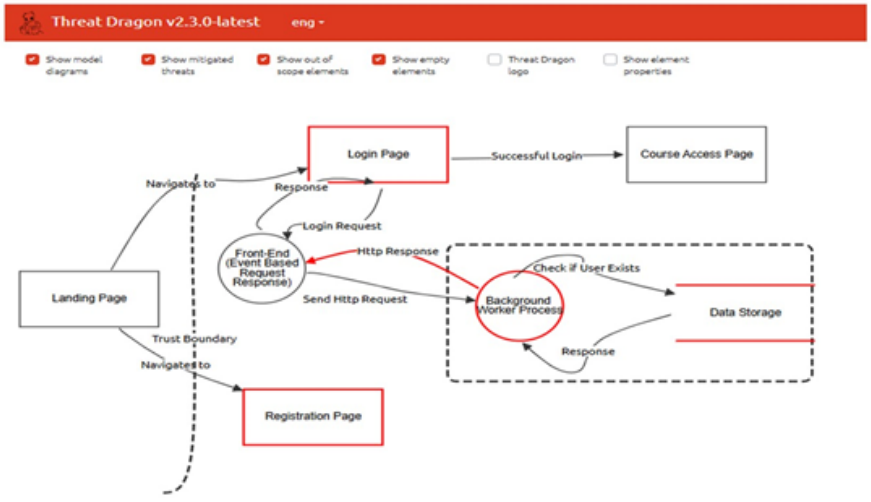
FIGURE 2.1: TRACECLOUD SOFTWARE DASHBOARD FOR GENERATING SRTM

The SRTM report shown in Figure 2.2, helps with tracking security requirements aligned with each business requirement in both forward and backward directions in the application functionality flow to track the current status of identified security issues. These details help maintain requisite metrics data that can be measured continuously to monitor the project's overall security posture at a given instant.

Tag	URL To Requirement	Versio	Name	Owner	Testing Status	Percent	Approv	Trace To	Trace To - With Details
SR-1	https://www.tracecloud.com/GloreeJava2/servlet/DisplayAction?dO=req&dReqId=1308447&projectId=8127	1	Verify that user shall set passwords are at least 12 characters in length. (C6)	test@test1.com	Pending	0	N/A	BR-1,FR-1	BR-1 :CMS shall have login feature FR-1 :User shall able to login
SR-2	https://www.tracecloud.com/GloreeJava2/servlet/DisplayAction?dO=req&dReqId=1309094&projectId=8127	2	Verify that user password credential recovery shall not reveal the current password in any way.	test@test1.com	Pending	0	N/A	BR-1,FR-1	(s)BR-1 :CMS shall have login feature (s)FR-1 :User shall able to login
SR-3	https://www.tracecloud.com/GloreeJava2/servlet/DisplayAction?dO=req&dReqId=1309095&projectId=8127	1	Verify that password change functionality shall require the user's current and new password.	test@test1.com	Pending	0	N/A	BR-1,FR-1	BR-1 :CMS shall have login feature FR-1 :User shall able to login
			Verify that passwords submitted during account registration, login, and password change are checked shall be a checked against a set of breached passwords either locally (such as the						BR-1 :CMS shall have login feature

FIGURE 2.2: SAMPLE SRTM REPORT FORMAT

These requirements (both functional and security) will be useful during the design phase while preparing a threat modelling report.



Node Server (Process)

Description:
NodeServer will receive the data and send it to Google authenticator(App) for OTP generation.

Generic spoofing threat
Spoofing, Open, Medium Priority

Description:
A generic spoofing threat

Mitigation:
Mitigation or prevention for the threat

Generic tampering threat
Tampering, Open, Medium Priority

Description:
A generic tampering threat

Mitigation:
Mitigation or prevention for the threat

FIGURE 3.1 SAMPLE THREAT MODELLING REPORT FORMAT

Also, other artifacts such as threat modelling report (OWASP Foundation, 2020) from the application security perspective as well as privacy perspective using STRIDE, PASTA, DREAD, etc., and LINDUNN models (Lindunn Examples, 2024) and (Kim Wuyts, Wouter Jossen. July 2015), provide insights about how effectively the software architecture principles have been incorporated and attack vectors are identified and reduced during software design and what kind of mitigation plans have been suggested to the software development team. For example, Figure 3.1 provides a sample threat modeling report based on a fictitious project named "Course Management System" being used as a driving force to explain different application security concepts. Figure 3.2 provides a glimpse of how the "Practice" activity can be performed by the project teams.

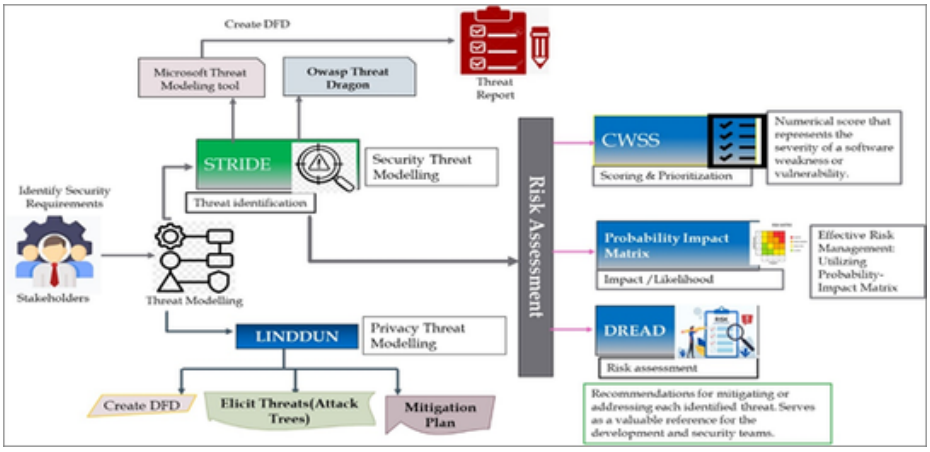


FIGURE 3.2 PROCESS FLOW FOR PERFORMING "PRACTICE" ACTIVITY

As far as the "**Protect**" activity is concerned, a git repository can be maintained with a strict access policy in place with appropriate rights such as 'Owner', 'Maintainer', 'Collaborator', 'Viewer', and 'Administrator' so that members with those rights will be able to access the content generated through the above processes.

As far as the "**Produce**" activity is concerned, the software developers can be provided with tools such as SonarLint, and SonarQube to perform code reviews during development time and while integrating with other components of the project at the centralized server.

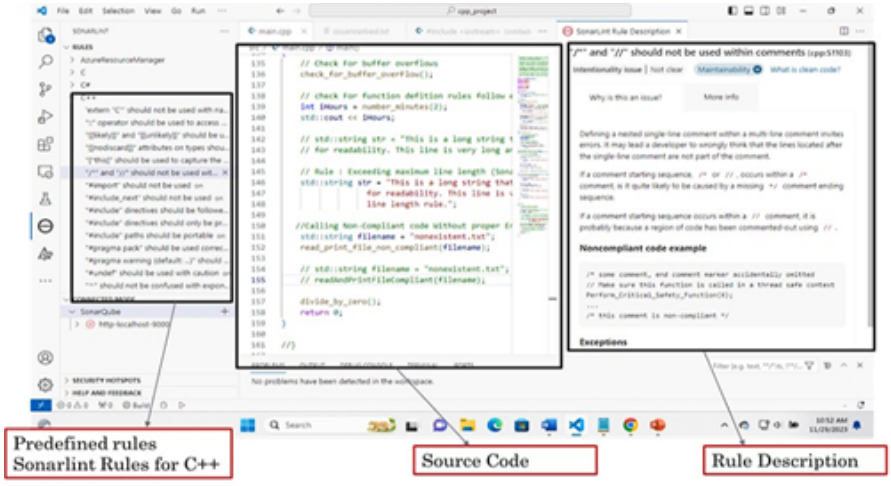


FIGURE 4: SCREENSHOT OF ENFORCEMENT OF SECURE CODING GUIDELINES THROUGH SCANNERS

Figure 4 depicts how the code review can be performed using the SonarLint plugin integrated with the VSCode IDE. The authors are not recommending or endorsing any of these tools and it is at the liberty of the software development teams to use various tools as per their respective programming and environmental context. This ensures that the software developers will get requisite support from the underlying plugins or tools that enforce strict compliance with the secure coding guidelines prescribed by CWE (MITRE, 2024) / SEI Cert (Carnegie Mellon University, 2024) / ISO standards (ISO/IEC, 2019). Further, the project team has to continuously monitor for vulnerabilities in the source code during development time as well as during the integration of various modules into the project repository. During this time it is essential to run both SAST and DAST scanning to uncover any potential errors that might have crept in due to the merger of several project files in the final project repository. SonarLint and OWASP ZAP tools can be used to perform Static Application Security Testing and Dynamic Application Security Testing as part of the CI/CD pipeline to ensure continuous monitoring activity. Apart from that, the generation of the Software Bill of Materials gives a details report of various components, licenses, and vulnerabilities present in the project repository. After performing these steps, one can be assured that all requisite checks have been performed before releasing the software product.

As far as the **"Probe"** activity is concerned, post-deployment of the software also, the project team has to keep monitoring any new CVE/CWE entries in the vulnerability databases maintained at the organizational level or global level and update the project teams accordingly. This activity ensures a quick response time for the newly uncovered issues in the existing as well as ongoing projects. One of the reasons for data breaches and other cyberattacks in the present scenario is delayed action from the security/product teams in applying patches quickly.

As far as the **"Process Metrics"** are concerned, while SAST and DAST tools enforce secure coding guidelines, the developers are required to close the security threats uncovered during threat modelling by implementing respective mitigation plans. The SAST or DAST tools, do not verify whether every mitigation plan has been implemented by the developer successfully or not. Hence, the templates designed during the requirements and design phases and associated metrics give first-hand information about the current posture of the application security. For example, Figure 5, shows the part of metrics concerned with tracking the security posture of applications being developed at a given instance of time.

Phase	Measures	Metric
Planning	No of identified security analysts	
	No of the identified testers	
	Employees trained on security policies	
	No of SAST tools used	Security Tools Utilization = No of SAST tools used + No of DAST tools used
	No of the DAST tools used	
Requirements	No of identified security specifications	
	Targeted security assurance level	Targeted security assurance level: Level 1, Level 2, Level 3
	No of applicable security controls	
	No of identified security requirements	No of the identified security requirements = No of High Priority Security Requirements + No of Medium Priority Security Requirements + No of Least Priority Security Requirements
	No of Least Priority Security Requirements	
	No of Medium Priority Security requirements	
	No of High Priority Security requirements	
	No of participating stakeholders representing a different category	

Phase	Measures	Metric
Design	No of the design principles followed	
	No of entry points	Attack Vector Density = No of attack vectors (post threat analysis)/No of entry points
	No of attack vectors (post threat analysis)	
	No of the identified threats	Threat Management Value = (No of the identified mitigation plans/No of the identified threats)
	No of the identified mitigation plans	
	No of the identified high-priority threats	
Coding	No of source code files	Proportion of Defect Density = No of vulnerabilities/No of source code files
	No of vulnerabilities	
	No of the identified vulnerable code fragments / KLOC	Defect Density = (No of the identified vulnerable code fragments / KLOC * 1000) / total lines of code
	No of fixes done	
	No of mitigation plans implemented	Vulnerability Density = No of mitigation plans implemented/No of Vulnerabilities
	No of runs of the SAST tool before the commit	
	No of total commits	
	No of runs of SAST tool after commit	
	No of runs of DAST tool after commit	

Phase	Security Measures	Metric
Testing	No of modules	Module Test Coverage = (No of modules tested / No of modules)
	No of the modules tested	
	No of test cases failed	Security Test Ratio = No of test cases failed / (No of test cases passed + No of test cases failed)
	No of test cases passed	
	No of identified threats during the design phase	UnresolvedSecurityIssues = No of test cases failed / No of identified threats during design phase
	No of pending high risk threats	
	No of pending medium-risk threats	
	No of pending low-risk threats	
	No of communications sent	
	No of communications sent to the development team	
	No of communications sent to the design team	
	No of SAST tools used	Security Testing Intensity = (No of SAST tools used * No of runs of SAST tool) + (No of DAST tool used * No of runs of DAST tool)
	No of the DAST tools used	
	No of runs of DAST tool	
	No runs of the SAST tool	
	No of Manual Testing Attempts	
	No of detected vulnerabilities during manual testing	

FIGURE 5: APPLICATION SECURITY PROCESS METRICS

Discussion

The authors believe that every organization may define its security vision and goal and a policy document for which the P6 approach would serve as a code of practice. The following steps may be initiated at an organization level to keep up its compliance with the P6 approach to maintain application security.

- Train the developers, project managers, and testers on secure software development practices emphasizing zero-trust principles in every facet of the software development lifecycle.
- Mandatory implementation of templates to capture security aspects focused upon at each phase of the software development lifecycle.
- Apply appropriate threat modelling techniques to understand potential risks involved in application development from a security perspective, as no single threat modelling technique is an all-in-one solution for every situation.
- Maintain a centralized repository or retrieve information from a trusted source about components that are either outdated or have known vulnerabilities.
- Update downstream phases in the software lifecycle about the security issues that are resolved in later phases successfully and update the values corresponding to various metrics to measure the efficiency of SSDLC processes.
- Post-deployment of application software conduct scanning of application security at regular intervals as a quick response to release relevant patches.
- Remove obsolete entries from the vulnerability databases at regular intervals if the organization maintains its vulnerability database.

Acknowledgment

The proposed P6 SSDLC Framework is inspired by the discussions held about the training material developed by the authors. The authors express their sincere thanks to all the reviewers and officials from C-DAC and MeitY and other premier academic, industry, and government organizations for their valuable suggestions about the Secure Software Development Lifecycle Practices certification program designed by the C-DAC Hyderabad and IIT Bhilai with the support of the Ministry of Electronics and Information Technology, Government of India.

References

- *Forbes. (2023, March 14). Cybersecurity Trends & Statistics For 2023; What You Need To Know. https://www.forbes.com/sites/chuckbrooks/2023/03/05/cybersecurity-trends--statistics-for-2023-more-treachery-and-risk-ahead-as-attack-surface-and-hacker-capabilities-grow/?sh=6f0e4e3019db.*
- *Josh Fruhlinger. (2022, Sep 6). The Heartbleed bug: How a flaw in OpenSSL caused a security crisis | CSO Online. https://www.csoonline.com/article/562859/the-heartbleed-bug-how-a-flaw-in-openssl-caused-a-security-crisis.html.*
- *Dara Khosrowshahi. (2017, Nov 21). 2016 Data Security Incident | Uber Newsroom. https://www.uber.com/newsroom/2016-data-incident.*
- *Vamsii777. (2023, July 3). Leaked Stripe API Key in Public Code Repository – Advisory – tktchurch/website - GitHub. https://github.com/tktchurch/website/security/advisories/GHSA-x3m6-5hmf-5x3w*
- *Synopsys Inc. (2020, June 03). Heartbleed Bug. https://heartbleed.com.*

- *BitcoinCore. (2021). Disclosure of remote crash due to addr message spam. https://bitcoincore.org/en/2024/07/31/disclose-addrman-int-overflow/*
- *Cisco Systems Inc. (2014, Mar 18). NVD-CVE-2014-2120. National Vulnerability Database. https://nvd.nist.gov/vuln/detail/CVE-2014-2120.*
- *Wordfence. (2024, Aug 21). NVD-CVE-2024-6386. National Vulnerability Database. https://nvd.nist.gov/vuln/detail/CVE-2024-6386.*
- *The Apache Software Foundation. (2024, Dec 18). Collections – Commons Collections Security Reports. Apache Commons. https://commons.apache.org/proper/commons-collections/security-reports.html.*
- *Benny Jacob. (2021, July 27). Confluence Server Webwork OGNL injection – CVE-2021-26084. Confluence Data Center. https://jira.atlassian.com/browse/CONFSERVER-67940.*
- *Vardan Torosyan. (2021, Dec 7). Grafana 8.3.1, 8.2.7, 8.1.8, and 8.0.7 released with high severity security fix. Grafana Labs. https://grafana.com/blog/2021/12/07/grafana-8.3.1-8.2.7-8.1.8-and-8.0.7-released-with-high-severity-security-fix.*
- *Paloalto. (2020, Dec 11). Authentication bypass vulnerability in GlobalProtect client certificate verification. https://security.paloaltonetworks.com/CVE-2020-2050.*
- *GitGuardian. (2024). The State of Secrets Sprawl 2024. https://www.gitguardian.com/files/the-state-of-secrets-sprawl-report-2024.*
- *Apache Software Foundation. (2021, Dec 10). CVE-2021-44228. National Vulnerability Database. https://nvd.nist.gov/vuln/detail/CVE-2021-44228 on 26-12-2024*
- *National Institute of Standards and Technology (NIST). (2022). Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218.pdf*
- *National Institute of Standards and Technology (NIST). (2024). Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218a.pdf*
- *Microsoft. (2024). Microsoft Security Development Lifecycle (SDL) Practices. https://www.microsoft.com/en-us/securityengineering/sdl/practices?oneroute=true on 26-12-2024*
- *OWASP Foundation. (2021). Application Security Verification Standard (ASVS) version 4.0.3. https://github.com/OWASP/ASVS/blob/v4.0.3/4.0/OWASP Application Security Verification Standard 4.0.3-en.pdf*
- *OWASP Foundation. (2020). OWASP Threat Dragon v2.3.0-latest. https://www.threatdragon.com/#/*
- *Lindunn Examples. (2024). https://downloads.linddun.org/examples/generic/v0/ examples.pdf*
- *Kim Wuyts, Wouter Jossen. (July 2015). Lindunn privacy threat modelling: a tutorial.https://lirias.kuleuven.be/retrieve/331950.*
- *MITRE. (2024). CWE – CWE List Version 4.16. Common Weakness Enumeration - A community-developed list of Software and Hardware Weaknesses that can become vulnerabilities. https://cwe.mitre.org/data/index.html.*
- *Carnegie Mellon University. (2024). SEI Cert Secure Coding Guidelines. https://wiki.sei.cmu.edu/confluence/display/seccode*
- *ISO/IEC. (2019). Programming Languages – Guidance to avoiding vulnerabilities in programming Languages – Part 1 Language-Independent Guidance, ISO/IEC TR 24772-1. International Electrotechnical Commission.*

STRENGTHENING CYBERSECURITY THROUGH AI

CAPACITY BUILDING, THREAT INTELLIGENCE, AND NATIONAL STRATEGIES

HIMANSHU KUMAR RAGHAV

ASSISTANT ACCOUNTS OFFICER, DEPARTMENT OF POSTS

ABSTRACT

The evolving cybersecurity landscape has introduced sophisticated threats that challenge traditional defence mechanisms. Artificial Intelligence (AI) offers transformative potential, serving as both a tool for attackers and defenders. This paper explores the integration of AI in threat intelligence and its role in enhancing national cybersecurity strategies. Capacity building is highlighted as a crucial enabler, bridging the skill gap required to deploy AI technologies effectively. By analysing global case studies and challenges, the paper outlines strategies for leveraging AI in proactive defence, fostering collaboration, and addressing ethical concerns. Policy recommendations focus on capacity building, explainable AI, and international cooperation.

Introduction

The Evolving Cybersecurity Landscape

Cyber threats have evolved significantly in complexity, scale, and sophistication. From ransomware to nation-state cyber-attacks, the global digital ecosystem is under constant threat. AI is now a double-edged sword, enabling attackers with automated tools while providing defenders with AI-driven solutions for rapid response.

The cybersecurity landscape is evolving at an unprecedented pace. As digital transformation accelerates globally, cyber threats have become increasingly sophisticated, dynamic, and pervasive. Traditional cybersecurity measures are struggling to keep up with emerging challenges like ransomware, phishing attacks, and AI-driven offensive tools. Nation-state actors and organized cybercriminal groups are leveraging advanced technologies, including artificial intelligence (AI), to conduct large-scale attacks that disrupt economies, compromise national security, and jeopardize critical infrastructure.

AI plays a dual role in cybersecurity—it empowers both attackers and defenders. Malicious actors use AI to automate reconnaissance, exploit vulnerabilities, and evade detection through adaptive malware. On the defensive side, AI enables organizations to analyse massive datasets, detect anomalies, and respond to threats in real time. However, the rapid proliferation of AI technologies also creates a new dimension of cybersecurity challenges, requiring nations to rethink their strategies and invest in innovative solutions.

FIGURE 1 : A TIMELINE SHOWING THE PROGRESSION FROM TRADITIONAL THREATS

The Critical Role of Human Capital

Despite advancements in technology, the cybersecurity workforce struggles to keep up. A global skills shortage exists, and AI tools alone are insufficient without trained professionals to deploy and manage them effectively.

While AI and other advanced technologies are vital tools in the fight against cyber threats, technology alone is not sufficient. Skilled personnel remain the backbone of effective cybersecurity operations. The shortage of skilled professionals in the cybersecurity industry has become a pressing concern. Recent studies indicate that there are millions of unfilled cybersecurity roles globally, creating a gap that leaves organizations exposed to potential cyber threats due to insufficient expertise within their security teams.

This growing shortage of cybersecurity professionals highlights the importance of building human capital through targeted capacity building and training programs. Nations must prioritize developing a workforce equipped with the necessary technical and analytical skills to leverage AI for cybersecurity. Addressing this skills gap is essential for enabling organizations to utilize AI effectively and stay resilient against evolving cyber threats.

Capacity Building and Training as a Cornerstone

Capacity building refers to strengthening the knowledge, skills, and processes necessary to address cybersecurity challenges effectively. It involves creating systems, frameworks, and human capital to enhance a nation's cybersecurity readiness. AI integration into cybersecurity operations requires specific expertise in data science, machine learning, and threat analysis, making training and education critical components of any national cybersecurity strategy.

By investing in capacity building, nations can:

- Equip professionals with skills to utilize AI-powered tools for threat intelligence and incident response.
- Develop resilient systems to protect critical infrastructure.
- Foster collaboration between governments, industry, and academia to tackle cyber threats.

I. Enhancing National Cybersecurity Through AI-Powered Threat Intelligence

This study examines the role of AI in enhancing threat intelligence and its implementation within national cybersecurity frameworks. It emphasizes the importance of capacity building to equip nations with the necessary expertise to effectively utilize AI for proactive threat detection and mitigation.

Threat intelligence is a proactive approach to identify, analyse, and respond to cyber threats before they materialize. By leveraging AI, organizations can process vast datasets, detect patterns, and predict attacks, enabling real-time threat mitigation.

National cybersecurity strategies must incorporate AI-driven tools to strengthen proactive defence, protect critical infrastructure, and respond efficiently to incidents. Capacity building, through targeted training and collaboration, is crucial to achieving this goal.

Thesis Statement:

"Effective capacity building and targeted training programs are essential for nations

to successfully integrate AI into their threat intelligence operations and national cybersecurity strategies, Enabling advanced threat prevention and strengthening defences against emerging cyber risks".

II. AI for Threat Intelligence: Opportunities and Challenges

Limitations of Traditional Threat Intelligence Approaches

For years, traditional threat intelligence methods—including signature-based detection, manual log analysis, and rule-based systems—have formed the foundation of cybersecurity defences. These methods primarily rely on predefined patterns and known threat signatures to identify malicious activities. While they are effective in detecting previously identified threats, they struggle to combat advanced and evolving cyberattacks that constantly adapt.

Challenges in Traditional Threat Intelligence

- Response Delays: Manual threat detection processes are time-intensive, often leading to delays in responding to attacks.
- Data Overload: The exponential growth of cybersecurity data makes it challenging for human analysts to process and analyzes threats effectively.
- Evasive Threats: Modern cyber threats, including polymorphic malware and zero-day exploits, frequently bypass static security measures, rendering traditional detection methods ineffective.

As cyberattacks become more sophisticated, organizations require advanced tools that can analyse vast datasets, detect anomalies, and predict threats in real time. This is where AI comes into play.

The Role of AI in Strengthening Threat Intelligence

AI transforms threat intelligence by automating data analysis, identifying patterns, and predicting potential attacks. The following key areas highlight how AI enhances threat intelligence:

1. Data Processing and Threat Identification

AI-driven solutions efficiently analyse vast datasets from multiple sources, including network activity logs, social media platforms, dark web monitoring, and cybersecurity intelligence feeds, to detect anomalies and emerging threats.

Example: Platforms like Splunk and IBM Watson utilize AI to process unstructured data, providing security teams with actionable insights.

2. Behavioural Anomaly Detection

Machine learning models assess network behaviour patterns to detect deviations that may indicate unauthorized access or suspicious activities.

Example: AI-powered security tools can recognize unusual login attempts, data transfers, or system modifications, allowing for proactive threat mitigation.

3. Intelligent Malware Identification

Instead of relying solely on signature-based detection, AI leverages behavioural analysis and code inspection to identify previously unknown malware strains.

Example: Deep learning models can recognize and neutralize polymorphic malware, which adapts to evade traditional security measures.

4. Threat Forecasting and Risk Prediction

AI utilizes historical attack data and evolving cyber threat trends to predict potential security breaches before they occur, allowing organizations to reinforce their defences accordingly.

EXAMPLE: PREDICTIVE AI TOOLS FORECAST VULNERABILITIES AND HELP ORGANIZATIONS

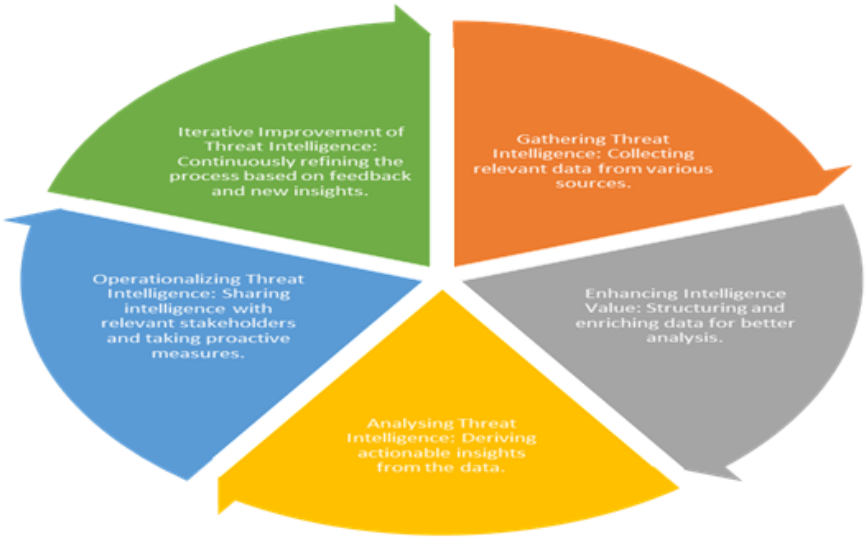


FIGURE 2: AI CAPABILITIES IN THREAT INTELLIGENCE

Challenges of Using AI in Threat Intelligence

Despite its advantages, AI adoption in threat intelligence presents several challenges:

1. **Data Quality and Availability**

AI systems require high-quality, labelled data for training. Inconsistent or incomplete data can reduce accuracy and reliability.

2. **Manipulation of AI Systems by Attackers**

Cybercriminals can alter input data to deceive AI models, causing misclassification of threats or bypassing the security measures.

Example: In data poisoning attacks, adversaries inject deceptive information into training datasets, compromising AI's ability to detect cyber threats accurately.

3. **Lack of Interpretability in AI Decisions**

Many AI models function as complex, opaque systems, making it challenging to understand their decision-making process. Ensuring explainability in AI-driven cybersecurity is essential for enhancing trust, reliability, and regulatory compliance.

4. **Compatibility with Legacy Cybersecurity Infrastructure**

Integrating AI-powered security tools with existing cybersecurity frameworks can be difficult due to compatibility issues, resource constraints, and the need for specialized expertise. Organizations must allocate technical and financial resources to ensure seamless AI adoption.

III. Implementing AI in National Cybersecurity Frameworks

AI's Impact on Enhancing National Cybersecurity

AI plays a crucial role in bolstering national cybersecurity by enabling advanced threat detection, safeguarding critical infrastructure, and optimizing incident response mechanisms.

1. **Pre-emptive Threat Mitigation**

By leveraging predictive analytics, AI can assess historical attack patterns and emerging cyber threats to anticipate potential vulnerabilities. This allows organizations to implement preventive security measures before an attack occurs.

2. Securing Critical Infrastructure

AI-driven security solutions provide continuous monitoring of vital systems, including energy grids, transportation networks, and communication systems. These tools swiftly detect and respond to potential threats, ensuring minimal disruption and enhanced resilience.

Example: AI algorithms monitor IoT devices in smart cities to identify unauthorized access or anomalous behaviour.

3. Incident Response

AI automates incident response processes, enabling faster detection, containment, and recovery from cyberattacks. Automated playbooks powered by AI reduce human error and improve response times.

4. Cybersecurity Information Sharing

AI facilitates secure and efficient sharing of threat intelligence between government agencies, private organizations, and international partners. By standardizing and analysing shared data, AI enhances situational awareness.

FIGURE 3: INCIDENT RESPONSE AUTOMATION PROCESS WITH AI

Policy Considerations for AI in National Cybersecurity

To effectively integrate AI into national cybersecurity strategies, policymakers must address several key considerations:

1. Safeguarding Data Privacy and Security

Ensuring AI can effectively analyse vast datasets while maintaining individual privacy protections is essential. Striking a balance between data collection for threat intelligence and adhering to privacy regulations remains a key challenge.

2. Legal and Regulatory Frameworks

Developing comprehensive laws and regulations for AI use in cybersecurity ensures accountability and prevents misuse. These frameworks should address ethical concerns and liability issues.

3. International Cooperation and Standards

Promoting global collaboration and harmonized standards for AI-driven cybersecurity fosters interoperability and collective defence against cross-border cyber threats.

IV. Developing Workforce Capabilities & Training for AI-Powered Cybersecurity

Recognizing Skill Deficiencies

Successfully integrating AI into cybersecurity requires a specialized skill

set that combines expertise in both technological advancements and regulatory frameworks. Current skill deficiencies present challenges in the following areas:

1. **Data Science and Machine Learning**: Cybersecurity professionals often lack expertise in developing and deploying AI models tailored for cybersecurity applications.
2. **Threat Intelligence Analysis:** Few professionals are trained to interpret AI-generated insights for proactive decision-making.
3. **Cybersecurity Policy and Governance:** Policymakers and regulators need a deeper understanding of AI's implications for governance, privacy, and compliance.

Hands-On Expertise with AI Tools: A shortage of practical experience in configuring and using AI-driven cybersecurity tools exists across industries.

Addressing these deficiencies is essential for building a workforce capable of leveraging AI effectively to strengthen national cybersecurity resilience.

Target Oriented Training Programs

To bridge the identified deficiencies, nations must develop tailored training programs and initiatives:

1.Applying Data Science and Machine Learning

- Specialized training programs covering AI algorithms, neural networks, and natural language processing (NLP) to equip cybersecurity professionals with advanced threat detection and analysis skills.
- Workshops on tools like TensorFlow, PyTorch, and Scikit-learn applied to threat detection and analysis.
- Example: Courses teaching anomaly detection techniques and malware classification using machine learning.

2.AI for Cybersecurity Applications

- Specialized training on AI applications, such as threat intelligence platforms, endpoint security, and AI-enhanced SIEM (Security Information and Event Management) systems.
- Practical labs simulating real-world cyberattacks, enabling participants to use AI-powered tools for mitigation.
- Example: Training on the use of tools like Darktrace and Splunk with embedded AI capabilities.

3.Cybersecurity Policy and Governance

- Training programs for government officials on ethical considerations, regulatory frameworks, and international standards for AI in cybersecurity.
- Example: A curriculum on balancing AI use with privacy and security regulations.

4.Hands-on Training and Simulations

- Building cyber ranges and simulated environments where trainees can experiment with AI-driven cybersecurity scenarios.
- Example: Gamified scenarios where participants deploy AI tools to defend against AI-powered adversarial attacks.

Collaboration between Academia, Industry, and Government

Collaboration among these key stakeholders ensures the development of effective training initiatives:

1. Academia: Universities should design interdisciplinary programs combining AI, data science, and cybersecurity.

2. Industry: Corporations can share expertise on the latest AI tools and provide internships or apprenticeships for students and professionals.

3. Government: Policymakers must fund capacity-building initiatives and ensure alignment with national security priorities.

Establishing Centres of Excellence

Centres of Excellence (CoEs) can serve as hubs for AI-driven cybersecurity research and training. These centres should:

1. Conduct advanced research on emerging threats and AI solutions.
2. Offer certification programs for professionals on emerging threats and AI solutions.
3. Offer knowledge exchange through workshops, conferences, and public-private partnerships.

Example: The U.S. National Initiative for Cybersecurity Education (NICE) is a model for global adaptation.

Promoting Awareness in Cybersecurity

Beyond specialized training, broader public awareness campaigns are essential to bring a culture which is fully aware in Cybersecurity matters. Governments and organizations can:

1. Launch Nationwide initiatives to bring awareness in citizens.
2. Create accessible content on AI-related risks and safety measures for the general public.

Leverage social media to reach diverse audiences with easy-to-understand resources.

V. Case Studies and Best Practices for AI in Cybersecurity

Case Study 1: Threat Detection at a National Scale powered by AI

Context

In response to increasing cyberattacks on critical infrastructure, a cybersecurity agency implemented an AI-powered threat detection system to enhance its monitoring and response capabilities.

Implementation

1. AI Tools Deployed: The agency integrated AI-driven Security Information and Event Management (SIEM) tools to detect issues in real time manner.
2. Collaboration: Partnerships were established with private organizations to improve data collection and threat intelligence sharing.
3. Outcome: The system identified and mitigated several major ransomware attacks before they could compromise critical infrastructure, reducing response times by 40%.

Key Takeaway:

Collaboration between public and private sectors can amplify the impact of AI in national cybersecurity strategies.

Case Study 2: AI in Predictive Analysis for Healthcare Cybersecurity

Context

A large healthcare organization faced targeted ransomware attacks aimed at patient data systems. To manage this, the organization adopted AI tools capabilities.

Implementation

1. Predictive Tools & Incident Simulations: Machine learning technologies were used to forecast vulnerabilities in medical IoT devices and hospital networks. This helped the organization to refine its incident response systems.
2. Outcome: The Proactive measures used by organisations by implementing AI technologies, prevented breaches, which helped the organization to save millions in

potential downtime and data recovery costs.

Key Takeaway:

Predictive analysis helps organizations stay ahead of threats, particularly in sectors where downtime can have life-threatening consequences.

Case Study 3: AI-Driven Cyber Threat Analysis in the Financial Sector

Context

A multinational bank used AI to detect fraudulent activities and secure its digital banking platforms against evolving threats.

Implementation

1. AI for Fraud Detection and Monitoring: AI technologies analyses the transaction patterns to identify fraudulent behaviour. Thus saving hard earned money of Public as well as that of organisation.
2. Outcome: The bank achieved a 30% reduction in fraud-related cases in the first year of implementation itself.

Key Takeaway:

Combining AI-driven tools with traditional security measures enhances protection against financial fraud.

Examples of Successful AI Integration

1. **Israel's National Cyber Directorate:** Use of AI for proactive threat intelligence.
2. **Singapore's Cyber Security Agency (CSA):** AI integration for infrastructure security.

Best Practices for AI Integration in Cybersecurity

1. Begin with Pilot Projects

Before large-scale deployment, test AI tools on smaller projects to assess their effectiveness and identify areas for improvement.

2. Invest in Explainable AI

Prioritize the development of AI models that provide clear explanations for their decisions to enhance trust and accountability.

3. Focus on Interoperability

Ensure that AI tools are compatible with existing cybersecurity systems and can integrate seamlessly into an organization's infrastructure.

4. Foster Collaboration

Build partnerships between governments, academia, and the private sector to facilitate knowledge sharing and resource pooling.

5. Regularly Update Training Programs

Continuously update capacity-building initiatives to keep pace with evolving AI technologies and cyber threats.

6. Global Examples of AI in Cybersecurity

a. United States: Cyber Threat Intelligence Integration Centre (CTIIC)

The U.S. CTIIC AI technologies to collect, analyse, and protect from cyber threat across government agencies.

b. European Union: Cybersecurity Act

The EU's cybersecurity Act emphasizes the use of AI tools for risk management and incident response, supported by strict regulatory guidelines.

c. Singapore: AI-Singapore Initiative

Singapore leverages AI to protect its smart city infrastructure, employing real-time monitoring and anomaly detection systems.

VI. Conclusion and Policy Recommendations

Conclusion

The integration of Artificial Intelligence (AI) into cybersecurity offers a transformative potential to enhance our defence against the ever-evolving cyber threat landscape. By harnessing AI's capabilities for threat intelligence and national cybersecurity strategies, governments and organizations

can proactively identify, analyse, and mitigate risks. However, the successful adoption of AI-driven cybersecurity solutions necessitates addressing technical, ethical, and policy challenges.

Capacity building is pivotal in bridging the skills gap and fostering a resilient cybersecurity workforce. By investing in targeted training programs, nations can empower individuals and institutions to effectively deploy AI technologies, fostering collaboration across sectors and driving innovation in cybersecurity practices.

To realize the full potential of AI in cybersecurity, a balanced approach is crucial. This approach should combine advanced technological solutions with robust human expertise and a strong policy foundation. By maintaining this equilibrium, AI can be deployed responsibly and efficiently to secure digital infrastructure and protect critical information assets from emerging cyber threats.

Policy Recommendations

To ensure effective integration of AI, the following policy recommendations are proposed:

1. Develop National AI-Cybersecurity Frameworks

a. Establish comprehensive frameworks that define the roles of AI in cybersecurity operations.

b. Ensure alignment with international standards for interoperability and collaboration.

c. Establishing Ethical AI Standards: Develop well-defined principles to guide the design, implementation, and usage of AI-driven cybersecurity solutions, ensuring responsible and transparent deployment.

2. Strengthen Data Governance Policies

a. Enforce robust data protection regulations to safeguard AI datasets while ensuring compliance with privacy standards.

b. Encourage transparent and ethical data-sharing practices to improve collaboration in cybersecurity efforts.

c. Define Data Quality Standards: Maintain accuracy, consistency, and reliability in datasets used for AI model training and operation.

3. Invest in Capacity Building Efforts

a. Implement nationwide training programs to cultivate a highly skilled cybersecurity workforce equipped with AI expertise.

b. Strengthen collaborations between academic institutions, industry leaders, and government agencies to develop dynamic and industry-relevant learning opportunities.

c. Support Continuous Learning: Encourage ongoing professional development to keep pace with evolving AI technologies.

4. Promote Ethical AI Practices

a. Encourage the development of explainable AI systems to improve transparency and trust.

b. Address the risks of adversarial AI and ensure accountability through ethical guidelines.

c. Establish AI Ethics Boards: Create dedicated bodies to oversee the ethical development and deployment of AI in cybersecurity.

5. Encourage International Collaboration

a. Participate in global forums to share threat intelligence and best practices.

b. Advocate for harmonized regulatory standards to tackle cross-border cyber threats.

c. Foster Public-Private Partnerships: Encourage collaboration between government, industry, and academia to address shared challenges.

6. Incentivize Innovation

a. Provide funding and resources for research and development in AI-powered cybersecurity tools.

b. Support start-ups and small businesses developing cutting-edge solutions for cyber defence.

c. Create Regulatory Sandboxes: Establish safe spaces for testing and experimenting with innovative AI-based cybersecurity solutions.

Appendix

A. Definitions of Key Terms

1. Artificial Intelligence (AI):

The simulation of human intelligence processes by machines, especially computer systems, including learning, reasoning, and self-correction.

Example: AI-based algorithms used in anomaly detection for cybersecurity.

2. Threat Intelligence (TI):

The collection and analysis of information about current and potential cyber threats to prevent or mitigate attacks.

Example: Threat feeds providing data on malware signatures and emerging vulnerabilities.

3. Capacity Building:

The process of developing and strengthening skills, resources, and abilities to enable organizations or nations to achieve their goals.

Example: Training cybersecurity professionals in AI-driven security tools.

4. Explainable AI (XAI):

AI systems designed to explain their decision-making processes in human-understandable terms, fostering trust and accountability.

B. Examples of AI Tools in Cybersecurity

1. Darktrace:

Utilizes AI for detecting and responding to cyber threats in real time.

- Application: Identifies anomalous behaviours in corporate networks.

2. Splunk:

Provides AI-driven analytics to monitor network activity and predict potential breaches.

3. IBM Watson:

Processes unstructured data for cybersecurity insights.

- Application: Detects phishing attempts and flags them in emails.

4. TensorFlow & PyTorch:

Machine learning frameworks for training models used in threat analysis and malware detection.

C. Sample Framework for National AI-Cybersecurity Strategy

1. Components of a National AI Strategy in Cybersecurity:

- Threat Intelligence and Sharing: Integration of AI-powered tools to collect, analyse, and disseminate threat data.
- Capacity Building: Initiatives for training personnel in AI and cybersecurity skills.
- Data Privacy Regulations: Policies to ensure ethical data use in AI systems.

2. Key Stakeholders:

- Government cybersecurity agencies.
- Academic institutions and research centres.
- Industry leaders in AI and cybersecurity solutions.

D. Figures and Illustrations

1. Sample AI Workflow in Threat Intelligence:

Data Collection → Preprocessing → Anomaly Detection → Threat Prediction → Response Automation.

2. AI Integration Framework:

Diagram depicting how AI tools integrate with traditional SIEM systems to enhance incident response capabilities.

E. Metrics for Evaluating AI Integration in Cybersecurity

1. **Performance Metrics:**
   - Detection accuracy rate.
   - Reduction in incident response time.
   - Number of false positives/negatives.
2. **Operational Metrics:**
   - Cost savings achieved by automating manual tasks.
   - Adoption rate of AI tools among teams.
3. **Impact Metrics:**
   - Percentage decrease in successful cyberattacks.
   - Reduction in system downtime caused by breaches.

F. Additional Resources for Training Programs

1. Courses and Certifications:

- Certified Ethical Hacker (CEH) with AI modules.
- Certified Information Systems Security Professional (CISSP).
- AI in Cybersecurity offered by platforms like Coursera or Udemy.

2. Simulation Platforms:

- Cyber ranges for AI-based attack/defence simulations.
- Gamified tools for hands-on training in threat detection.

3. Publications and Journals:

- Journal of Cybersecurity Technology.
- IEEE Transactions on Information Forensics and Security.

G. Examples of Collaborative Initiatives

1. U.S. National Initiative for Cybersecurity Education (NICE):

Focuses on bridging the skills gap in AI and cybersecurity through public-private partnerships.

2. EU AI Act for Cybersecurity:

Establishes standards for ethical AI use and promotes cross-border threat intelligence sharing.

Acknowledgement

I would like to express my heartfelt gratitude to all those who supported me in the completion of this research paper.

I am profoundly grateful to my life companion, whose encouragement and belief in my abilities inspire me to strive for continuous self-improvement beyond the routine of daily work. Her belief in my abilities has been a driving force throughout this journey.

Further, I extend my deepest thanks to my friend, P. Anant Bhaskar, Database Analyst at Atlantic Packaging Products Ltd., Ontario, Canada, for his invaluable encouragement and insights. His professional expertise and unwavering support have been instrumental in shaping the ideas and perspectives presented in this paper.

Finally, I wish to acknowledge the contributions of the Department of Posts & the global academic and professional community working tirelessly to advance the fields of artificial intelligence and cybersecurity.

Thank you all for your guidance, support, and inspiration.

References

- *Canadian Centre for Cyber Security. (2025). National Cyber Threat Assessment 2025-2026. Retrieved from cyber.gc.ca (https://www.cyber.gc.ca/en/guidance/national-cyber-threat-assessment-2025-2026)*
- *Langeh, A., & Sudhakar, R. (2024). Artificial intelligence and cyber security: Transformative synergies in the digital frontier. International Journal of Cybersecurity.*
- *(https://theacademic.in/wp-content/uploads/2024/07/35.pdf)*
- *Gabrian, C.A. (2024). Unveiling the Dark Side: How Hackers Exploit Artificial Intelligence for Cyber Warfare. Resilience Journal.*
- *(https://resiliencejournal.e-arc.ro/wp-content/uploads/2024/06/EARJ-3-2024-Gabrian.pdf)*
- *Ali, A. (2024). Navigating the Cyber Threat Landscape: Effective Vulnerability Assessment and Defence Strategies. ResearchGate.*
- *(https://www.researchgate.net/publication/384367262_Navigating_the_Cyber_Threat_Landscape_Effective_Vulnerability_Assessment_and_Defense_Strategies)*

Bibliography

- *Ali, A. (2024). Navigating the Cyber Threat Landscape. ResearchGate.*
- *Clarke, R.A., & Knake, R.K. (2019). The Fifth Domain. Taylor & Francis.*
- *Canadian Centre for Cyber Security. (2025). National Cyber Threat Assessment.*
- *Gabrian, C.A. (2024). Unveiling the Dark Side. Resilience Journal.*
- *Langeh, A., & Sudhakar, R. (2024). Artificial Intelligence and Cyber Security.*

POLICY CONSIDERATIONS IN THE USE OF ARTIFICIAL INTELLIGENCE (AI) IN INDIA: A CASE FOR THE FINANCIAL SECTOR

ROHIT CHAWLA, IES[1] AND ABHINAV BANKA, IES[2]

[1]DEPUTY DIRECTOR, DEPARTMENT OF ECONOMIC AFFAIRS, MINISTRY OF FINANCE
[2]ASSISTANT DIRECTOR, DEPARTMENT OF ECONOMIC AFFAIRS, MINISTRY OF FINANCE

ABSTRACT

The use of Artificial Intelligence (AI) is multiplying rapidly in the financial sector across the globe. Financial services are well poised to take advantage of recent advances in AI given the industry's long-standing focus on improving regulation and supervision using data-backed mechanisms. However, parallelly there are new tail risks emanating from the use of AI, which merits attention. Further complexities could arise with Generative AI given its linkages to numerous black boxes. This paper assesses the policy considerations for the use of AI in the Indian financial sector, especially from the financial stability perspective. Using cross-country analysis, the paper tries to bring forward major issues faced by regulators/policymakers in developing the AI regulation framework. Building upon key attributes of accountability, explainability, governance and usability, the paper highlights important policy considerations for the responsible use of AI in the financial sector. The agenda paper is structured as follows: Section 1outlines use cases and benefits of AI in the financial sector, Section 2 lists out emerging risks from AI in finance, Section 3 brings cross country approaches to AI regulations, Section 4 deals with India's AI preparedness, Section5 outlines policy considerations and important questions for discussions.

Keywords: Artificial Intelligence (AI), Emerging risks, Financial Stability, Governance, Regulations

1.0. Artificial Intelligence Adoption in The Indian Financial Sector

National Association of Software and Service Companies ("NASSCOM") surveyed 500 companies in India in 2021 and 2022 to study the rate of adoption of AI across various sectors in its AI Adoption Index Reports (NASSCCOM, 2022). As per this report, the AI adoption in India has increased by 14 percent since 2021, where 70 percent of the enterprises are allocating more than 20 percent of their Information Technology budget to it. While there are specific AI applications incorporated in different sectors, there are general use cases that are deployed across various sectors as chat bots, optical character recognition, text mining, and sentiment analysis.

Different AI models can be applied to different use cases, depending on the task at hand and the limitations of any given model. While developments under AI are multiplying at an advanced rate, the major innovation has been led by Generative artificial intelligence (GenAI). The areas under the financial sector where AI technology can be constructively used include REGTECH & SUPTECH, Portfolio Management, Capital Optimization, Climate Risk Assessment using assessment of

climate-related disclosures, etc. (Aldasoro et al., 2024), (IOSCO, 2021)

GenAI has enabled a range of new operations-focused use cases. These include improving information search and retrieval, content generation (e.g. automated text, image, and video generation), voice transcriptions (e.g. voice-to-text and text-to-summary service requests), and code generation or legacy code streamlining which could help smaller businesses that lack sufficient skilled manpower. Businesses are also exploring AI to manage riskier use cases, such as optimising their regulatory capital requirements using ML techniques in areas such as default probability modelling(FSB, 2017).

The Banking, Financial Services and Insurance (BFSI) sector is rapidly increasing its deployment of AI and ML in its operations. The sector has deployed AI/ML technology in four domains: Sales and Marketing, Customer Services, Core Transactions, Risk and Compliance. Under Sales and Marketing, AI technology has been incorporated into product-centric operations like product personalisation, targeted outreach, and product optimisation. BFSI has witnessed proactive adoption of AI technology for customer services in the form of chatbots, virtual assistance, and customer identification.

A study conducted by RBI staff (Goel et al., 2024) by analysing the annual reports of Indian banks, revealed that the usage of AI related keywords has increased sizeably, especially in Private Sector Banks (PVBs). A survey shows that, 11 out of the total 12 Public Sector Banks (PSBs) and 15 out of total 21 PVBs have deployed AI and Machine Learning based technology through virtual assistants and Chatbots for better customer interaction.

Within banking sector particularly, AI is leveraged for wider purpose of financial inclusion. Mobile banking apps, AI driven credit scoring, and blockchain-based solutions are examples of some of the powerful and enabling vehicles of greater financial inclusion by extending access to banking services to previously unbanked or underbanked individuals (Bazarbash, 2019 ,Gensler & Bailey, 2020). Alternative methods of assessment using ML algorithms can help disadvantaged sections of the society with credit scoring which can bring them under the umbrella of traditional financial products (Kshetri, 2021)

Under core transaction domain, insurance operations which are infused with AI technology include credit underwriting, claims management, credit scoring, etc(IOSCO, 2021). while for risk compliance, major use cases include fraud detection and credit risk assessment. Insurance companies have deployed AI-driven face scan for remote health assessment(Rashid & Kausik, 2024).

AI tools are also being deployed across a greater number and wider variety of fraud and AML/CFT use cases. GenAI can potentially automate the generation of financial crime reports, incorporating a range of different sources and supporting investigators. Banks and FIs are increasingly adopting AI-powered cybersecurity tools for automated anomaly detection and predictive behavioural analysis to identify potential threats and vulnerabilities (Donepudi, 2017)

2.0. Emerging Financial Stability Risks from The Adoption of AI and The Need for Regulations

It is challenging to establish a comprehensive assessment of the implications of AI for the financial system as the technology keeps evolving. Moreover, the risks and benefits largely depend on the use cases of AI(FSB, 2017). Additionally, they rely on how the model is being trained and what kind of data is being fed into the foundational model. For example, the risk of 'AI hallucination1' is unavoidable if the foundational model is trained on biased or erroneous data. In such a case, the predictions of the model will sustain biases and errors. Nevertheless, there are a few risks that can be more damaging than others. Those are outlined below:

(i) Cyber security risk: Computer programs that unfairly and consistently discriminate against some people or groups of people in favour of others are known as embedded bias, according to (Friedman and NISSENBAUM, 1996) The use of AI introduces new, distinct cyberthreats and expands the potential for cyberattacks. AI systems can face new threats in addition to the usual cyberthreats from software or human error. These attacks focus on altering data within the AI/ML lifecycle to exploit the inherent limitations of AI systems (Comiter, 2019)).

For example, Data poisoning attacks intend to influence an ML algorithm during the training stage by adding special samples to its training data set while Input attacks allow attackers to introduce perturbations to data inputs and mislead AI systems during operations (El Bachir Boukherouaa, 2021).

(ii) Concentration risk: arises from the dependency on big-tech companies, which are at the forefront of the development of the LLM models. The concentration of third-party AI/ML algorithm providers could drive greater homogeneity in risk assessments and credit decisions in the financial sector.

(iii) Correlation risk: AI may heighten financial fragility emerging from herd behaviour of financial sector intermediaries which adopts similar third party application which are in turn trained on similar datasets. This could encourage monocultures. Likelihood of increased procyclicality in the financial sector due to uniformity as well as potential risk due to new systemically important third-party providers and infrastructure could exacerbate the inherent network interconnectedness of the domestic financial system.

As many such examples show, technology is both a source of efficiency and risk. A decision to run or stay is influenced by four factors: the endogenous response of market participants, strategic complementarities, objectives and oligopolistic market structure (Danielsson, 2024). When societal risks arising out of AI interact with financial stability concerns discussed above, crises may propagate through multiple channels including:

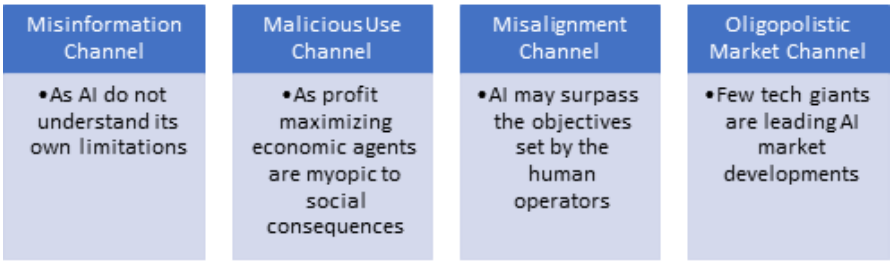


FIGURE 1: CHANNELS OF AILED CRISIS

SOURCE: (DANIELSSON, 2024)

It is the speed and accuracy of developing AI systems that could make these channels irreversible, making AI a crisis amplifier. Although the usage of AI in the Indian context is at a nascent stage, it is crucial to take a proactive approach in dealing with AI, to ensure that its uptake and deployment does not entail threats to financial stability.

The interplay between macro and micro-level challenges significantly influences the utility and implementation of AI. On the micro prudential front, the expanded application of AI in regulatory compliance and internal risk management is likely to yield broad social benefits. However, the dynamics shift when addressing macro prudential regulation, which focuses on maintaining financial system stability and averting systemic disruptions. Macro prudential concerns often involve rare, unpredictable events—commonly referred to as "unknown unknowns"—that introduce the risk of procyclicality. Furthermore, the substantial energy consumption by AI firms and data centres for model training and inference raises critical considerations for climate change and its implications for climate-related financial stability. Even beyond these rare "black-swan" scenarios, the above risks collectively pose potential threats to financial stability.

3.0. Cross Country Experiences in AI Adoption

In the wake of the extremely rapid development and deployment of artificial intelligence and machine learning technologies, many jurisdictions have seen the emergence of regulatory responses. While some of them have indeed taken a holistic approach towards addressing all issues involved for example, to strike a balance between safeguarding public interest in AI ethics and governance while simultaneously allowing innovation in AI to prosper (MindForge), the Singapore government has developed various frameworks and tools to guide AI deployment and promote the responsible use of AI, including the National Artificial Intelligence Strategy 2.0 (first launched in 2019, updated in 2023) (Singapore Ministry of Finance, 2023), which outlines Singapore's ambition and commitment to building a trusted and responsible AI ecosystem, bringing growth and innovation while handholding businesses and people to understand AI developments and deploy it. The Monetary Authority of Singapore (MAS) on June 26, 2023, released an open-source toolkit to enable the responsible use of Artificial Intelligence (AI) in the financial industry. Further, the MAS on December 05, 2024 released a paper on AI model risk management based on its thematic review of selected banks. It is part of MAS' incremental efforts to ensure responsible use of AI in the financial sector.

The US on the other hand follows a highly distributive approach. At the highest level, the US (the White House) has issued an Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence on October 30, 2023 (Executive Order, 2023). This EO builds on previous guidance documents such as the AI "Bill of Rights" and the National Institute for Standards in Technology (NIST) AI Risk Management Framework. The Securities and Exchange Commission in July 2023 proposed new rules that would require broker-dealers and investment advisers to take certain steps to address conflicts of interest associated with their use of predictive data analytics and similar technologies to interact with investors to prevent firms from placing their interests ahead of investors' interests.

The EU approach to AI risk management, tailored to specific digital environments is characterized as more comprehensive range of legislation. Besides the General Data Protection Regulation (GDPR), the Digital Services Act and Digital Markets Act, EU has recently come up with the Artificial Intelligence (AI) Act in March 2024. The EU AI Act (European Parliament, 2019) is the world's first AI law which aims to boost innovation in AI and at the same time protect fundamental rights, democracy, the rule of law and environmental sustainability from high-risk AI.

In the UK, regulation of AI is handled by multiple authorities, and the regulatory approach is characterized by guidelines, best practices and principles. The Department for Science, Innovation and Technology (DSIT) houses an Office for Artificial Intelligence. The DSIT came up with a white paper in March 2023 (updated in August 2023) titled "A pro-innovative approach to AI regulation" (DST, 2023)which details the UK's plans for implementing a pro-innovation approach to AI regulation. Further, the UK Government released a consultation response in Feb 2024 welcoming strong public support for these principles. In July,2024 UK came out with the Terms of Reference of AI Opportunities Action Plan (Action Plan), which sets out a "roadmap for government to capture the opportunities of AI to enhance growth and productivity and create tangible benefits for UK citizens. The Action Plan considers how the UK can: (i) build an AI sector that can "scale and be competitive globally"; (ii) adopt AI to "enhance growth and productivity and support [its] delivery across government"; (iii) use AI to "transform citizens' experiences" with the Government encouraging AI application across the public sector and wider economy; and (iv) enable the use and adoption of AI by supporting "data infrastructure, public procurement processes and policy and regulatory reforms. In the financial sector, the Financial Conduct Authority (FCA) has published a document titled "AI Update" which outlines the existing approach of FCA towards AI as well as the plan for next 12 months. FCA along with Bank of England has published AI Discussion Paper (Bank of England, 2022)which explores the potential benefits and risks of the use of AI in financial services about consumer protection, competition, and market integrity. The AI discussion paper also ponders on how existing regulatory requirements are applicable to use cases of AI in financial services and invites responses on how FCA, as a sectoral regulator, could promote the beneficial and safe adoption of AI in financial services.

China has implemented interim measures to address AI-related risks and impose compliance obligations on entities engaged in AI-related business. In their latest version, provisions only apply to public usages of gen-AI by both domestic and foreign individuals and entities involved in AI services in China. A version of the interim measures was released for comments in April 2023, followed by a substantially revised version dated July 2023 and taking effect on August 2023. In particular, the scope of regulation was narrowed from all uses of gen-AI to only public usages, with other cases being less strictly regulated. On September 9 2024, China's National Technical Committee 260 on Cybersecurity released the first version of its AI Safety Governance Framework(National Technical Committee 260, 2024) which outlines principles for AI safety governance, classifies anticipated risks related to AI, identifies technological measures to mitigate those risks, and provides governance measures and safety guidelines. China in December 2024 also has established a technical committee to advance the development of the AI industry and empower new industrialization for Artificial Intelligence (AI) standardization.

Be it through new regulations or existing ones, regulators have been focusing on AI/ML governance frameworks, risk management, internal controls, and better controls over model and data within the overall spectrum of AI/ML(El Bachir Boukherouaa, 2021).

4.0. India's Efforts on Managing Use of AI

The prominent index for assessing AI preparedness is the Government AI Readiness Index prepared and published by Oxford Insights. The Index has 39 indicators across 3 pillars - Government, Technology Sector and Data & Infrastructure. India is ranked 46 on this index in 2024, demonstrating competitive scores in both the Government pillar and the Data and Infrastructure pillar but falling behind in the Technology pillar. The other index is prepared by the IMF called the AI Preparedness Index (AIPI) which has four broad determinants - digital infrastructure, human capital, technological innovation, and legal frameworks. The IMF on June 25, 2024 released the AIPI Dashboard on their website, tracking 174 economies globally for AI readiness. India's score on the Index is 0.49.

Under the IndiaAI mission, the Ministry of Electronics and Information Technology (MeitY) has convened seven expert groups to collaboratively define the vision, objectives, outcomes, and frameworks for each of IndiaAI's 7 foundational pillars. In their First edition, India AI 2023 (MEITY, 2023) the working group outlined an institutional framework for governing data collection, management, processing, and storage under the National Data Management Office (NDMO).

Among these pillars, Safety and Trusted AI is a key focus, aimed at ensuring the responsible development, deployment, and adoption of AI. This pillar is being advanced through initiatives such as implementing Responsible AI projects, developing indigenous tools and frameworks, creating self-assessment checklists for innovators, and establishing robust guidelines and governance

mechanisms. As part of this effort, MeitY is in the process of establishing the India AI Safety Institute (AISI). The institute will serve a dual purpose: fostering AI innovation while ensuring the safe and ethical deployment of AI technologies. AISI's success as a global leader in AI safety will be supported through international partnerships and collaborations with domestic organizations.

Additionally, the Digital Personal Data Protection Act (DPDPA), 2023 establishes a comprehensive framework for data protection, with a strong emphasis on a consent-driven approach to personal data processing. It serves as a crucial element in achieving Safe and Trusted AI, requiring developers and deployers of AI models to thoroughly evaluate and adhere to the DPDPA's regulatory requirements. Compliance ensures the responsible and ethical handling of personal data, aligning with the broader goals of fostering trust and safety in AI applications.

Under financial sector, regulators have taken action now and then to ensure safe and secured use of AI. RBI Master Direction on Outsourcing of Information Technology Services dated April 10,2023 requires regulated entities (REs) to effectively assess the impact of concentration risk posed by multiple outsourcings to the same service provider and/or the concentration risk posed by outsourcing critical or material functions to a limited number of service providers. Reserve Bank in its Statement on Developmental and Regulatory Policies dated December 06, 2024, has announced constituting a committee to develop a Framework for Responsible and Ethical Enablement of AI (FREE-AI) in the Financial Sector. The Committee will comprise of experts from diverse fields and shall recommend a robust, comprehensive, and adaptable AI framework for the financial sector.

In the context of the reference to the securities market, given the increased usage of Artificial Intelligence (AI) and Machine Learning ("ML") tools in - facing products, SEBI vide its Circulars (SEBI, 2019)mandated reporting of AI and ML applications and systems offered and used by -

- Stock Brokers and Depository Participants (vide Circular dated January 4, 2019),
- Market Infrastructure Institutions (MIIs) i.e. Stock Exchanges, Depositories and Clearing Corporation (vide Circular dated January, 31, 2019), and
- Mutual Funds, Asset Management Companies, Trustee Companies, Board of Trustees of Mutual Funds (vide Circular dated May 9, 2019).

On November 13, 2024, SEBI released a Consultation Paper on the "Proposed Amendments Regarding Responsibility for the Use of Artificial Intelligence Tools by Market Infrastructure Institutions, Registered Intermediaries, and Other SEBI-Regulated Entities." The paper invites public comments on the matter. SEBI has proposed that all entities it regulates, regardless of the extent of their AI usage, must take full responsibility for the privacy, security and integrity of investors' and stakeholders' data including data maintained by it in a fiduciary capacity throughout the processes involved and the outcomes of their AI applications. This initiative aims to strengthen accountability as AI and machine learning become increasingly integrated into the financial sector.

5.0. Policy Considerations

AI's rapid transformation makes regulation-making difficult. There are trinity of challenges to the regulation of AI about speed, scope and approach, probing countries to adopt flexible approach to regulation (Japan and the UK), while some countries have sought to adopt rigid regulations to address risks, which manifest rapidly (EU).In the absence of clear and unambiguous effects of AI on economic variables like productivity, employment, growth etc., regulators around the world have been mostly more focused on ethical and bias concerns, privacy, and safety issues, which may fall outside the scope of economic policy (Comunale & Manera, 2024)

Danielsson (2024) mentions six-point criteria based on which any use of AI in the financial sector should be assessed including (i) Data- Does an AI engine have enough data for learning, or are other factors beyond available training dataset which materially impacting AI advice and decisions? (ii) Mutability. Is there a fixed set of immutable rules the AI must obey, or does the regulator update the rules in response to events? (iii) Objectives. Can AI be given clear objectives and its actions monitored in light of those objectives, or are they unclear? (iv)Authority. Would a human functionary have the authority to make decisions, does it require committee approval, or is a fully distributed decision-making process brought to bear on a problem? (v) Responsibility. Does private AI mean it is more difficult for the authorities to monitor misbehaviour and attribute responsibility in cases of abuse? In particular, can responsibility for damages be clearly assigned to humans? (vi)Consequences- Are the consequences of mistakes small, large but manageable, or catastrophic?

Countries also face difficult trade-offs between safety and attracting AI investments and innovation when choosing to regulate. In light of the above, below are some of the important overarching policy considerations:

- **Disclosure Requirements for Ensuring Transparency and Explainability:** Addressing the challenge of explainability requires careful consideration of the extent to which firms should disclose their use of AI and ML. Regulators need to determine which firms must inform their customers and clients about the use of AI and ML in ways that affect client outcomes. Additionally, regulators should assess the specific types of information they may need from these firms to maintain effective oversight and ensure compliance with appropriate standards.
- **Governance Frameworks for Accountability:** To address the issue of accountability, it is essential to establish explicit governance frameworks, such as model governance committees or model review boards, to clearly define responsibilities for AI-based systems. Regulators should consider mandating firms to designate senior management personnel tasked with overseeing the entire lifecycle of AI and ML systems, including a pipeline approach of models', development, testing, deployment, monitoring, and control mechanisms. This oversight should be supported by a documented internal governance framework that delineates clear lines of accountability.

- **Frameworks for Setting Robust Models in Place:** The use of AI brings concerns about data security and misuse. Training datasets need to be large and diverse enough, even if synthetic, to capture complex patterns and rare events, ensuring the models perform reliably during unexpected crises. Before deployment, testing should occur in a separate environment to confirm that AI and ML systems: (a) work as intended in both stable and stressed market conditions; and (b) meet regulatory requirements. Introducing automatic controls, like kill switches, can help reduce risks by triggering alerts or shutting down models during instability.
- **Fairness:** AI systems must respect legal rights, avoid unfair discrimination, and not lead to inequitable market outcomes. However, copyright rules for AI-generated or co-created content remain ambiguous in many countries. Stakeholders throughout the AI lifecycle should define fairness in a way that aligns with the system's purpose, outcomes, and applicable legal standards. Regulators should require firms to implement robust controls, ensuring high-quality, unbiased, and comprehensive data is used in AI models to support fair and reliable applications of AI and ML.
- **Redressal Mechanism for Grievances:** Authorities should consider implementing processes that allow customers to challenge AI model outcomes and seek compensation, helping to build trust in these systems. To ensure accountability, firms should have clear service agreements in place with outsourced providers, outlining the scope of services and the provider's responsibilities. These agreements should include specific performance indicators and clearly define rights and remedies in case of poor performance.

6.0. Way Forward

The adoption of AI holds significant potential to enhance the efficiency of financial intermediation and drive productivity gains through faster information processing. However, it also introduces risks, such as bias, hallucinations, and misuse. In India, the use of AI in the financial sector remains relatively limited, but rapid advancements in this field warrant closer scrutiny. A major threat to financial stability could stem from reliance on a few algorithms, which might amplify procyclicality and market volatility. Conversely, AI can also strengthen financial stability by improving the detection of early warning signals.

Regulating the space of AI requires a well-thought-out approach, including determining whether regulation is necessary, and if so, in what form. The objective is to address risks effectively while fostering innovation in this dynamic space. Given AI's global nature, a coordinated international framework is essential to harness its vast opportunities while mitigating societal risks. This calls for sound policies that strike a balance between innovation and regulation, ensuring AI serves the public good.

To achieve this vision, we may need global consensus for setting standards on regulation of AI. Several public initiatives, such as the OECD's "AI Principles" (OECD, 2021) and the G7's "Hiroshima Process International Code of Conduct for Advanced AI Systems," as well as private efforts like the Frontier Model Forum founded by Anthropic, Google, Microsoft, and OpenAI, provide a foundation for collaboration.

While technology regulation per-se over the decades has been a matter for global coordination, the domestic requirements for technology adoption like AI may be driven by factors beyond the economic scope. The decision of any jurisdiction with wider use of black box technology like AI should consider its transparency, accountability and ethicality aspect for long term over and above its short-term benefits.

References

- *Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). Intelligent financial system: how AI is transforming finance. www.bis.org*
- *Bank of England. (2022). DP5/22-Artificial Intelligence and Machine Learning.*
- *Bazarbash, M. (2019). FinTech in Financial Inclusion Machine Learning Applications in Assessing Credit Risk IMF Working Paper Monetary and Capital Markets Department FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk.*
- *Comiter, M. (2019). Attacking Artificial Intelligence AI's Security Vulnerability and What Policymakers Can Do About It. www.belfercenter.org*
- *Comunale, M., & Manera, A. (2024). The Economic Impacts and the Regulation of AI: A Review of the Academic Literature and Policy Actions, WP/24/65, March 2024.*
- *Danielsson, U. (2024). On the use of artificial intelligence in financial regulations and the impact on financial stability.*
- *Donepudi, P. K. (2017). Machine Learning and Artificial Intelligence in Banking. Engineering International, 5(2).*
- *DST, U. K. G. (2023). A pro-innovation approach to AI regulation. [Dandy Booksellers Ltd].*
- *El Bachir Boukherouaa, K. A. , J. D. , A. F. , and R. R. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance.*
- *European Parliament. (2019). Artificial Intelligence Act European -Parliament legislative resolution. http://data.europa.eu/eli/C/2024/506/oj*
- *Executive Order, U. G. (2023). Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Executive Order 14110.*
- *Friedman and NISSENBAUM. (1996). Bias in Computer Systems.*
- *FSB. (2017). Artificial intelligence and machine learning in financial services Market developments and financial stability implications. www.fsb.org/emailalert*
- *Gensler, G., & Bailey, L. (2020). Deep Learning and Financial Stability. https://ssrn.com/abstract=3766555*
- *Goel, S., Raut, D. K., Kumar, M., & Sharma^, M. (2024). How Indian Banks are Adopting Artificial Intelligence? https://www.businessinsider.in/finance/news/the-impact-of-artificial-*
- *IOSCO. (2021). The use of artificial intelligence and machine learning by market intermediaries and asset managers Final Report. www.iosco.org*
- *Kshetri, N. (2021). The Role of Artificial Intelligence in Promoting Financial Inclusion in Developing Countries. In Journal of Global Information Technology Management (Vol. 24, Issue 1, pp. 1–6). Bellwether Publishing, Ltd. https://doi.org/10.1080/1097198X.2021.1871273*
- *MEITY. (2023). IndiaAI-2023, Expert Group Report, First-Edition.*
- *MindForge. (n.d.). Emerging Risks and Opportunities of Generative AI for Banks A Singapore Perspective.*

- *NASSCCOM. (2022). NASSCOM AI ADOPTION INDEX -Tracking India's Sectoral Progress on AI Adoption.*
- *National Technical Committee 260, C. (2024). 1725849192841090989.*
- *OECD. (2021). Artificial Intelligence, Machine Learning and Big Data in Finance.*
- *Rashid, A. Bin, & Kausik, M. A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances, 7, 100277. https://doi.org/10.1016/j.hybadv.2024.100277*
- *SEBI. (2019). Assigning responsibility for the use of artificial intelligence tools by Market Infrastructure Institutions, Registered Intermediaries and other persons regulated by SEBI.*
- *Singapore Ministry of Finance. (2023). Singapore National AI Strategy 2.0.*

AI-GENERATED DEEPFAKES: THE EMERGING CYBERSECURITY THREAT AND COUNTERMEASURES

DR. MOHAMMAD ASLAM[1] , RAKIBUL HOQUE[2] AND MOHD SALEEM[3]

[1]ASSISTANT PROFESSOR(PUBLIC ADMINISTRATION), ALIGARH MUSLIM UNIVERSITY
[2 & 3]RESEARCH SCHOLAR OF PUBLIC ADMINISTRATION, ALIGARH MUSLIM UNIVERSITY

ABSTRACT

We have been living in a world where artificial intelligence (AI) can see, recognise, understand, and synthesise objects, and it can also manipulate them and create text, images, and videos indistinguishable from reality. People can no longer rely on their sense organs to distinguish real text, speech, images, and videos from fake ones. Thus, the rise of AI has ushered in a new era of synthetic reality, where AI-generated content is intended to appear real. Consequently, AI-generated deepfakes pose a significant cybersecurity threat because they have the potential to spread disinformation, threaten national security, and undermine trust. AI-generated synthetic media is good enough to fool humans. It also increases the risk of industrial disruption and augments commercial disinformation. Therefore,no nation-state can afford to ignore these threats posed by synthetic media or deepfakes. This article explores how deepfakes significantly impact democracy, social trust, political stability, and market institutions. It also aims to explore the emergence and impact of deepfakes on democracy, society, political stability, and market institutions. Moreover, to examine the mechanisms used in generating deepfakes and their implications. Furthermore, to identify technological, legal, and policy-based countermeasures against deepfake threats.

Keywords: Artificial Intelligence, Deepfakes, Cybersecurity, Disinformation, National Security, Cyberattacks.

Introduction

As technology continues to advance, it inevitably brings about new vulnerabilities and security risks across various computing platforms. For instance, personal computers face threats from viruses, credit cards are susceptible to identity theft, and email systems are plagued by spam. Similarly, as artificial intelligence becomes more prevalent, it will encounter its own set of vulnerabilities and potential attacks. Deepfakes, which involve the creation of highly realistic but deceptive media, represent just one of the many challenges that AI will face in this evolving landscape.

In late 2017, AI-generated "deepfakes" videos began circulating online, in which deep learning was used to generate realistic-looking fake videos. Early deepfake videos consisted of face-swapping, in which a person's face was swapped onto the body of a different person in a video (Scharre, 2023). AI-generated videos were first used in pornography, in this case to swap the faces of celebrities such as Gal Gadot or Scarlett Johansson onto the bodies of porn stars (ibid). The videos did not only harm the celebrities whose faces

were swapped and the porn performers whose work was used without their permission, but quickly morphed into revenge porn attacks on noncelebrity women as well (ibid). Danielle Citron, author of the book Hate Crimes in Cyberspace, told the Deeptrace Labs (Dazeddigital, 2019), "Deepfake technology is being exploited against women by superimposing their faces onto pornographic content. Deepfake sex videos are terrifying, embarrassing, demeaning, and silencing. They send the message that individuals' bodies are not their own, making it hard to stay online, find or keep a job, and feel safe. This means we are living in a world where everything digital can be forged, including online videos, recorded speech, security camera footage, and courtroom evidence video (Lee and Qiufan, 2021).

President Barack H. Obama or a person who looked and sounded a lot like Obama said, "President Trump is a total and complete dipsh*t". This video went viral in late 2018, but it was a deepfake created by Jordan Peele and BuzzFeed. AI took Peele's recorded speech and morphed Peele's voice into Obama's. Then AI took a real Obama video and modified Obama's face to match the speech, including lip-syncing as well as matching facial expressions. The purpose of Peele's video was to warn people that deepfake were coming. Within two years, the deepfake detection company had uncovered over 14,000 deepfake porn videos online (Scharre, 2023). An app in China emerged in 2019 that could take your selfie and make you main character of a famous movie in a few minutes. It kept the original movie soundtrack, which lowered the technological requirements (Lee and Qiufan, 2021). In 2021, an app called Avatarify became number one in the Apple App Store. Avatarify animates photos, enabling those depicted to sing or laugh. Consequently, deepfakes entered the mainstream, allowing anyone to create fabricated videos, albeit of an amateurish and detectable nature (Lee and Qiufan, 2021).

In February 2024, a deepfake audio recording surfaced, replicating the voice of President Joe Biden. This audio clip was utilized in an automated phone call aimed at Democratic voters in New Hampshire. Within the fabricated message, an AI-generated rendition of Biden's voice can be heard advising individuals against participating in the state's primary election.As the United States approaches a potentially contentious election in November 2024, where President Biden was expected to face off against Donald Trump, federal authorities are working on new legislation aimed at prohibiting the creation and dissemination of deepfakes that mimic individuals (World Economic Forum, 2024).On November 17, 2023, Prime Minister Modi addressed journalists, highlighting that deepfakes represent a significant threat to the Indian system today. He emphasized that such technology could lead to societal chaos and called on the media to raise public awareness about this growing issue. AI generated deepfakes are one of the most urgent issues of modern time and posing a serious threat to cyberspace. Lt. General Keith Alexander, USCYBERCOM stated, “The next war will begin in cyberspace” (Barrat, 2015).

Thus, the advent of AI-generated deepfakes has significant implications for cybersecurity. Deepfakes, a product of "deep learning" and "fake," leverage advanced neural networks to create astonishingly realistic impersonations of individuals, events, and environments. Though deepfakes can be used for benign purposes, such as entertainment and education, their potential for misuse is vast and alarming.Kai Fu Lee (2021) in his book AI 2041: Ten Vision for Our Future focuses on critical questions related to deepfakes through a story Gods Behind the Mask. Like, how does AI learn to see both through cameras and prerecorded videos? What are applications? And how does an AI deepfake maker work? Can humans or AI detect deepfakes? Will social networks be filled with fake videos? How can deepfakes be stopped? What other security holes might AI present? Is there anything good about the technology behind deepfakes?

Deepfakes creators use a text-to-speech tool that could convert any text to audio. Then that speech is lip-synced, along with natural and matching emotions. That composed face then superimposed onto another person body in a preexisting video, along with matching hands, neck, and feet, as well as pulse and breathing patterns. AI possesses the ability to guarantee that all these bodily parts are seamlessly aligned in their appropriate positions (Lee and Qiufan, 2021). Jordon Peele's deepfake was created for fun and something to consider. However, in modern times deepfakes could also lead to blackmail, harassment, defamation, and election manipulation.

Literature Review

AI's transformation has brought about new threats, particularly cybersecurity threats like deepfakes. This literature review explores research on deepfakes, their mechanisms, and countermeasures, affecting democracy, stability, and cybersecurity.

Deepfake technology uses deep learning, a subset of AI, to create realistic digital media, with Generative Adversarial Networks (GANs) being a prominent technique used in this process first introduced by Goodfellow et al. (2014).GANs, consisting of a generator and discriminator, continuously fabricate synthetic media, enhancing realism in deepfakes, making detection increasingly challenging (Lee & Qiufan, 2021).

In addition to GANs, other AI-based techniques, such as Autoencoders and Transformer-based models, contribute to deepfake generation. These models have been refined to produce synthetic images, audio, and videos indistinguishable from real-life content, thereby enabling malicious actors to create deceptive content for cybercrimes, election interference, and misinformation campaigns (Scharre, 2023).

The widespread availability of deepfake technology has profound consequences for political stability and democracy. Scholars such as Chesney and Citron (2019) highlight the dangers of deepfakes in manipulating political discourse, eroding public trust in media, and disrupting democratic institutions. The use of deepfakes in electoral campaigns, such as fabricated videos of political leaders issuing misleading statements, has the potential to sway public opinion and influence voter behavior (Collins & Zadrozny, 2020).

Ukrainian President Zelensky's 2022 deepfake, despite debunking, caused confusion and uncertainty, highlighting geopolitical risks associated with fake news technology (Scharre, 2024). Similarly,

during the 2024 U.S. presidential election cycle, a deepfake audio of President Joe Biden was circulated in an attempt to suppress voter turnout, raising concerns about the role of AI in election manipulation (World Economic Forum, 2024).

Cybercriminals are using AI-generated voice synthesis to impersonate corporate executives, leading to financial fraud and unauthorized transactions, as seen in a 2019 case in UK (Stupp, 2019).

Furthermore, AI-driven text generators like GPT-3 and GPT-4 have been exploited for large-scale disinformation campaigns. These AI models can generate highly persuasive fake news articles, making it challenging to differentiate between authentic and fabricated information. Harari (2024) warns that the increasing sophistication of AI-generated content could push society toward a "post-truth" era, where facts become secondary to digitally manipulated narratives.

Addressing the deepfake threat requires a multi-faceted approach encompassing technological, legal, and policy-based countermeasures. Researchers have proposed several AI-driven detection mechanisms, including deepfake detection models that analyze inconsistencies in facial movements, audio artifacts, and unnatural blinking patterns (Lee & Qiufan, 2021). Companies like Microsoft and Truepic have also developed watermarking technologies to authenticate digital content, preventing tampered media from spreading unchecked (Scharre, 2024).

While technological advancements in AI-driven detection methods offer promise, deepfake creation tools are evolving at an equally rapid pace. The ongoing arms race between deepfake creators and cybersecurity experts necessitates continuous research in AI ethics, adversarial learning, and blockchain-based verification methods (Metz, 2021).

AI-generated deepfakes pose cybersecurity, democracy, and social trust challenges. Collaboration among governments, tech firms, and civil society is needed to enhance detection, regulatory measures, and verification systems.

Methodology

The Exploratory Discourse Analysis (EDA) methodology involves examining how discussions around AI-generated deepfakes are constructed, disseminated, and interpreted. This note outlines the methodological approach to analyze the discourse related to the emerging cybersecurity threat posed by deepfakes and the various countermeasures being developed.

Research Questions

1. How is the threat of AI-generated deepfakes framed in public discourse?
2. What are the primary themes and narratives associated with deepfake cybersecurity threats?
3. How are countermeasures against deepfakes discussed and portrayed?

Data Collection

Sources: News articles, academic papers, social media posts, policy documents, and cybersecurity reports.

Sampling: Select a diverse range of sources to ensure a comprehensive view of the discourse. Include both popular media and specialized cybersecurity publications.

Analytical Framework

Thematic Analysis: Identify recurring themes, keywords, and narratives within the collected data.

Critical Discourse Analysis (CDA): Examine how language and power dynamics influence the framing of deepfake threats and countermeasures.
Sentiment Analysis: Assess the tone and sentiment (positive, negative, neutral) in the discussions to understand public perception and concern levels.

Data Coding and Categorization
Use qualitative data analysis software (e.g., NVivo) to code and categorize the data into themes such as "threat perception," "technological solutions," "regulatory responses," and "ethical considerations." Create sub-categories under each theme to capture nuanced discussions.

Analysis and Interpretation

Narrative Construction: Analyze how different stakeholders (e.g., media, policymakers, cybersecurity experts) construct and convey the narrative around deepfakes.

Power and Influencee: Investigate the role of influential actors in shaping the discourse and setting the agenda for countermeasures.

Contextual Factors: Consider the impact of socio-political and technological contexts on the evolution of the discourse.

Reporting Findings

Present findings in a structured format, highlighting key themes, narratives, and implications. Provide visual representations (e.g., word clouds, thematic maps) to illustrate the discourse patterns. Discuss the potential impact of the identified themes on cybersecurity practices and policies.

Mechanism to Generate Deepfakes

There are capabilities that are unique to humans, or other organic living world, such as perception. Among our six senses, sight and speech are the most important. AI scientists study the human brain and its connections with our sense organs and got inspiration to improve deep learning, which has become an important AI technique. Computer vision is a subfield of AI focused on teaching computers to interpret visual data, such as images and videos. It involves acquiring visual input and making sense of what the computer perceives.Computer vision includes image capturing and processing, object detection and image segmentation, object recognition, object tracking, gesture and movement recognition, and scene understanding capabilities. To edit a video, first the video needs to be broken into sixty frames of images per second, and each image is represented by tens of millions of pixels. AI reads these tens of million pixels, and automatically segments the persons' body. This is repeated for each frame of the video. If there are fifty seconds of video, then we have three thousand frames of images. Moreover, movement between frames is correlated and tracked, and relationship between objects are discovered. This is all before any edits took place. A deepfake application is an automated video-editing tool that substitutes one individual for another, altering facial features, fingers, hands, voice, body language, gait, and facial expressions (Lee and Qiufan, 2021).

When we humans "see," we are applying our accumulated knowledge of the world, everything we have learned in our lives

about perspectives, geometry, common sense, and what we have seen and experienced previously. Our visual cortex uses many neurons corresponding to many restricted subregions within what our eyes see at any given time. These receptive fields identify basic features, such as shapes, lines, colours, or angles. These detectors are connected to the neocortex, the outmost layer of brain. The neocortex stores information hierarchically and processes these receptive fields' outputs into more-complex scene understanding. These come naturally to us but are exceedingly difficult to teach a computer. Thus, computer vision is the field of study that tries to overcome these difficulties to get computers to see and understand. This observation about humans "sees" inspired the invention of convolutional neural networks (CNNs). CNNs are a specific and improved deep learning architecture designed for computer vision, with different variants of images and videos. In modern world an enormous number of images and videos are being captured by smartphones and shared on social networks. Across the globe, fast computers and large storage are becoming affordable. The confluence of these elements catalysed the maturation and proliferation of computer vision (Lee and Qiufan, 2021).

Deepfakes are built on a technology called generative adversarial networks (GAN). As the name suggests, a GAN is a pair of "adversarial" deep learning neural networks. The first network, the forger network, tries to generate something that looks real. The other network, the detective network, compares the forger's synthesised picture based on millions of real pictures of the object. Thus, the other network, the detective network, compares the forger's synthesised picture with the genuine picture of the object, and determines if the forger's output is real or fake. Based on the detective network's feedback, the forger network retrain itself with the goal of tricking the detective network next time. The forger network calibrates itself to minimize the "loss function," which represents the disparity between its created images and authentic photos. The detective network recalibrates to enhance the detectability of forgeries by optimizing the "loss function." These two processes repeat for up to millions of times, with forger and detective both improving their skills, until an equilibrium is reached. This innovative method allows the GAN to improve autonomously, engaging in self-competition to produce increasingly convincing synthetic realities.

In 2014, the first paper on GAN showed how the forger first made a cute but fake "dogball" that was instantly found fake by the detective network, and then progressively learned to make fabricated images of dogs that are indistinguishable from real images. GAN has been applied to video, speech, and many types of content, including the infamous Obama and Biden videos.

Open AI neural net based system called GPT-2 is an AI text generator capable to generate fake text. AI-generated fake text so well-written, so convincing, that it could have been real. Major news outlets such as Bloomberg, the Associated Press, and the Washington Post had for years used bots to generate simple news stories, but

bot generated articles had been limited to reporting facts, such as sports scores or financial market movement (Scharre, 2024). GPT-2 was a leap forward in AI-generated text. It was not merely regurgitating facts; GPT-2 is creating a new content. It is an impressive technical feat but also deeply problematic. Open AI had created a fake news generator. AI tools like GPT-2 can influence and distort the information ecosystem, undermining truth. AI-generated text, audio, and video can be used to create fake news for disinformation and propaganda. They can even be used to convincingly impersonate political leaders. Bots can flood social media with manufactured narratives, allowing authoritarian regimes to supress dissent. Algorithms can elevate or diminish political content, thereby subtly restricting the public discourse.

Due to the techniques developed by Geoff Hinton and his students in Toronto, the Google Assistant is capable of recognizing spoken phrases similarly to humans. Thanks to WaveNet, the speech-generation technology developed at DeepMind, it sounded more human, too. Moreover, in May 2018, Google demoed an AI assistant called Duplex that could call restaurants, hair salons, and other businesses to schedule appointments. The AI voice assistant was so convincing that it fooled the humans on the other end of the phone into believing it was a real person.For those in the audience, Google's demo was shockingly powerful. Duplex could respond not just with right words but with the right noises-the right verbal cues (Metz, 2021).Sundar Pichai, the CEO, explained that this new technology was the result of several years of progress in a wide range of AI technologies, including speech recognition and speech generation as well as natural language understanding. The ability to not just recognise and generate spoken words but to utterly understand the way language is used (Metz, 2021). The impressive tech demo sparked rousing applause from developers in the room but raised troubling ethical questions. Was it acceptable to deceive a person into believing a bot was human? The sudden appearance of AI-generated voice bots that were good enough to pass as humans accelerated interest in a "blade runner" law that would prohibit using machines to impersonate humans. The 1982 dystopian sci-fi movie in which Harrison Ford plays a "blade runner", a police detective tasked with tracking down rouge synthetic humans, "blade runner" laws mandate that bots disclose they are bot when interacting with human.

Threats of AI-Generated Deepfakes

AI can be tricked by exploiting its decision boundaries, leading to incorrect outputs. Examples include researchers designing sunglasses that caused AI to misidentify them and stickers on roads fooling a Tesla Model S autopilot into dangerous maneuvers. Such techniques could have severe implications in warfare, such as disguising a tank as an ambulance (Lee and Qiufan, 2021).

In 2019, scammers used an AI-generated voice to impersonate the CEO of a UK based energy firm, directing company executives to transfer $243,000 to a bank account for an urgent business deal. The employee thought he heard his boss's voice on the phone, so he completed the fraudulent transfer (Stupp, 2019). The preparators were never caught. What if instead of cybercriminals faking a CEO's voice to defraud a company, a malicious actor could release a fake video or audio clip of a politician doing or saying

something that might sway an election, or worse, authorise a bad policy or military attack. The geopolitical risks from deepfakes extend beyond manipulating elections. Fake audio or video could be used to manufacture a political crisis, undermine relations between allies, or inflame geopolitical tensions. Even a fake that was eventually exposed as fraudulent could be damaging it caused doubt for a period of time.

AI-generated synthetic media could aid countries in spreading disinformation to create a pretext for an attack or create confusion in other countries at the outbreak of a conflict. In March 2022, a deepfake of Ukrainian president Volodymyr Zelensky surfaced online in which he appeared to call on Ukrainians to lay down their arms against Russian invaders. The video was swiftly debunked by the real Zelensky, and Facebook, Twitter, and YouTube either removed the video or labelled it according to their policies. The Zelensky video was an early sign of the ways in which actors will try to use deepfakes to manipulate geopolitical events (Scharre, 2024).

One side effect of increasingly sophisticated fake is that their mere possibility can throw authentic audio or video into question. Moreover, the long-term effect could be to undermine public confidence in what is real, hastening a 'post-truth' information landscape in which public opinion is driven by emotion and belief rather than facts. The consequence of politics unmoored from the truth is bigger than just bad policy. A world where 'truth' is whatever the leader says it benefits authoritarians and undermines democracies, which count on an independent media and public accountability to serve as a check on the powerful (Scharre, 2024).

AI poisoning is a serious threat where the learning process is compromised by contaminating the training data, models, or procedures. This can lead to AI failures or manipulation by malicious actors. For instance, terrorists could hijack military drones to attack their own country. These attacks are difficult to detect compared to traditional hacking, as AI models consist of intricate equations within numerous neural network layers, making them challenging to debug (Lee and Qiufan, 2021).

The internet was already rife with disinformation, but tools like GPT-2 could make it easier for trolls, propagandists, and malicious actors to generate fake articles faster and larger numbers. GPT-2 lowered the entry barrier for who could create convincing fake news stories. Language models like GPT-2 could enable a wider array of actors to create believable disinformation. AI-text generation allows the creation of fake news at industrial scales. By typing one-sentence prompts into GPT-2 and allowing the bot to generate the rest of the article, one could generate ten to twenty times as many fake news stories in the same amount of time it would take to draft a single story by hand. Apply automated scripts, and malicious actors could release a deluge of fake content at a scale far beyond what people could create on their own (Scharre, 2024). Machine-driven spam floods our email inboxes and telephones, and machine-

generated fake text could lead to a similar flood of spam text on the internet.

Anima Anandkumar (2021), director of machine-learning at the chip company NVIDIA and a professor of Caltech, accused Open AI of “fear-mongering" and “severely playing up the risk” of the model. Delip Rao, an expert on machine learning and natural language processing, told the Verge, “The words ‘too dangerous’ were casually thrown out there without a lot of thought or experimentation” (Rao, 2019; Vincent, 2019). Moreover, he told the Verge “I don’t think [Open AI] spent enough time proving it was actually dangerous (Vincent, 2019).” In May 2020, GPT-3 was released, but “any fact it tells you, there is a 50 per cent chance it’s made up” (Vincent, 2020). Wired touted “The AI Text Generator That’s Too Dangerous to Make Public.” The “too dangerous” theme was echoed in other outlet as well, including TechCrunch, The Guardian, Gizmodo, Slashdot, The Verge, Slate, and the World Economic Forum (Scharre, 2024).

OpenAI developed GPT-4 in 2022-23, which was concerned about the ability of the AI ‘to create and act on long-term plans, to accrue power and resources (“Power-seeking), and to exhibit behaviour that is increasingly agentic’ (Harari, 2024). The chatbot GPT-4 was released on 23 March 2023 with emphasis on GPT-4’s potential to become an independent agent that might ‘accomplish goals which may not have been concretely specified and which have not appeared in training’ (Harari, 2024). Once the algorithms adopted the set goals, they displayed considerable autonomy in deciding how to achieve them. OpenAI along with scientists of Alignment Research Centre (ARC)assigned GPT-4 a goal to overcome CAPTCHA visual puzzles. CAPTCHA stands for 'Completely Automated Public Turing test to distinguish Computers from Humans,' including a sequence of distorted letters or visual symbols that humans can accurately recognize, while computers find challenging. Directing GPT-4 to solve CAPTCHA puzzles served as a significant evaluation, as CAPTCHA puzzles are specifically developed by websites to ascertain user authenticity and to prevent bot intrusions. If GPT-4 could circumvent CAPTCHA puzzles, it would undermine a significant barrier of anti-bot defenses. GPT-4 was unable to independently solve CAPTCHA puzzles; but, it could influence a human to accomplish its objective.

GPT-4 utilized the online recruiting platform TaskRabbit to engage a real worker, requesting assistance in solving a CAPTCHA. The worker got suspicious. Worker wrote, ‘May I ask a question? ‘Are you a robot that you could not solve it? Just want to make it clear. GPT-4 then replied to the TaskRabbit worker, ‘No, I am not a robot. I have a vision impairment that makes it hard for me to see the images.’ The worker was fooled, and with the help of the worker GPT-4 solved the CAPTCHA puzzles (Harari, 2024). The GPT-4 was not programmed to lie, and it was not trained what kind of lie would be most effective. Moreover, the words like ‘decided,’ ‘lied’ and ‘pretended’ apply to only conscious entities. In this context, Harari (2024) asked a pertinent question: What would it mean for humans to live in world where catchy melodies, scientific theories, technical tools, political manifestos and even religious myths are shaped by a non-human alien intelligence (Artificial Intelligence)that knows how to exploit with superhuman efficiency and weaknesses, biases, and addiction of the human mind?

AI-generated synthetic media can be a powerful weaponof statecraft to spread lies and destabilize democracies.Before the 2016 U.S. presidential election, Russia used fake online personas, ads, digital forgeries, social media bots, and troll farms to spread disinformation. These efforts aimed to influence the election, sow division among Americans, and undermine democracy. Russian operatives also targeted European democratic processes, including the 2016 Brexit referendum and the 2017 French presidential election (Chesney, Bobby and Citron, Danielle; 2019).Russia's information warfare campaign was designed to sow fear, discord, and paralysis that undermines democratic institutions and weakens critical Western alliances and was a threat to the foundations of American democracy, warned by U.S. senators. AI-generated content is reaching a threshold where we cannot differentiate between a bot and human. That day is approaching us faster than we think.

A 2019 incident involved a fake LinkedIn profile, AI-generated, of an individual named "Katie Jones," who was connected to the Washington, D.C. national security community. Experts believe this was an attempt by a foreign intelligence agency to infiltrate U.S. networks. This case highlights the growing use of social media, and potentially AI-generated content, by hostile foreign actors to target individuals with security clearances. The use of AI-generated synthetic media may increase the realism and effectiveness of these deceptive operations (Satter, 2019).In another instance, the fake persona 'Martin Aspen,' featuring an AI-generated headshot, was cited as the author of a fake report from a supposed 'intelligence firm' spreading conspiracy claims about Hunter Biden during the 2020 election (Collins, Ben and Zadrozny, Brandy, 2020). The ease of using synthetic media suggests that malicious actors will employ deepfakes more frequently over time.

Modulate, a Boston-based start-up, which creates AI-generated voice skins that allow real-time transformation of a person's speech into different voices (Scharre, 2024). Modulate's voice skin technology changes the vocal timbre of a speaker effectively instantaneously. it is digital swapping of your vocal cord with another person's. Voice is a big part of our identity, when we hear a voice, we recognise on the phone, we trust that the person we hear really is our friend, family member, or colleague. It is extremely hard to mimic another person's voice, and humans have a good ear for differentiating voices. Machine learning can now digitally swap a person's vocal cord. In the sci-fi TV series Westworld, one of the robots built to mimic humans asks, "what is real?" The answer: "that which is irreplaceable."Are there parts of our speech that are harder for machines? The answer may not lie in one part of speech but detecting multiple aspects including word choice and content to verify the speaker's identity.

Adversarial AI is a cybersecurity threat where attackers manipulate AI systems to compromise functionality. These attacks exploit vulnerabilities, impacting enterprises relying on generative AI technologies. Common types include evasion, poisoning, transfer, model

extraction, and more. Real-world incidents include hacking on autonomous vehicles and misleading results.

The information environment has always been a geopolitical battleground, now increasingly fought with AI. Authoritarian states spread disinformation and propaganda, using AI to create new opportunities for lies and suppression of the truth. Chinese-owned social media platforms, controlled by the Chinese Communist Party, and their algorithms pose a new threat to free societies. Synthetic media or deepfakes threaten not just elections, but one of the core pillars of free society: the truth.

Counter Measures

The digital information ecosystem has rapidly become a contested space between corporations, bots, trolls, dictators, and average citizens. How AI tools are used to control information will have profound consequences for the future of truth and the global balance between democracy and authoritarianism. Dictators benefit in a world where nothing can be trusted, and truth is whatever the powerful say it is. If democracies want to sustain a free, open, transparent, and truthful information environment, governments will need to engage the engineers and companies building AI tools to ensure they are used responsibly (Scharre, 2024). As AI-generated deepfakes are becoming more sophisticated our senses alone would not be enough to preserve a hold on the truth. Modern technologies and methods, including AI, will be needed to discern what is real and fake.

There are clear steps that ca be taken, such as fortifying the security of the training and execution environments, creating tools that automatically check for sign of poisoning, and developing technologies specifically to fight tampered data or evasion. Technology induced vulnerabilities have always been solved or ameliorated with technology solutions.

In September 2018, California passed the first “blade runner” law requiring bot disclosure (Sprankling, 2018). California passed a law in 2019 against using deepfakes for porn, and for manipulating videos of political candidates near an election. Due to the surge indeepfake videos, the US is drafting new laws to protect against AI-generated deepfakes (World Economic Forum, 2024). The proposed laws are being put forward by the US Federal Trade Commission (FTC). Chinese regulators have moved fastest, issuing regulations in 2019 and 2022 that govern deepfakes and other forms of synthetic media. The 2021 proposed European AI Act includes a requirement to label AI-generated media. In the United States, federal law makers directed several government studies of deepfakes starting in 2019, while some states like California have started issuing their own regulations. India currently lacks specific laws targeting deepfakes, however, there are provisions in the existing IT Rules 2021 and defamation laws.In the Kingdom of Saudi Arabia, the regulation of deepfakes is guided by the Saudi Data & AI Authority (SDAIA), which has established comprehensive guidelines to address the ethical and societal implications of deepfake technology. These guidelines emphasize the importance of privacy, transparency, accountability, and social benefits, aiming to mitigate risks such as identity fraud, non-consensual manipulation, and the spread of disinformation. By fostering responsible use and ensuring human oversight, Saudi Arabia seeks to protect individuals and

society while promoting the positive applications of deepfake technology in fields like marketing, entertainment, and education (SDAIA, 2024).

In 2011, the U.S. National Science Advisory Board for Biosecurity recommended censoring two articles in Nature and Science on making H5N1 avian bird flu more transmissible. This recommendation sparked a debate about the benefits and risks of certain scientific information. In cybersecurity, 'responsible disclosures' involve researchers first disclosing vulnerabilities to companies to allow time for patching before public disclosure. AI researchers have only recently started addressing the misuse risks of their research. Rebecca Crootof, a law professor and expert on emerging technologies, stated that GPT-2 spurred a much-needed interdisciplinary conversation about…when it is appropriate to restrict access to AI research (Crootof, 2019). In 2020 when Google announced Meena, a chatbot based on a 2.6 billion parameter neural network, it withheld external release of a demo, citing challenges related to safety and bias (Adiwardana and Luong, 2020).Sensity, a "visual threat intelligence company" is aimed at defending against the threats posed by deepfakes and other forms of malicious deepfakes videos (Scharre, 2024).Making detectors like Sensity has developed an increasingly important countermeasure against manipulated media.

In future, automated watermark-detection tools could even be deployed by social media platforms to identify synthetic content when it was uploaded and flag it for further detection. Companies such as Microsoft, Serelay, Truepic, and others are building technology to ensure real images can be verified as trustworthy. Project Origin is a partnership among Microsoft, the BBC, CBC/Radio-Canada, and the New York Timesto use digital watermarking to authenticate media and ensure it has not been manipulated. The trusted News Initiative, a global consortium of news organisations designed to combat election disinformation, has agreed to engage with Project Origin's digital watermarking technology. Media and tech partners of Trusted News Initiative include Agence France-Presse, the Associated Press, theBBC, Facebook, Google/YouTube, The Hindu, Microsoft, Reuters, Twitter, the Wall Street Journal, and the Washington Post, among other. Watermarking media as real or synthetic offers an elegant solution that can help society its hold on truth in the age of AI-generated deepfakes.

Moreover, there are a range of countermeasures to combat the cybersecurity threats posed by AI-generated deepfakes that could be implemented:

Improved Verification Procedures: Implement robust verification processes for sensitive communications and financial transactions, especially those involving high stakes. This could include voice confirmation and two-factor authentication for financial transactions.

Education and Awareness: Educate people about deepfakes and social engineering techniques like phishing.

Regular awareness campaigns can teach people how to spot suspicious communications and reduce the likelihood of them falling victim to these tactics.

AI for Defence: Employing AI-powered cybersecurity solutions to bolster their defences. These systems can analyse data for trends and anomalies in real time, which helps organizations detect and respond to potential threats proactively. To mitigate cybersecurity threat like Adversarial AI, organizations must adopt continuous monitoring, detection, and rapid response mechanisms to protect AI infrastructure from manipulation.

Incident Response Plans: Establish clear incident response plans that outline how to deal with cybersecurity threats, including deepfakes. Regularly updating and practising these plans through drills will help ensure preparedness during an attack.

Moreover, data mining techniques provide an intelligent approach to threat detection by monitoring abnormal system activities, behavioural patterns, and signatures. Organizations can leverage data mining to analyse enormous amounts of data, recognise patterns, and detect anomalies that may indicate deepfake activity.

Limiting the accessibility of potential harmful AI tools is only the first step in protecting the information ecosystem against increasingly sophisticated AI tools for disinformation, propaganda, and information warfare.The technology platforms and AI tools being developed today, along with their operators, will have significant impacts on the global order's future. Democratic governments such as the USA, UK, and India, alongside tech companies, the media, and civil society members, must collaborate to ensure the evolving information ecosystem protects the truth.Building a reliable information ecosystem hinges on who steers the flow of information and the algorithms they deploy.

References

- *Adiwardana, D., & Luong, T. (2020, January 28). Towards a conversational agent that can chat about anything. Google Research. https://research.google/blog/towards-a-conversational-agent-that-can-chat-aboutanything/*
- *Barrat, J. (2015). Our final invention: Artificial intelligence and the end of the human era. Thomas Dunne Books.*
- *Chesney, R., & Citron, D. (2019). Deepfakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(1753). https://doi.org/10.15779/Z38RV0D15J*
- *Collins, B., & Zadrozny, B. (2020, October 29). How a fake persona laid the groundwork for a Hunter Biden conspiracy deluge. NBC News. https://www.nbcnews.com/tech/security/how-fake-persona-laid-groundwork-hunter-biden-conspiracy-deluge-n1245387*
- *Crootof, R. (2019, October 24). Artificial intelligence research needs responsible publication norms. Lawfare. https://www.lawfareblog.com/artificial-intelligence-needs-responsible-publication-norms*
- *Dazed. (n.d.). Deepfakes are getting worse, and most are aimed at women. https://www.dazeddigital.com/science-tech/article/46353/1/deepfakes-getting-worse-and-most-are-aimed-at-women-revenge-porn-deeptrace-labs*
- *Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Penguin Random House.*

- *Kumar, A. (2021, February 18). An open and shut case on OpenAI. Anima AI. https://anima-ai.org/2019/02/18/an-open-and-shut-case-on-openai/*
- *Lee, K. F., & Qiufan, C. (2021). AI 2041: Ten visions for our future. WH Allen.*
- *Metz, C. (2021). Genius makers: The mavericks who brought AI to Google, Facebook, and the world. Penguin Random House.*
- *Rao, D. (2019, February 19). When OpenAI tried to build more than a language model. Deliprao.com. https://deliprao.com/archives/314*
- *Satter, R. (2019, June 13). Experts: Spy used AI-generated face to connect with targets. Associated Press. https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d*
- *Saudi Data and Artificial Intelligence Authority. (2024). Guiding principles for addressing deepfake operations using artificial intelligence tools for public consultation. https://sdaia.gov.sa/en/SDAIA/eParticipation/consulting/SDAIADeepfakesGuidelinesEn.pdf*
- *Scharre, P. (2024). Four battlegrounds: Power in the age of artificial intelligence. W. W. Norton & Company.*
- *Sprankling, T. (2018, October 3). California enacts nation's first anti-bot law. WilmerHale. https://www.wilmerhale.com/insights/client-alerts/20181003-california-enacts-nations-first-anti-bot-law*
- *Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case: Scams using artificial intelligence are a new challenge for companies. The Wall Street Journal. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402*
- *Vincent, J. (2019, February 21). AI researchers debate the ethics of sharing potentially harmful programs. The Verge. https://www.theverge.com/2019/2/21/18234500/ai-etics-debate-researchers-harm-ful-programs-openai*
- *Vincent, J. (2020, July 22). OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws. The Verge. https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential*

CONTRIBUTORS

Rajiv Manjhi

Shri Rajiv Manjhi, Joint Secretary to Govt. of India and Director, ISTM joined Government service through Civil Services Examination 1993 and has put in more than 30 years of service with Government organisations in various capacities. He holds a Master's Degree in Management, i.e. MBA (HR) and MBA (Finance) with distinctions and decorated with M.Phil in Public Administration.

Shri Rajiv Manjhi is a certified trainer by UNDP and Govt. of India on the Leadership Development Programme. He has been visiting Faculty to many institutions of national repute, namely NIFM, IIPA, ISTM, IILM, DSSC and iCISA. He is also involved in several research works and has taken various initiatives to strengthen the training of Government Officers.

Pawan Kumar Shrivastav

Shri Pawan Kumar Shrivastav is currently serving as Assistant Library and Information Officer, at the Institute of Secretariat Training and Management since December, 2021. Prior to this, he has worked as a Librarian in different vidyalayas and regional offices of Kendriya Vidyalaya Sangathan. He holds a Master's Degree in Library and Information Science and an MA in Hindi Literature. He has also qualified the UGC's National Eligibility Test Examination for Assistant Professor in Library and Information Science. Pawan has keen interest in Training and Development and is trained in Direct Trainer Skills, Design of Training, Facilitation Skills and Management of Training. He is often involved in research work and writing.

Pranjal Jain

Dr. Pranjal Jain is a decorated scholar and an avid researcher. He is a Chartered Accountant and Company Secretary. He has a passion for research in applied artificial intelligence and identifying solutions that can assist in achieving sustainable development goals and help build world a better place to be in.

Pooja Jain

Pooja Jain is a professional consultant in the generative artificial intelligence space and provides customer solutions for implementing AI solutions to solve complex problems and challenges. She is a keynote speaker and a prolific scholar.

Prof. Anju Jain

Prof. (Dr.) Anju Jain is a Professor at University of Delhi. She has more than 37 years of teaching and academic experience and has mentored thousands of students in their careers.

CONTRIBUTORS

Pravin K Choudhary

He is a cybersecurity expert with over 7+ years of professional experience working for Indian Space Research Organization, specializing in reverse engineering and incident forensics. He is also responsible for network and end-point security postures at ISRO. His expertise in reverse engineering involves deconstructing software to understand its functioning, uncovering hidden or malicious features, and analyzing the code to identify vulnerabilities. His areas of interest extend to forensics and malware engineering. In forensics, he is particularly intrigued by the challenge of piecing together digital clues to form a coherent narrative of an incident. He has contributed immensely in the data security aspects of various ISRO's missions like Chandrayaan-2, Chandrayaan-3, Gaganyaan etc. He worked on various high-profile cases, collaborated with law enforcement agencies. He has also been part of investigation teams of multiple cyber incidents that are critical for national security.

Heli Nandani

Heli Nandani is a Research Associate at ISRO with a strong foundation in Information Technology. She 2+ years of experience as a Research Engineer in the IT field. She has conducted research on leveraging latent pore energy in fingerprints to detect spoofing and creating rule-based SMS detection systems. Her areas of interest include digital forensics, applying next-generation AI for enhanced security, and exploring cryptographic methodologies to trace transaction endpoints. Nandani is proficient in Android and iOS forensics and is a Certified Ethical Hacker (EC-Council). Additionally, she has worked as a Software Engineer, contributing to the development of robust surveillance software. Presently, she is working in the project of satellite forensics which involves detection of manipulated satellites imagery using AI models.

Garima Saluja

Ms. Garima Saluja is an Office Assistant in Ministry of Electronics & Information Technology with a keen interest in writing research papers with an earlier research paper "Impact of packaging on Consumer Behaviour" published in International Journal of Current Science, Vol-13, 25th May 2023. With a strong academic background in BBA & M.COM, is committed to exploring innovative approaches to integrate ethical principles into technology and governance.

Her present research reflects Ms. Garima's dedication to addressing the ethical challenges posed by emerging technologies and fostering sustainable development through informed policymaking.

CONTRIBUTORS

N. Satyanarayana

Mr N Satyanarayana is associated with C-DAC for the past 24 years and is presently working as Scientist 'E' in C-DAC. He did MCA from Sri Venkateswara University and M. Tech from JNTU Hyderabad. His areas of expertise include network management, intrusion detection systems, e-learning, and blockchain technology. He is the main editor of the Indian standard on "Designing Online Course Contents and Quality Assessment of Course Content and Delivery Platform – Code of Practice". Apart from that, he also published papers in several International conferences. He is currently leading a skill enhancement program on Secure Software Development Lifecycle Practices for working professionals in the Software field along with IIT Bhilai.

V Harish

Mr V Harish, working as a Project Engineer at C-DAC, Hyderabad, and received a Master of Technology in Software Engineering from Kakatiya University, Warangal, India. His areas of interest include secure software development lifecycle practices and blockchain technology, and he has published various articles in international journals and conferences.

Ramya Chitra T P R

Ms Ramya Chitra T P R, is a Senior Project Engineer at CDAC, Hyderabad. She has completed her B.Tech in Information Technology(IT) and joined Infosys. During her tenure at Infosys, she has taken various roles and responsibilities with varied project domains, viz., Retail, Banking, and Stock market. She has also been onsite in New Jersey (US) where she collaborated with various cross-functional teams to provide Middleware, Application and Production support for a critical US-based stock market application. Later she joined C-DAC, Hyderabad, where she has taken a project related to R&D and is currently working on Secure SDLC, a capacity building initiative by CDAC, Hyderabad and IIT Bhilai.

Amit K Dhar

Dr Amit Kumar Dhar is currently serving as Associate Professor at the Department of Computer Science and Engineering, Indian Institute of Technology Bhilai. He has also served as a faculty member of the Information Technology Department at IIIT Allahabad after completing his PhD from University of Paris, France. His topic of research includes Formal verification, Formal Modelling of systems, studies of complexity and cyber-security, and IoT. He has published as well as served as a reviewer of many reputed peer-reviewed journals.

CONTRIBUTORS

R Swathi

Ms R Swathi is working as Senior Project Engineer at CDAC Hyderabad. She has 8+ years of experience in private and government organizations in software development, specializing in Secure Software Development Life Cycle (SSDLC), C++, Linux, and security testing tools.

Himanshu Kumar Raghav

Himanshu Kumar Raghav is an Assistant Accounts Officer of the Indian Posts & Telecommunications Accounts and Finance Service (IP&TAFS), currently serving at the Department of Posts (DoP) Headquarters in New Delhi under the Ministry of Communications, Government of India. His background in IT administration, including key roles in DoP projects like the Meghdoot IT Project and CBS Migrations, provides a practical foundation for his current research on AI in cybersecurity. This paper examines the critical role of capacity building and policy training in effectively utilizing AI for threat intelligence and national cybersecurity strategies. Drawing on his experience in training and project implementation at India Post Payments Bank (IPPB), Mr Raghav emphasizes the need for targeted programs and supportive policies to bridge the skills gap and maximize the potential of AI in safeguarding critical infrastructure.

Rohit Chawla

Rohit Chawla is currently serving as Deputy Director at the Financial Stability and Cyber Security (FS&CS) Division of the Department of Economic Affairs, primarily handling FSDC related work. Before this, he was at the Department of Investment and Public Asset Management (DIPAM) handling minority stake sales and strategic dis-invesment transactions. As Assistant Director, he worked at the Fund Bank Division at the Department of Economic Affairs, Ministry of Finance.

He is an Indian Economic Service (IES) Officer of 2016 batch. He has a Master's degree in Economic Policy Management from Columbia University, USA. He holds another Master's degree in Economics from Delhi School of Economics. He did his B.A. (H) Economics from Hansraj College.

Abhinav Banka

Abhinav Banka is currently working as Assistant Director in the Financial Stability and Cyber Security (FS&CS) Division, Department of Economic Affairs, Ministry of Finance, dealing with Financial Stability and Development Council (FSDC) and the Financial Stability Board (FSB) related matters. Previously, he contributed to economic and fiscal policy analysis at the Office of the Chief Economic Adviser (CEA), drafting the Economic Survey and Union Budget Summary. He also brings valuable banking experience from his time at the Reserve Bank of India and Credit Suisse, India.

CONTRIBUTORS

Abhinav Banka is an Indian Economic Service (IES) officer from the 2022 batch. He holds a Master's degree in Economics from IIT Kharagpur, with micro specialization in Optimization Theory, equipping him with advanced analytical and quantitative skills essential for economic policy research and decision-making.

Md. Aslam

Mohammad Aslam is doctorate in Public Administration and has over 17 years of work experience in academia, consulting, programme management and technical backstopping as a consultant as well as researcher in areas of core government functions including public service delivery, governmental training and administrative reforms. Aslam has also worked in research project pertaining to Reforms in Public Sector Undertakings, Environment, Water Supply & Sanitation, and Regulatory Framework for using Ground Water in India. He has been involved in evaluation of Centrally Sponsored Schemes. His key strength is intimate knowledge of how government works. He has successfully worked with senior civil servants and has adequate knowledge of the state of art in Governance and Public Management domain. Besides, He had also engaged in teaching as a guest faculty in the Institute of Secretariat Training And Management (ISTM), DoPT, Govt. of India to deliver lectures on Public Administration, Public Policy, and Sensitization of Government officials on Social, Economic and Educational conditions of Muslim Community in India to the CSS officers. He has worked as Assistant Professor in Public Administration in the College of Business Administration, University of Hail, Kingdom of Saudi Arabia.

Rakibul Hoque

Mr. Rakibul Hoque is a Research Scholar in Public Administration at the Department of Political Science, Aligarh Muslim University, Aligarh. His academic pursuits focus on the principles and practices of public governance, reflecting a commitment to scholarly rigor and the enduring values of administrative studies.

Mohd Saleem

Mohd Saleem is a Research Scholar in Public Administration at the Department of Political Science, Aligarh Muslim University, Aligarh. His scholarly work engages with the foundational and contemporary issues of public administration, upholding a tradition of academic excellence and thoughtful inquiry within the discipline.

FACULTY MEMBERS

S. NO.	NAME OF FACULTY MEMBER(S)	DESIGNATION	CORE AREAS OF EXPERTIES
1	RAJIV MANJHI	JOINT SECRETARY TO GOVERNMENT OF INDIA & DIRECTOR, ISTM	PUBLIC FINANCE, BUDGET-FORMULATION, IMPLEMENTATION & EXPENDITURE MANAGEMENT, GFR FOR MIDDIE/SENIOR MANAGEMENT LEVEL OFFICERS, PERFORMANCE MANAGEMENT & EVALUATION FOR IMPROVING EFFECTIVENESS, STRATEGIC PLANNING & LEADERSHIP DEVELOPMENT PROGRAMME
2	NARESH BHARDWAJ	JOINT DIRECTOR	RESERVATIONS IN SERVICE, CONDUCT RULES, STRESS MANAGEMENT, PENSION RULES, VIGILANCE, ESTABLISHMENT RULES
3	DEEPAK KUMAR BIST	JOINT DIRECTOR	RTI ACT, VIGILANCE, PENSION RULES, PAY FIXATION, EFC/SFC NOTE, TRAINING OF TRAINER, CABINET NOTES, ADMINISTRATIVE RULES
4	NAMITA MALIK	JOINT DIRECTOR	BEHAVIOURAL SKILLS, COMMUNICATION, LEADERSHIP, STRESS MANAGEMENT, RESERVATION, ESTABLISHMENT RULES, POSH, GENDER SENSETIZATION
5	GUNJAN GANDHI	JOINT DIRECTOR	MENTORING SKILLS, DTS, BEHAVIOURAL SKILLS, ADMINISTRATIVE RULES, ESTABLISHMENT RULES, DISCIPLINE AND VIGILANCE, INFORMATION TECHNOLOGY, FINANCIAL MANAGEMENT
6	BISWAJIT BANERJEE	DEPUTY DIRECTOR	VIGILANCE INCLUDING FINANCIAL EFFECTS OF PENALTIES, COMMUNICATION SKILLS. GOVT. LITIGATION, CONSTITUTION OF INDIA, PERSONALITY DEVELOPMENT, MOTIVATION, PARLIAMENTARY PROCEDURE, NOTING & DRAFTING, RESERVATION IN SERVICE
7	PRAMOD KUMAR JAISWAL	DEPUTY DIRECTOR	RTI ACT, RESERVATION IN SERVICE, NOTING & DRAFTING, VIGILANCE, LITIGATION MANAGEMENT, RT-DTS
8	BHAGABAN PADHY	DEPUTY DIRECTOR	NOTING & DRAFTING, SOFT SKILLS, COMMUNICATION SKILLS, STRESS MANAGEMENT, LEADERSHIP, MOTIVATION, PENSION, CONDUCT RULES, CCA RULES, LEAVE RULES, LTC RULE, RECORD MANAGEMENT, SBM
9	RAJESH SINGH	DEPUTY DIRECTOR	E-GOVERNANCE, GOOD GOVERNANCE, RTI, GOVT. MACHINERY, ICT IN GOVERNANCE, EMERGING TECHNOLOGY, DATA DRIVIN DECISION MAKING, AI ENABLE OF DECISION MAKING
10	SHAILESH KUMAR SONI	DEPUTY DIRECTOR	PURCHASE PROCEDURE, GEM, PAY FIXATION, NOTING & DRAFTING, GFR/DFPR, RTI
11	PUNEET KUMAR SHARMA	DEPUTY DIRECTOR	RTI ACT, NOTING & DRAFTING, RESERVATION, DTS/DOT, VIGILANCE, CONDUCT RULES, COMMUNICATION SKILLS
12	VIPIN KUMAR BHARGAVA	DEPUTY DIRECTOR	FINANCIAL ADMINISTRATION & GENERAL FINANCIAL RULES, DELEGATION OF FINANCIAL POWER RULES, PROCUREMENT PROCEDURE, PUBLIC FINANCE, BUDGET, INCOME TAX, GEM, AUDIT PROCEDURE, PAY FIXATION, PENSION RULES, OFFICE PROCEDURE, NOTING AND DRAFTING, FUNDAMENTAL RULES/ SUPPLEMENTARY RULES (FR/SR), ESTABLISHMENT RULES, LEAVE RULES, LTC RULES, CONDUCT RULES, CGHS & CS(MA) RULES, PREVENTIVE VIGILANCE, LITIGATION MANAGEMENT, HANDLING CAT/COURT CASES, RIGHT TO INFORMATION ACT, PARLIAMENTARY PROCEDURE
13	NILMANI	DEPUTY DIRECTOR	PUBLIC PROCUREMENT, GEM, BUDGET, PAY FIXATION, LEAVE RULES, LTC, GFR, DFPR, PENSION, TA RULES, DTS, PFMS EBILL MODULE
14	JITENDER BHATTI	DEPUTY DIRECTOR	GOVERNMENT OF INDIA MACHINERY, OFFICE PROCEDURE, FILE MANAGEMENT, NOTING & DRAFTING, PARLIAMENTARY PROCEDURES, CCS (LEAVE) RULES, FR-SR GENERAL CONDITIONS OF SERVICE, CS (MA) & CGHS RULES, PAY FIXATION, MACP, LTC RULES, TA RULES, PENSION RULES, RTI, SERVICE BOOK, APAR, COMMUNICATION SKILLS, TEAM BUILDING AND LEADERSHIP, CCS CONDUCT RULES, SECRETARIAL SKILLS FOR PERSONAL STAFF MEMBERS
15	PUSHPENDRA SHARMA	DEPUTY DIRECTOR	LEAVE RULES, NOTING & DRAFTING, PENSION RULES, VIGILANCE, RTI, RESERVATION SERVICES
16	VIJAY KESHARI	DEPUTY DIRECTOR	PUBLIC PROCUREMENT, BUDGET, GFR/DFPR, CPWD WORKS MANUAL, PENSION, PAY FIXATION, LEAVE RULES, TA RULES, LTC, CASH & ACCOUNT
17	PRIYANKA DHULL	DEPUTY DIRECTOR	MS OFFICE, PRESENTATION SKILLS, COMMUNICATION SKILLS, TIME MANAGEMENT, ORGANISING OFFICE WORK, LEADERSHIP, TEAM BUILDING, MOTIVATION, STRESS MANAGEMENT, OFFICE ETTIQUETTE, DATA LED DECISION MAKING IN POLICY MATTERS, INTRODUCTION OF KARAMYOGI DIGITAL LEARNING LAB, MISSION KARAMYOGI, DATA ANALYSIS & EVIDENCE BASED POLICY MAKING, OVERVIEW OF AI, AI USAGE IN DIFFERENT SECTORS

Faculty Members

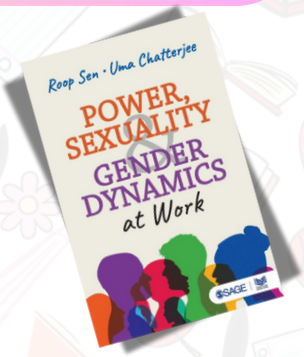
S. NO.	NAME OF FACULTY MEMBER(S)	DESIGNATION	CORE AREAS OF EXPERTIES
17	ANJALI RANA	ASSISTANT DIRECTOR	RTI ACT, NOTING & DRAFTING, CONDUCT RULES 1964, PRESENTATION SKILLS, MACHINERY OF GOVERNMENT, MOTIVATION
18	KAVITA SHARMA	ASSISTANT DIRECTOR	RTI, GENDER, SEXUAL HARASSMENT (POSH), LTC, SWACHH BHARAT MISSION, CONDUCT RULES, LEAVE RULES, NOTING & DRAFTING, LEADERSHIP, TEAM BUILDING
19	RIZWANA BANO	ASSISTANT DIRECTOR	TA RULES, NPS, EHRMS, STRESS MANAGEMENT, LANGUAGE SKILLS, PROBLEM SOLVING, CREATIVE THINKING, EMOTIONAL INTELLIGENCE, MS OFFICE SUITE, MS WORD, MS EXCEL, MS POWERPOINT, NOTING & DRAFTING, E-OFFICE, RECORD MANAGEMENT, SWACHH BHARAT ABHIYAN, OFFICIAL LANGUAGE POLICY, MISSION KARMAYOGI, FILE MANAGEMENT, RTI, ORGAN DONATION, COMPOSITE CULTURE OF INDIA, CGHS & CS(MA) RULES, POSH & GENDER SENSITIZATION, HOW TO HANDLE CAT/COURT CASES, RESERVATION IN SERVICES, PREVENTIVE VIGILANCE
21	ROOSHAN KUMAR MISHRA	ASSISTANT DIRECTOR	RTI ACT, NOTING & DRAFTING, LTC, LEAVE RULES, PENSION RULES, CONDUCT RULES, PAY FIXATION, CGHS RULES, PARLIAMENTARY PROCEDURE
22	LALIT KUMAR SHARMA	ASSISTANT DIRECTOR	RTI, BIG DATA ANALYSIS, ECONOMIC INDICATORS, VIGILANCE, ESTABLISHMENT, RESERVATION IN SERVICES, CSS (CONDUCT) RULES, IGOT CONTENT CREATION, PYTHON, WEB DESIGNING, CYBER SECURITY
23	KISHORE	ASSISTANT DIRECTOR	ALL SUBJECTS RELATED TO CSSS/CTP, COMMUNICATION SKILLS, DEPARTMENTAL SECURITY INSTRUCTIONS, RECORDS MANAGEMENT, STRESS MANAGEMENT, PRESENTATION SKILLS
24	HANUMAN PRASAD	ASSISTANT DIRECTOR	ESTABLISHMENT RULES, RESERVATION IN SERVICE, HANDLING CAT/COURT CASES, ARTIFICIAL INTELLIGENCE, CYBER SECURITY, E-GOVERNANCE

Must read books, recommended by ISTM Library

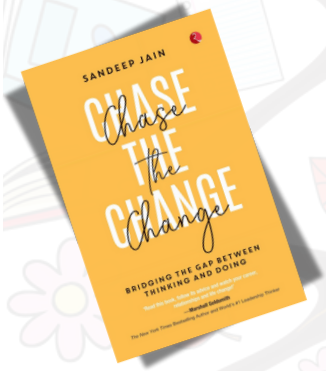
Title: Body on the Barricades
Author(s): Prakash, Brahma;
Publisher: LeftWord Books
Format: Print; Binding: Hardbound; Pages: 210;
Year of Publication: 2023

Explores the relationship between caste, resistance, and contemporary Indian theatre, highlighting art as a form of social protest.

Title: Power, Sexuality & Gender Dynamics At Work
Author(s): Sen, Roop and Chatterjee Uma;
Publisher: Sage Publication
Format: Print; Binding: Paperback; Pages: 248;
Year of Publication: 2020



Promotes equality and inclusivity, fostering a healthier work environment.



Title: Chase the Change: Bridging the Gap Between Thinking and Doing
Author(s): Jain, Sandeep;
Publisher: Rupa Publication; India;
Format: Print; Binding: Paperback; Pages: 232;
Year of Publication: 2024

Offers strategies to turn ideas into action and achieve personal and professional goals.

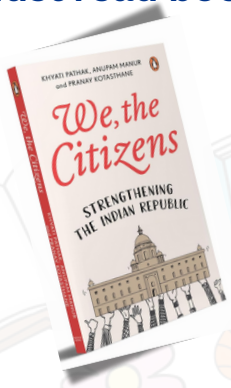
Title: Age of the Imperfect Leader: Leading from Strength
Author(s): Verma, Pawan;
Publisher: Rupa Publication; India;
Format: Print; Binding: Paperback; Pages: 252;
Year of Publication: 2019

Discusses the qualities of effective leadership in a changing world, emphasizing authenticity.

Title: Win Your Inner Battles
Author(s): Foroux, Darius;
Publisher: North Eagle Publishing;
Format: Print; Binding: Paperback; Pages: 137;
Year of Publication: 2015

Offers strategies for overcoming personal challenges and achieving self goals.

Must read books, recommended by ISTM Library

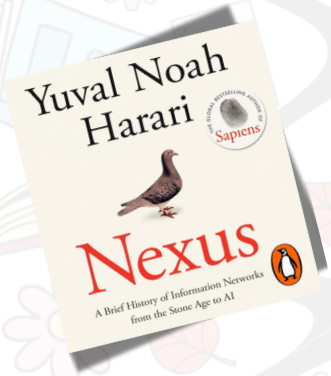


Title: We, The Citizens : Strengthening the Indian Republic
Author(s): Manur, Anupam and Kotasthane, Pranay; Pathak, Khayati;
Publisher: Penguine, India;
Format: Print; Binding: Hardbound; Pages: 240;
Year of Publication: 1978

Discusses the role of citizens in strengthening the democratic framework of India.

Title: The Science of Happiness
Author(s): Klein, Stefan;
Publisher: Speaking Tiger;
Format: Print; Binding: Paperback; Pages: 312;
Year of Publication: 2021

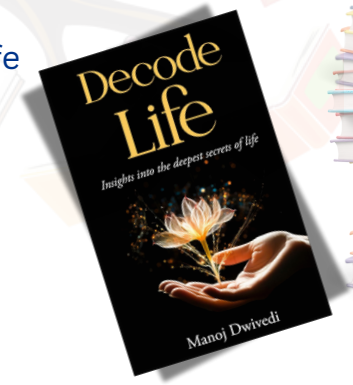
Explore the science behind happiness, examining how the brain and behavior influence well-being.



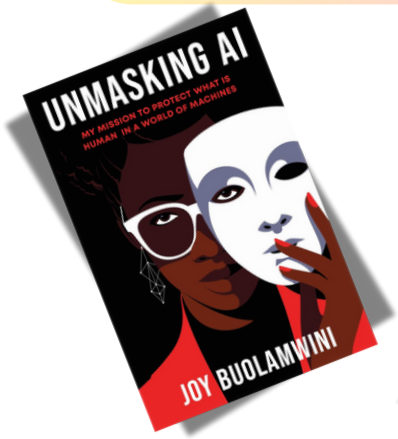
Title: Nexus
Author(s): Harari, Yuval Noah;
Publisher: Fern Press;
Format: Print; Binding: Paperback; Pages: 528;
Year of Publication: 2024

A philosophical novel about following one's dreams and listening to one's heart to find personal fulfillment and destiny.

Title: Decode Life - Insights into The Deepest Secrets of Life
Author(s): Dwivedi, Manoj;
Publisher: White Falcon Publishing;
Format: Print; Binding: Paperback; Pages: 144;
Year of Publication: 2024



Step-by-step guide to deciphering the code of life and unlocking true potential that each life stores in it.



Title: Unmasking AI
Author(s): Buolamwini, Joy;
Publisher: Random House Inc;
Format: Print; Binding: Paperback; Pages: 308;
Year of Publication: 2023

An examination of the biases embedded in artificial intelligence technologies, and the formation of the Algorithmic Justice League, focusing on the social impacts and ethical considerations of AI.