Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Secure AI: A Comprehensive Review on Security and Privacy Challenges and the Potential of Decentralized Approaches

Aryender Singh (member, IEEE)1, Deepak Chhabra2 , Senthilvelpalani Balavignesh3, Manu Rathee 4, Ashish Kaushik5, and Sarthak Singh Tomar6

1Salesforce, San Francisco, CA, USA (e-mail: aryendersingh2109@gmail.com)
2University Institute of Engineering & Technology, Maharshi Dayanand University, Rohtak, Haryana, India (e-mail: deepak.chhabra@mdurohtak.ac.in)
3 Prosthodontics Department, Post Graduate Institute of Dental Sciences, Pandit Bhagwat Dayal Sharma University of Health Sciences, Rohtak, Haryana, India (e-mail: balavignesh151199@gmail.com).
4 Prosthodontics Department, Post Graduate Institute of Dental Sciences, Pandit Bhagwat Dayal Sharma University of Health Sciences, Rohtak, Haryana, India (e-mail: ratheemanu@gmail.com).
5 Mechanical Engineering Department, Shree Guru Gobind Singh Tricentenary University, Gurgaon, Haryana, India. (e-mail: ashishkaushik6789@gmail.com).
6 Prosthodontics Department, Post Graduate Institute of Dental Sciences, Pandit Bhagwat Dayal Sharma University of Health Sciences, Rohtak, Haryana, India (e-mail: sarthakjay560@gmail.com).

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Background: Artificial Intelligence (AI) is transforming diverse sectors, yet its rapid deployment brings significant concerns regarding data security and user privacy. Centralized AI architectures, which rely on aggregating data into singular locations, are inherently vulnerable to breaches, surveillance, and manipulation. These limitations call for a shift toward more secure, privacy-preserving alternatives.

Objective: This review explores the major security and privacy challenges inherent in current AI systems and evaluates the potential of decentralized approaches, such as Federated Learning, Decentralized AI, and Zero-Knowledge Proofs, to address these limitations effectively.

Scope: The paper surveys key AI vulnerabilities, including data leakage, adversarial attacks, and trust issues in centralized models. It then discusses how decentralized architectures can improve resilience, confidentiality, and trustworthiness in AI systems. Special attention is given to privacy-preserving computation, distributed trust models, and cryptographic enhancements.

Findings: Decentralized AI approaches show strong potential to mitigate risks without compromising performance. Federated Learning enables collaborative model training without raw data sharing. Decentralized AI systems, often built on blockchain or peer-to-peer networks, eliminate single points of failure. Zero-Knowledge Proofs and similar cryptographic tools add verifiability without compromising sensitive data. However, practical adoption faces obstacles in terms of communication overhead, scalability, and interoperability.

Conclusion: Decentralized approaches represent a foundational shift in designing secure AI systems. While they are not a panacea, they provide powerful tools to build AI architectures that are robust, transparent, and privacy-aware. Future work should focus on enhancing scalability, standardizing protocols, and integrating multiple privacy-preserving techniques for real-world deployment

Keywords: AI security, privacy-preserving AI, Federated Learning, Decentralized AI, Zero-Knowledge Proofs, cryptographic AI, adversarial defense, secure computation.

INTRODUCTION

Artificial Intelligence (AI) has rapidly evolved into a transformative force, reshaping nearly every industry from healthcare and finance to education, business, and engineering. Its capacity to enhance efficiency, automate complex processes, and support decision-making has revolutionized how technology interacts with human needs. Generative AI has emerged as a groundbreaking development within the expansive domain of AI. Distinct from traditional AI

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

systems that primarily analyze, interpret, and classify data, Generative AI introduces the remarkable ability to create entirely new content—text, images, music, and even code based on learned patterns from massive datasets. This generative capacity has not only widened the functional scope of AI but has also unlocked new possibilities for creativity, personalization, and human-AI collaboration. Generative AI models such as GPT (Generative Pre-trained Transformer), BERT, and image-generation tools like DALL·E, Midjourney, and Stable Diffusion, have demonstrated an extraordinary ability to produce human-like outputs. Trained on vast and diverse corpora, these models can generate coherent written narratives, photorealistic images, original compositions, and operational code. As a result, tools like ChatGPT, GitHub Copilot, and Alpha Code have become indispensable across domains like content creation, education, customer service, software development, and even medical diagnostics. The integration of Generative AI into everyday workflows has made it a digital co-pilot for millions, assisting with tasks that range from drafting professional correspondence to aiding in complex clinical decision-making. In healthcare, for instance, Generative AI is already contributing significantly by providing personalized treatment suggestions, predicting disease progression, and automating routine documentation—all of which improve efficiency and reduce human error. The widespread proliferation of Generative AI has been further accelerated by the ubiquity of smartphones, social media platforms, and cloud computing. People now harness AI-powered tools to generate stylized artwork, compose music, write essays, and even create hyper-realistic deepfake videos. These applications highlight the immense creative potential of Generative AI, while also surfacing profound concerns regarding privacy, security, and ethical use. Users often interact with these tools by submitting highly personal data—photos, voice recordings, financial details, or health records—without fully understanding how that data is stored, processed, or possibly reused. For example, when a user uploads personal photos into AI art generators to produce stylized images, those inputs may be retained and used to further train the model, potentially without their knowledge. Similarly, conversational data handled by chatbots or virtual assistants might be stored in centralized databases that are vulnerable to breaches or misuse. In clinical settings, where the data involves sensitive health records, the stakes are even higher, raising critical questions about how to preserve patient confidentiality while still leveraging AI for improved outcomes. Despite its immense benefits, the adoption of Generative AI is accompanied by a range of pressing privacy and security risks. One major concern is the collection and use of data, much of which may contain personally identifiable information (PII). These datasets, often scraped from publicly available sources or user inputs, are not always collected with informed consent, raising ethical red flags. Moreover, certain technical threats like inference attacks where adversaries can deduce sensitive training data from model outputs further complicate the privacy landscape. Users are frequently left in the dark about how their data is being utilized, stored, or repurposed, fostering a lack of transparency and control. On the security front, Generative AI systems are increasingly vulnerable to various forms of cyberattacks. Prompt injection attacks, for instance, involve crafting malicious inputs that manipulate the AI's response, potentially leading to the leakage of confidential information. Similarly, model poisoning attacks target the training process itself, corrupting datasets in ways that bias or destabilize outputs. Deepfakes represent another dimension of concern: AI-generated videos or audio recordings that convincingly mimic real individuals can be exploited for misinformation campaigns, identity theft, and other fraudulent activities. Beyond technical vulnerabilities, Generative AI raises significant ethical and legal dilemmas that current regulatory frameworks are struggling to keep pace with. AI models can inadvertently reinforce societal biases present in their training data, leading to discriminatory or unjust outputs. Furthermore, the question of intellectual property rights surrounding AI-generated content remains contentious who owns the output, the user, the developer, or the model itself? Perhaps most crucially, the issue of accountability remains unresolved. If an AI system generates harmful, offensive, or misleading content, it is unclear where responsibility lies: with the developers, the deployers, or the AI itself? In response to these growing challenges, a variety of privacy-preserving and security-enhancing technologies have been developed. Federated learning, for example, allows AI models to train on decentralized data sources without directly accessing raw inputs, thus preserving privacy at the source. Techniques like differential privacy introduce algorithmic noise to datasets, making it statistically improbable to identify individual users. Homomorphic encryption enables computations on encrypted data, allowing systems to function securely without ever exposing the underlying information. AI auditing and interpretability tools are also being implemented to monitor model behavior, identify biases, and detect anomalies or misuses.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

At the same time, regulatory and ethical frameworks are beginning to emerge globally. The European Union's General Data Protection Regulation (GDPR) has set a high standard for data privacy, influencing global practices. Ethical guidelines developed by organizations such as the IEEE and OpenAI advocate for transparency, fairness, and accountability in AI systems. Legislative proposals like the EU AI Act aim to classify AI systems by their risk level and impose stricter compliance requirements for high-risk applications, particularly those involving human rights or critical infrastructure. In this context, the objective of this review is to provide a comprehensive exploration of the privacy and security implications of Generative AI. It examines the major threats posed by these technologies, evaluates current technological solutions aimed at safeguarding personal data, and discusses the ethical and legal challenges facing researchers, developers, and policymakers. As Generative AI continues to evolve and integrate more deeply into daily life, addressing these challenges will be crucial to ensuring that its benefits are realized without compromising the fundamental rights and safety of individuals and societies. This review seeks to inform ongoing discussions and future directions for secure and responsible AI development, aiming to balance innovation with accountability in an increasingly AI-driven world.

AI-SPECIFIC CYBERSECURITY VULNERABILITIES

Alongside data misuse, AI systems are increasingly targeted by cybercriminals due to the rich information they hold. These systems are not only repositories of sensitive data but also potential gateways to larger databases and critical infrastructures. AI introduces a new class of cyber threats that traditional cybersecurity models are not fully equipped to manage. Prompt injection attacks, for example, involve feeding malicious inputs into AI chatbots or virtual assistants to manipulate responses and potentially extract confidential information. This type of vulnerability was exemplified in incidents where AI customer service bots disclosed personal user details after being prompted in specific ways. Model inversion attacks take this a step further by allowing adversaries to reconstruct training data from the outputs of AI models, potentially revealing individual user records, including images or sensitive textual content. Similarly, adversarial attacks—where subtle, intentional modifications to inputs cause AI to make incorrect or dangerous decisions highlight the fragility of AI systems under hostile conditions. These vulnerabilities are particularly concerning in high-stakes environments such as healthcare, finance, and public safety. For instance, adversarial attacks on diagnostic AI models could result in misclassification of medical images, leading to incorrect treatments or delayed diagnoses. Moreover, the growing use of deepfakes and synthetic media generated by AI has added a layer of social and psychological vulnerability. Deepfakes can convincingly replicate an individual's appearance or voice, making them a potent tool for identity theft, misinformation, and social manipulation. In recent years, political figures, celebrities, and ordinary citizens alike have become targets of deepfake videos and audio clips that spread false narratives or damage reputations. The societal implications of such technologies, when misused, can be vast undermining public trust, promoting disinformation, and eroding democratic processes. Despite these risks, many AI systems are deployed with minimal oversight, limited transparency, and inadequate user protections, making them ripe for exploitation. Even unintentional security breaches, like the 2023 incident where ChatGPT users briefly saw each other's chat histories due to a technical flaw, reveal how vulnerable AI platforms can be when internal safeguards fail.

LEGAL SAFEGUARDS AND THE NEED FOR GLOBAL AI GOVERNANCE

To address the multifaceted privacy and security challenges posed by AI, regulatory bodies around the world have begun to implement frameworks aimed at protecting individuals' rights in the digital age. The European Union's General Data Protection Regulation (GDPR) remains the most comprehensive and influential legal standard, establishing strict rules for data handling, transparency, and user consent. Under GDPR, data must be collected for specific, legitimate purposes and cannot be repurposed without additional consent (purpose limitation). Only data necessary for those purposes should be collected (data minimization), and it must be deleted once it is no longer needed (storage limitation). Users also retain the right to access, correct, or delete their personal data and must be clearly informed about how their information is being used. These principles aim to give control back to individuals, ensuring that companies and AI developers operate with accountability and transparency. Other jurisdictions have followed suit, including the California Consumer Privacy Act (CCPA), which grants residents rights similar to those under GDPR. At the same time, forward-looking legislation such as the proposed EU AI Act seeks to address the unique risks posed by AI by classifying systems into risk categories and imposing stricter rules on high-risk

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

applications such as emotion recognition technologies and social scoring systems. These legal measures are vital steps toward creating a responsible AI landscape, but they also expose gaps in global governance. Many countries still lack comprehensive data protection laws, and cross-border data flows complicate enforcement efforts. Additionally, current regulations often struggle to keep pace with rapid technological advancement, leaving loopholes that can be exploited. To move forward, the development of harmonized international standards is essential. Stronger consent mechanisms, such as explicit opt-ins and informed disclosure policies, must become standard practice. AI developers should implement regular audits to detect and mitigate algorithmic bias, especially in critical areas like hiring, lending, and law enforcement. Robust cybersecurity protocols including encryption, adversarial testing, and anomaly detection are necessary to defend against both known and emerging threats. Most importantly, there must be a shift in culture from reactive regulation to proactive ethical design, where privacy and security are not afterthoughts but foundational pillars of AI development. This will require collaboration among governments, corporations, technologists, and civil society to create transparent, fair, and secure AI systems that uphold human dignity and autonomy. While AI continues to offer extraordinary benefits, such as personalized healthcare, efficient customer service, and enhanced educational tools, the risks associated with privacy violations and security breaches demand urgent attention. Without robust legal protections and ethical frameworks, the very technologies designed to improve lives could end up compromising the rights and safety of those they serve.

EUROPEAN UNION ARTIFICIAL INTELLIGENCE ACT

The European Union Artificial Intelligence Act (EU AI Act), officially enacted in June 2024, marks a pivotal development in global AI governance, constituting the first comprehensive legislative framework designed to regulate artificial intelligence across the European Union. Adopting a risk-based regulatory architecture, the Act classifies AI systems into four distinct categories based on their potential to harm fundamental rights, safety, and societal values: unacceptable risk, high risk, limited risk, and minimal risk. AI systems deemed to pose an unacceptable risk are strictly prohibited. These include applications involving manipulative subliminal techniques, government-implemented social scoring, untargeted scraping of facial images from the internet, and emotion recognition technologies deployed in workplace settings—except where medically or occupationally justified. High-risk AI systems, encompassing technologies used in sectors such as critical infrastructure, law enforcement, border control, and healthcare, are subject to rigorous regulatory oversight. These systems must undergo conformity assessments to ensure compliance with technical, ethical, and procedural safeguards. Requirements include robust data governance practices, transparency in system operation and intended use, documented risk management protocols, human oversight mechanisms, and continuous post-market monitoring to assess real-world performance and emergent risks. These obligations aim to prevent adverse impacts on public safety and individual rights while fostering trust in AI-enabled services. AI systems categorized as limited risk, such as AI-generated content platforms, virtual assistants, and interactive chatbots, are primarily subject to transparency requirements. Developers of such systems must clearly inform users when they are interacting with or receiving content generated by AI. Of particular significance is the Act's approach to general-purpose AI (GPAI) models. These models, capable of performing a wide range of tasks across multiple domains, are mandated to meet specific regulatory standards, including the provision of detailed technical documentation, disclosure of training data provenance to ensure compliance with intellectual property rights, and transparency measures related to model capabilities and limitations. For GPAI systems deemed to pose systemic risks defined in part by exceeding certain computational thresholds additional obligations are imposed. These include enhanced cybersecurity measures, mandatory risk assessments, and the implementation of risk mitigation strategies to prevent unintended or harmful outcomes. In recognition of the need to balance regulatory control with technological innovation, the EU AI Act includes provisions for regulatory sandboxes, which allow AI developers to test and refine systems in controlled environments under supervisory guidance. Furthermore, AI systems developed exclusively for research, development, or military applications are exempt from the scope of the Act, thereby preserving space for scientific progress and national defense imperatives. Enforcement of the Act is distributed among national competent authorities, coordinated by the newly established EU AI Office, which is responsible for ensuring consistent implementation across Member States. Non-compliance may result in substantial administrative fines, reaching up to 7% of an organization's global annual turnover, underscoring the EU's commitment to accountability in AI deployment. Despite its ambitious scope, the EU AI Act has not been without critique. Concerns have been raised regarding the potential financial burden imposed on small and medium-sized

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

enterprises (SMEs), the broad and evolving definition of AI within the legislation, and the reliance on self-assessment mechanisms for high-risk systems, which may undermine effective regulatory enforcement. Nevertheless, the Act is widely regarded as a foundational effort to align AI development with European values, including the protection of human dignity, democratic integrity, and social justice. Its phased implementation, scheduled over a 6 to 36-month timeline, is intended to allow stakeholders adequate time to adapt to new compliance requirements. The framework is also designed to be forward-looking, with anticipated amendments to address emerging technological and societal challenges, including the proliferation of generative AI and the environmental sustainability of large-scale AI systems. In essence, the EU AI Act represents a paradigm shift in the governance of artificial intelligence, seeking to operationalize ethical principles through enforceable legal standards. By integrating safeguards for fundamental rights with mechanisms that support innovation, the legislation aspires to position the European Union as a global leader in trustworthy AI.

GLOBAL REGULATION FOR DATA PROTECTION

Globally, data protection regulations have evolved differently depending on regional priorities, legal traditions, and political systems, yet there is a growing convergence in principles such as transparency, accountability, privacy, and fairness. In the United States, a fragmented, sector-specific approach is adopted toward data protection and AI regulation. While no comprehensive federal data privacy law exists, several targeted laws such as the Health Insurance Portability and Accountability Act (HIPAA) for healthcare data and the Children's Online Privacy Protection Act (COPPA) provide protection in specific domains. At the state level, California has led the way with the California Consumer Privacy Act (CCPA) and its successor, the California Privacy Rights Act (CPRA), which grant residents the rights to access, delete, and opt out of the sale of their personal data. These laws require businesses to disclose data practices and maintain transparency in data processing, significantly influencing data governance across the U.S. On AI governance, the U.S. relies on institutional guidance from bodies such as the Federal Trade Commission (FTC), which oversees consumer protection and emphasizes preventing algorithmic bias and ensuring algorithmic transparency. The National Institute of Standards and Technology (NIST) offers voluntary AI risk management frameworks, while the White House's AI Bill of Rights outlines citizens' rights in an AI-driven society, focusing on protection from discrimination, data misuse, and lack of accountability. Despite these efforts, the lack of a unified federal AI regulation presents a regulatory gap as AI adoption accelerates. Meanwhile, China adopts a centralized and security-oriented framework for both data and AI governance. Key laws include the Cybersecurity Law, Data Security Law, and Personal Information Protection Law (PIPL), which collectively establish stringent mandates for data localization, user consent, and governmental access. Additionally, China's Generative AI Regulation and AI Algorithm Regulation require AI developers to follow government-approved ethical standards and provide the state with oversight over recommendation and content-generation algorithms. These measures reflect China's dual commitment to AI leadership and national control, embedding data governance into its broader goals of social stability and geopolitical influence. In the United Kingdom, post-Brexit regulations continue to reflect principles of the EU's GDPR through the UK GDPR, preserving strong data protection while fostering a pro-innovation environment. The UK government is in the process of implementing an AI Governance Framework focused on transparency, accountability, fairness, and human oversight. The UK's approach seeks to balance robust ethical oversight with encouragement for business innovation, ensuring that AI development remains responsible and rights-oriented. Canada has taken a similarly structured approach, with the Personal Information Protection and Electronic Documents Act (PIPEDA) serving as the cornerstone of federal data privacy legislation. PIPEDA mandates meaningful consent, data security, and individual access to data while ensuring organizational accountability through the Office of the Privacy Commissioner. To address AI-specific challenges, Canada has proposed the Artificial Intelligence and Data Act (AIDA), which targets high-impact AI systems, promotes transparency, and mandates organizational accountability to prevent harm. This dual-structured approach reflects Canada's alignment with global trends in ensuring trustworthy AI systems while upholding citizens' data rights. Japan, another technologically advanced nation, promotes ethical AI through frameworks such as the AI R&D Guidelines and the AI Governance Principles, which emphasize human-centric AI development and transparency in automated decision-making. Japan encourages global cooperation, invests heavily in AI for sectors like robotics and healthcare, and prioritizes fairness and accountability in the deployment of intelligent systems. In the Asia-Pacific region, Australia enforces data privacy under the Privacy Act 1988, which regulates how personal information is collected, used, and stored by public and

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

private sector organizations. Recent amendments have introduced stricter rules for handling data breaches and reinforced the requirement for informed consent, bringing Australian privacy law closer to international standards. Additionally, Australia supports regional cooperation through the Asia-Pacific Economic Cooperation (APEC) Cross-Border Privacy Rules (CBPR) framework, which facilitates secure data flows among member economies while encouraging mutual recognition of privacy protections. Brazil has emerged as a leader in Latin America with its General Data Protection Law (LGPD), which reflects the principles of the GDPR by establishing legal bases for data processing, ensuring data subject rights, and creating a national data protection authority. The LGPD imposes accountability obligations on organizations and aligns Brazil with global best practices in data governance, enhancing its digital economy's credibility. India, on the other hand, is gradually shaping its data and AI landscape. The recently passed Digital Personal Data Protection Act provides a foundation for regulating digital privacy by enforcing data processing rules, requiring consent, and providing for cross-border data transfer protocols. Alongside this, existing frameworks such as the Information Technology Act (2000) and AI guidelines issued by NITI Aayog promote responsible innovation in areas like healthcare, education, and governance. India's approach reflects a strategic attempt to harness AI for socioeconomic development while safeguarding privacy and security. Across all these jurisdictions, several common principles guide both data and AI regulations. Transparency is a foundational element, requiring that organizations inform users about how their data is used and how automated systems make decisions. Accountability ensures that data controllers and AI developers are held responsible for the ethical and safe operation of their systems. Bias prevention is another central concern, especially in AI applications that influence critical decisions in areas like employment, healthcare, and criminal justice. Many countries now require fairness audits, bias testing, or human oversight to mitigate discriminatory outcomes. Data privacy remains a pillar of digital rights, enforced through consent requirements, breach notification obligations, and data minimization principles. Human oversight is increasingly mandated to prevent over-reliance on automated decision-making, ensuring that AI systems support rather than replace human judgment in sensitive areas. While there is a growing global consensus around these principles, several challenges persist. The absence of harmonized international standards creates legal uncertainty, especially for multinational companies operating across different legal regimes. Rapid technological advancements also outpace legislative developments, leading to regulatory gaps and inconsistent enforcement. In addition, resource constraints, limited technical expertise, and jurisdictional fragmentation hinder effective monitoring and compliance enforcement. Addressing these issues requires coordinated global efforts involving governments, industry leaders, academia, and civil society. Looking ahead, the future of data and AI governance will likely involve deeper international collaboration, stronger ethical guidelines, and more adaptive regulatory mechanisms. Emerging trends suggest increased focus on explainability, with mandates for interpretable AI models that allow users to understand and contest decisions. Regulatory sandboxes and public-private partnerships are also gaining traction as flexible tools to test new technologies under regulatory supervision. Additionally, ongoing discussions at forums such as the OECD, G7, and the United Nations are laying the groundwork for shared global standards on data protection and AI ethics. Ultimately, effective governance must strike a delicate balance between fostering innovation and upholding fundamental rights. As AI and data-driven technologies continue to evolve, governments and organizations must remain agile, ensuring that legal frameworks are responsive, inclusive, and capable of guiding ethical technology deployment across diverse social and cultural contexts.

CENTRALIZED ARTIFICIAL INTELLIGENCE (CEAI): DOMINANCE AND RISKS

Despite groundbreaking progress in Artificial Intelligence (AI), the field remains heavily reliant on centralized architectures—collectively known as Centralized Artificial Intelligence (CEAI). In CEAI, major technology corporations control the entire AI lifecycle: from data collection and model training to infrastructure and deployment. This centralization allows for high efficiency through access to vast computing power and integrated cloud platforms. However, it also introduces profound vulnerabilities and socio-technical issues. One key concern is monopolistic control. A handful of corporations dominate the AI landscape, accumulating immense datasets, proprietary algorithms, and infrastructural leverage. This dominance not only suppresses competition and innovation but also concentrates economic power and decision-making influence in a narrow segment of society. Additionally, centralized systems present significant security risks. Consolidating data and computation in central hubs creates attractive targets for cyberattacks. A single breach can jeopardize the privacy and safety of millions. Furthermore, the

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

central control model results in single points of failure—if one server or organization fails, entire AI systems can collapse.

Centralized AI is also prone to perpetuating biases. When AI systems are trained on data reflecting narrow worldviews—often shaped by their creators they risk reinforcing and amplifying societal inequities. Algorithmic bias can lead to discrimination in critical areas such as hiring, credit scoring, and law enforcement, especially when oversight is limited and transparency is lacking. From a governance standpoint, centralized entities wield disproportionate influence over ethical norms, public discourse, and technological priorities. Decisions are often driven by commercial incentives rather than public interest, with little accountability to end-users or affected communities. In regions with weak data protections or authoritarian regimes, the surveillance capabilities of CEAI can be weaponized to suppress dissent and infringe on civil liberties.

DECENTRALIZED ARTIFICIAL INTELLIGENCE (DEAI): A WEB3-INSPIRED SHIFT

Decentralized Artificial Intelligence (DEAI) emerges as a countermodel to CEAI, inspired by the principles of Web3 decentralization, transparency, and user sovereignty. DEAI envisions AI development distributed across diverse participants rather than controlled by a centralized authority. In such systems, data remains locally owned, and AI models are collaboratively trained and maintained by multiple independent actors. This approach enhances privacy by eliminating the need to centralize sensitive data, while also encouraging inclusivity by allowing a broader range of contributors to participate in model development. DEAI aligns with other decentralized innovations such as Decentralized Finance (DeFi), Decentralized Autonomous Organizations (DAOs), and peer-to-peer infrastructure, suggesting a broader movement toward more democratic digital ecosystem. Despite its potential, DEAI is still largely in the exploratory phase. Most systems are at the conceptual or prototype stage, lacking full-scale deployment or comprehensive integration across components. Addressing this gap requires not only technical innovation but also the creation of robust standards, governance models, and incentive mechanisms.

Federated Learning (FL): A Stepping Stone to Decentralization

Federated Learning (FL) provides an important foundation for DEAI by enabling collaborative model training without sharing raw data. In FL, decentralized clients (e.g., smartphones or edge devices) train local models and share only model updates with a central server for aggregation. This preserves user privacy, minimizes data movement, and can help meet regulatory requirements such as the GDPR.However, traditional FL still depends on a central aggregator, reintroducing a single point of failure. It also faces practical challenges, such as handling non-independent and identically distributed (non-IID) data, managing communication overhead, and protecting against malicious participants.

Blockchain-Enhanced Federated Learning (BCFL): Toward Full Decentralization

Blockchain-Enhanced Federated Learning (BCFL) aims to overcome these limitations by integrating blockchain technology with FL. Blockchain serves as a decentralized, immutable ledger that enables participants to validate and synchronize model updates without relying on a central server. This architecture eliminates centralized trust bottlenecks, improves resilience, and ensures that contributions are traceable and verifiable.In BCFL, updates from clients can be recorded on-chain, allowing for transparent aggregation and tamper-resistant logging. Smart contracts can enforce security policies, manage incentive mechanisms (e.g., token rewards), and verify the integrity of model contributions. This fosters trust and encourages honest participation, even among mutually untrusted parties.

Models of Coupling in BCFL

The integration of FL and blockchain varies in tightness, generally falling into three models:

Fully Coupled BCFL- Both FL and blockchain processes run on the same infrastructure. While this ensures maximal decentralization and reduces vulnerabilities, it demands high computational resources from each participating node.

Flexibly Coupled BCFL- FL and blockchain components operate on separate infrastructures. This design balances decentralization with performance, reducing latency and resource strain while preserving privacy and integrity.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Loosely Coupled BCFL- Blockchain plays a minimal role, such as verifying model updates or maintaining reputations. Although easier to implement, this model retains some centralization risks and lacks the robustness of fully decentralized systems.

Reputation-Based Client Selection

To improve model reliability and security, BCFL can incorporate reputation-based participant selection. Using blockchain, each client's past performance is transparently recorded and scored. These reputation scores based on direct contributions and peer evaluations can guide the inclusion or exclusion of participants. This mechanism strengthens the learning process by filtering out low-quality or malicious contributions and accelerates convergence by prioritizing trustworthy nodes. It also defends against attacks like Sybil or data poisoning, ensuring a more secure federated ecosystem.

Building Blocks of a DEAI Ecosystem

For DEAI to become viable at scale, a comprehensive ecosystem of interoperable components must be established. According to systematic analyses, key building blocks include:

Registry and Identity: Decentralized identity verification to ensure trusted participation.

Data and Ownership: Mechanisms for secure data control and provenance.

Training and Inference: Distributed computational resources for model development and execution.

Governance: Community-driven decision-making frameworks, potentially enabled via DAOs.

Marketplace and Incentivization: Platforms for exchanging data, models, and compute, powered by token-based rewards.

Reputation and Discoverability: Tools to evaluate and locate reliable contributors and datasets.

Cryptography and PrivacyTechniques like homomorphic encryption and zero-knowledge proofs to safeguard sensitive data.Although prototypes and proof-of-concept systems exist, most DEAI implementations lack full integration across these domains. Bridging these gaps requires interdisciplinary collaboration, standardization efforts, and rigorous validation through real-world deployments.

Zero-Knowledge Proofs (ZKPs): Privacy Without Exposure

Zero-Knowledge Proofs (ZKPs) are an advanced cryptographic technique that enables one party (the prover) to demonstrate to another (the verifier) that a specific statement is true without disclosing any additional information beyond the validity of the claim itself. This unique property makes ZKPs an essential building block for privacy-preserving technologies in Decentralized Artificial Intelligence (DEAI), where distributed collaboration must be grounded in trust and transparency, yet constrained by the need for confidentiality.

ZK-SNARKs: Compact and Efficient Privacy Layers

Among the most prominent implementations of ZKPs are ZK-SNARKs (Succinct Non-Interactive Arguments of Knowledge), which offer compact proof sizes and rapid verification without the need for interaction between prover and verifier. These characteristics make ZK-SNARKs particularly suitable for resource-constrained environments such as edge devices, mobile nodes, or blockchain-based smart contracts. In decentralized AI systems, ZK-SNARKs enable the verification of model integrity, data ownership, or user identity in a way that is cryptographically secure and computationally scalable.

Applications in Decentralized AI Systems

ZKPs have broad applications across the DEAI ecosystem, supporting both data sovereignty and computational integrity:

Privacy-Preserving Model Training- In federated and blockchain-enhanced federated learning environments, ZKPs can confirm that a local model was trained on valid data without revealing the data itself, ensuring that sensitive information (such as personal health records) never leaves the user's device.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Secure Identity Verification-Through ZKP-enabled decentralized identity systems, participants in an AI network can verify their credentials (e.g., being a certified medical practitioner) without disclosing their full identity or sensitive documents. This reduces identity theft risks while maintaining system trustworthiness.

Confidential Inference and Model Auditing- ZKPs allow AI services to perform inference on encrypted inputs or verify model behavior without exposing either the data or the internal workings of the model. This is especially important in regulated industries such as finance, healthcare, or law.

Decentralized Voting and Governance- In DAOs and community-driven AI governance frameworks, ZKPs can be used to build anonymous yet verifiable voting systems, ensuring fair participation without revealing individual choices.

Supply Chain and Data Provenance- In collaborative AI applications such as environmental monitoring or supply chain optimization, ZKPs can certify the authenticity and integrity of data contributions from multiple sources, even when some sources require anonymity.

Strengthening Trust in a Distributed Environment

By enabling verifiable computation and authenticated interactions without data leakage, ZKPs serve as a cornerstone of trustless collaboration in DEAI systems. Participants can confidently share computational outcomes or engage in data-driven workflows while preserving autonomy and privacy. This trust model is particularly crucial in settings where the participating entities—such as hospitals, research institutions, or international NGOs—may not fully trust one another or lack a centralized oversight mechanism.

REFRENCES

[1] C. Williams, “A Brief Introduction To Artificial Intelligence,” Proceedings OCEANS’83, 1983
[2] K. S. Kaswan, J. S. Dhatterwal, K. Malik and A. Baliyan, ”Generative AI: A Review on Models and Applications,” 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 699-704, doi: 10.1109/ICCSAI59793.2023.10421601.
[3] 3. A. Golda et al., “Privacy and Security Concerns in Generative AI: A Comprehensive Survey,” IEEE Access, vol. 12, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3381611.
[4] 4. Abebe Diro, Shahriar Kaisar, A. Saini, S. Fatima, Pham Cong Hiep, and F. Erba,“Workplace security and privacy implications in the GenAI age: A survey,” Journal of Information Security and Applications, vol. 89, pp. 103960–103960, Jan. 2025, doi:https://doi.org/10.1016/j.jisa.2024.103960.
[5] 5. J. Pitt, ”Deepfake Videos and DDoS Attacks (Deliberate Denial of Satire) [Editorial],” in IEEE Technology and Society Magazine, vol. 38, no. 4, pp. 5-8, Dec. 2019, doi:10.1109/MTS.2019.2948416.
[6] 6. M. Alloghani et al., “A systematic review on the status and progress of homomorphic encryption technologies,” Journal of Information Security and Applications, vol. 48,p. 102362, Oct. 2019, doi: https://doi.org/10.1016/j.jisa.2019.102362.
[7] 7. C. J. Hoofnagle, B. V. D. Sloot, and F. Z. Borgesius, “The European Uniongeneral data protection regulation: what it is and what it means,” Information & Communications Technology Law, vol. 28, no. 1, pp. 65–98, 2019, doi: https://doi.org/10.1080/13600834.2019.1573501.
[8] 8. S. Shahriar, S. Allana, S. M. Hazratifard and R. Dara, ”A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle,” in IEEE Access, vol. 11, pp. 61829-61854, 2023, doi: 10.1109/ACCESS.2023.3287195.
[9] 9. A. Boulemtafes, A. Derhab and Y. Challal, ”A review of privacy-preserving techniquesfor deep learning”, Neurocomputing, vol. 384, pp. 21-45, Apr. 2020.
[10] 10. M. ´ Angel and A. Meuwese, “Regulating AI from Europe: a joint analysis of the AI Act and the Framework Convention on AI,” The Theory and Practice of Legislation,pp. 1–20, Apr. 2025, doi: https://doi.org/10.1080/20508840.2025.2492524.
[11] 11. I. Kusche, “Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk,” Journal of Risk Research, pp. 1–14, May 2024, doi:https://doi.org/10.1080/13669877.2024.2350720.

[12] 12. D. Rezaeikhonakdar, "AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors," Journal of Law, Medicine & Ethics, vol. 51, no. 4, pp.988–995, Jan. 2023, doi: https://doi.org/10.1017/jme.2024.15.
[13] 13. B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," Government Information Quarterly, vol. 39, no. 4, p. 101685, Mar. 2022, doi: https://doi.org/10.1016/j.giq.2022.101685.
[14] 14. S. McLean, G. J. M. Read, J. Thompson, C. Baber, N. A. Stanton, and P. M. Salmon,"The risks associated with Artificial General Intelligence: A systematic review," Journal of Experimental & Theoretical Artificial Intelligence, vol. 35, no. 5, pp. 1–17, Aug. 2021, doi: https://doi.org/10.1080/0952813x.2021.1964003.
[15] 15. P. Slattery et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence," arXiv (Cornell University), Aug. 2024, doi: https://doi.org/10.48550/arxiv.2408.12622.
[16] 16. B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," Government Information Quarterly, vol.June 11, 2025 4 39, no. 4, p. 101685, Mar. 2022, doi: https://doi.org/10.1016/j.giq.2022.101685.
[17] 17. L. Cao, "Decentralized AI: Edge Intelligence and Smart Blockchain, Metaverse, Web3, and DeSci," in IEEE Intelligent Systems, vol. 37, no. 3, pp. 6-19, 1 May-June 2022, doi: 10.1109/MIS.2022.3181504.
[18] 18. Z. Wang et al., "SoK: Decentralized AI (DeAI)," arXiv (Cornell University), Nov.2024, doi: https://doi.org/10.48550/arxiv.2411.17461.
[19] 19. I. Varlamis et al., "Using Big Data and federated learning for generating energy efficiency recommendations", Int. J. Data Sci. Anal., pp. 1-17, 2022.
[20] Betul Yurdem, Murat Kuzlu, Mehmet Kemal Gullu, Ferhat Ozgur Catak, and M. Tabassum, "Federated Learning: Overview, Strategies, Applications, Tools and Future Directions," Heliyon vol. 10, no. 19, pp. e38137–e38137, Sep. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e38137.
[21] 21. K. Dasaradharami Reddy and T. R. Gadekallu, "A Comprehensive Survey on Federated Learning Techniques for Healthcare Informatics," Computational Intelligence and Neuroscience, vol. 2023, pp. 1–19, Mar. 2023, doi: https://doi.org/10.1155/2023/8393990.
[22] 22. W. Ning et al., "Blockchain-Based Federated Learning: A Survey and New Perspectives," Applied Sciences, vol. 14, no. 20, pp. 9459–9459, Oct. 2024, doi: ://doi.org/10.3390/app14209459.
[23] 23. A. Qammar, A. Karim, H. Ning, and J. Ding, "Securing federated learning with blockchain: a systematic literature review," Artificial Intelligence Review, Sep. 2022, doi:https://doi.org/10.1007/s10462-022-10271-9.
[24] 24. H. Zhang, S. Jiang, and S. Xuan, "Decentralized federated learning based on blockchain: Concepts, framework, and challenges," Computer Communications, Jan. 2024, doi: https://doi.org/10.1016/j.comcom.2023.12.042.
[25] 25. M. H. ur Rehman, K. Salah, E. Damiani, and D. Svetinovic, "Towards Blockchain Based Reputation-Aware Federated Learning," IEEE Xplore, Jul. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9163027
[26] 26. S. Panja and B. Roy, "A Secure End-to-End Verifiable E-Voting System Using Zero Knowledge Proof and Blockchain," Indian Statistical Institute Series, pp. 45–48, 2021, doi: https://doi.org/10.1007/978-981-33-6991-7 6.
[27] 27. X. Yang and W. Li, "A zero-knowledge-proof-based digital identity management scheme in blockchain," Computers & Security, vol. 99, p. 102050, Dec. 2020, doi: https://doi.org/10.1016/j.cose.2020.102050.
[28] 28. T. Chen, H. Lu, T. Kunpittaya, and A. Luo, "A Review of zk-SNARKs," arXiv (Cornell University), Jan. 2022, doi: https://doi.org/10.48550/arxiv.2202.06877.
[29] 29. D. A. Luong and J. H. Park, "Privacy-Preserving Identity Management System on Blockchain Using zk-SNARK," IEEE Access, pp. 1–1, 2023, doi:https://doi.org/10.1109/access.2022.3233828.
[30] 30. Anatoly Konkin and Sergey Zapechnikov, "Zero knowledge proof and ZK-SNARK for private blockchains," Journal of Computer Virology and Hacking Techniques, vol. 19, no. 3, pp. 443–449, Mar. 2023, doi: https://doi.org/10.1007/s11416-023-00466-1