Volume 02 | Issue 01 | March 2026 | Quarterly

CyberPeace

CYBERPEACE

MAGAZINE

BUILDING A SAFER DIGITAL FUTURE FOR ALL HUMANITY

TECHNOLOGY
POLICY
GOVERNANCE
DIPLOMACY

AGENTIC AI

- Autonomy,
- Control,
- Accountability

https://magazine.cyberpeace.click

Emerging technologies, digital laws and governance, trust and safety, cyber diplomacy

Volume 02 | Issue 01 | March 2026 | Quarterly

CYBERPEACE MAGAZINE

OUR SUPPORTERS

SHAH & ANCHOR

CyberPeace
CENTER OF EXCELLENCE

Army Institute of Technology, Pune
Onward to Glory

PARTNERS

Autobot Infosec

SysTools®
Simplifying Technology

Cyber 4N6 Consultants

CYBERPEACE MAGAZINE

This is a digital publication. An ISSN has been applied for.

The views and opinions expressed in this magazine are those of the respective authors and do not necessarily reflect the official policy or position of CyberPeace Foundation.

While every effort has been made to ensure the accuracy of the information presented, the publisher makes no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, or suitability of the content.

CyberPeace Magazine was formerly published as CyberPeace and is being rebranded to reflect an expanded scope and global focus.

Published by:

CyberPeace Foundation

For inquiries, collaborations, or permissions:

secretariat@cyberpeace.net

Website:

www.cyberpeace.org

OUR TEAM

LEADERSHIP. VISION. COMMITMENT.

EDITOR-IN-CHIEF

Maj Vineet Kumar
Founder and Global President
CyberPeace

EDITORIAL BOARD

Lt Gen (Dr.) Rajesh Pant (Retd)
Global Advisor CyberPeace, PSM, AVSM, VSM
Former National Cybersecurity Coordinator

Shri SN Pradhan
IPS (Retd.), Global CEO, CyberPeace
Former Director General, Narcotics Control Bureau

Dr. Bhaskar Chatterjee
Former IAS, Global Advisory Council, CyberPeace

Prof. (Dr.) Manish Prateek
Professor and Pro Vice-Chancellor, DBS Global University

Prof. (Dr.) Scott Shackelford
Associate Vice President & Vice Chancelloror for Research, Indiana University Bloomington, (IUB)

OUR TEAM

LEADERSHIP. VISION. COMMITMENT.

TECHNICAL COMMITTEE

Mr. MAKP Singh

Chief Technology Officer,
CyberPeace

Capt. Suresh Chandra Joshi

Sr Researcher, Innovation and Research,
CyberPeace

Dr. Nilakshi Jain

Head of Department, Shah & Anchor
Kutchhi Engineering College, Mumbai

EDITORIAL COORDINATOR

Ms Ayndri

Research Analyst, Policy & Advocacy,
CyberPeace

EDITORIAL TEAM

Ms Hithika Kohli

Consultant, Policy & Advocacy,
CyberPeace

Ms Anupma Prakash

Consultant, Policy & CPC,
CyberPeace

DESIGN & PRODUCTION

Mr Ajeet Kumar

Sr Graphic Designer, CyberPeace

TABLE OF CONTENTS

TABLE OF CONTENTS

EDITOR'S NOTE

Maj Vineet Kumar

Founder and Global President

CyberPeace

Welcome to the April 2026 edition of CyberPeace Magazine. India is currently experiencing one of its most transformative phases in technological growth. The recent AI Impact Summit held in Delhi is a shining example of the country's proactive approach to harnessing the potential of Artificial Intelligence (AI) to achieve the vision of "Welfare for All, Happiness for All." The next phase of AI, in the form of agent-led systems, is reshaping the technological landscape by driving revolutionary changes across critical industries such as healthcare, finance, and software development. Andrew Ng, a British-American computer scientist, is widely credited with popularising the concept of Agentic AI, shifting the focus of AI from a passive tool to a more autonomous, reasoning-based system.

This issue has been thoughtfully curated with the aim of educating readers, encouraging dialogue, and raising awareness to prepare for the third wave of AI evolution. The articles featured in this edition explore major Agentic AI trends in 2026, offering insights into the anticipated expansion of AI technologies across multiple sectors, the risks associated with them, and the evolving regulatory framework.

As AI becomes more human-like through Agentic systems, CyberPeace remains committed to promoting cybersecurity awareness and safeguarding digital spaces. This issue also examines the factors driving rural youth in India's hinterlands toward cybercrime. It highlights not only the root causes behind such financial and cyber frauds but also emphasises the importance of rehabilitating young individuals through a compassionate and empathetic approach.

Agentic AI is here to stay and is expected to make significant inroads into all aspects of life. Its impact on the professional landscape is still unfolding. Evaluating its success will depend on whether efficiency gains outweigh the emerging cybersecurity risks. Given the autonomous nature of these systems, accountability remains a critical concern, particularly in cases of faulty decision-making or algorithmic errors. These challenges present ethical and legal barriers to the fair and effective implementation of Agentic AI.

The full impact of Agentic AI is yet to be realised, but the future certainly looks promising. At CyberPeace, we remain vigilant and committed to addressing the evolving technological landscape, along with the risks and challenges that accompany it.

THE BRIEF

CYBERPEACE PULSE

Agentic AI doesn't just respond; it decides, plans, and acts. For example, a smart travel assistant could automatically plan a trip by noticing your long weekend, checking your preferences, comparing prices, and booking within your budget. It anticipates your needs while still responding to your commands. It works by securely using your personal data like your calendar and past choices to make informed, autonomous decisions on your behalf.

This section provides a curated snapshot of global developments across technology, security, policy, and cyber diplomacy shaping the rise of autonomous AI systems.

GLOBAL SIGNALS ON AGENTIC AI

Autonomous AI agents capable of reasoning, planning, and executing actions are rapidly moving from research labs into real-world systems. Governments, security researchers, and technology companies are now grappling with the governance, cybersecurity, and strategic implications of these emerging systems.

1. Technology and industry

✦ Samsung Explores Agentic AI Manufacturing

Samsung is reportedly developing AI-driven factories where autonomous agents coordinate production processes, manage logistics, and optimise industrial operations with minimal human intervention. The initiative reflects a broader push toward self-managing industrial systems powered by autonomous AI decision-making.

https://www.techbuzz.ai/articles/samsung-bets-on-agentic-ai-to-transform-global-factories-by-2030

AI AGENTS ENTER DIGITAL COMMERCE

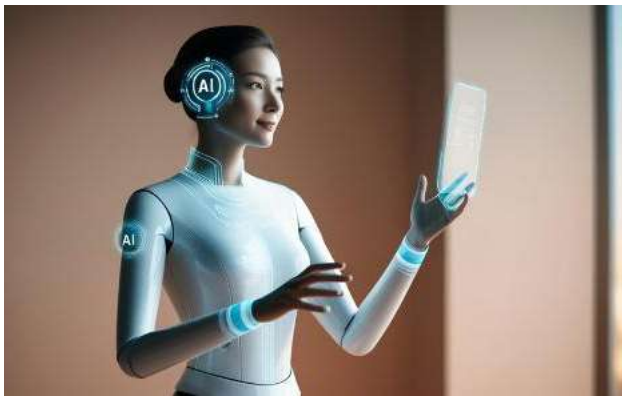
Autonomous AI agents are beginning to independently conduct online purchases and financial transactions on behalf of users and companies. Experts warn that this could create **new legal challenges around liability, identity verification, and consumer protection.**

https://www.digitalcommerce360.com/2026/02/24/why-ai-agents-are-reaching-checkout-without-clear-rules/

Enterprises Deploy AI "Digital Workers"

Businesses are increasingly deploying autonomous agents capable of planning workflows, interacting with software systems, and coordinating with other agents. Some analysts predict that **AI agents may soon function as "digital employees" within enterprise environments.**

https://www.computerworld.com/article/3843138/agentic-ai-ongoing-coverage-of-its-impact-on-the-enterprise.html



2. SECURITY AND CYBER RISK

Autonomous AI Exploit Risks Emerge

Cybersecurity researchers have raised concerns about experimental autonomous agents capable of identifying vulnerabilities and executing cyber operations independently. Such systems could significantly accelerate the speed and scale of cyberattacks if weaponised.

https://fortune.com/2026/02/12/openclaw-ai-agents-security-risks-beware/

OWASP Releases Top 10 Risks for Agentic Systems

Security researchers have identified new vulnerabilities unique to autonomous AI agents, including agent impersonation, uncontrolled tool access, and multi-agent manipulation attacks.

https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/\

3. POLICY AND GOVERNANCE

Global Push for Agentic AI Standards

Technology firms and policy institutions are working toward interoperability standards for AI agents. Experts argue that common governance frameworks will be essential as autonomous systems begin interacting across platforms, services, and national jurisdictions.

https://www.pymnts.com/artificial-intelligence-2/2026/openai-and-cisco-back-standards-to-scale-agentic-ai/

Industry Divided on AI Safety

A widening debate has emerged within the AI industry between proponents of rapid technological deployment and advocates calling for stronger regulatory oversight and safety guardrails for advanced AI systems.

https://www.vox.com/politics/481229/anthropic-pentagon-openai-amodei-ai

4. CYBER DIPLOMACY AND STRATEGIC SECURITY

Autonomous Cyber Capabilities in National Security

Defence institutions are exploring AI systems capable of autonomously identifying cyber vulnerabilities, conducting network reconnaissance, and supporting defensive cyber operations.

🔗 https://www.ft.com/content/a56d70b5-669c-4bcc-8541-a4961fc99802

Agentic AI and Global Technology Competition

Autonomous AI systems are increasingly viewed as a strategic capability within global technology competition. Governments are investing heavily in AI research, talent development, and computing infrastructure as countries seek advantages in economic competitiveness, national security, and technological sovereignty.

https://www.oecd.org/en/publications/competition-in-artificial-intelligence-infrastructure_623d1874-en

THE SHIFT

TOP AGENTIC AI TRENDS FOR 2026

In 2026, AI is evolving from a helpful assistant into an autonomous digital colleague. Here are the defining trends reshaping the enterprise landscape:

- From "Copilots" to Autonomous Coworkers: Enterprises are moving beyond chat-based assistants to goal-driven agents that "own" outcomes rather than just "helping with" tasks. Gartner predicts that 40% of enterprise applications will embed task-specific AI agents by the end of 2026.

- Multi-Agent Orchestration (The "AI Team"): Single agents are being replaced by networks of coordinating autonomous agents. Specialised agents (e.g., a researcher, a coder, and an analyst) collaborate autonomously to solve complex, cross-functional workflows.

- The "Protocol Revolution": New standards like Anthropic's Model Context Protocol (MCP) are becoming the "USB-C of AI," allowing agents to plug-and-play across different tools and databases without custom integrations.

- Governance-by-Design: As agents gain autonomy, governance is being coded directly into their DNA. Forrester predicts that 50% of Enterprise Resource Planning (ERP) vendors will launch autonomous governance modules in 2026 to provide real-time compliance and audit trails.

- Domain-Specific Small Language Models (SLMs): Organisations are shifting toward smaller, specialised models trained on proprietary data for high-frequency, narrow tasks (like claims triage or inventory management) to reduce costs and latency.

- Physical AI Integration: Agentic AI is moving from screens to the physical world, coordinating robots and sensors in warehouses and factories for intelligence that responds to environmental changes in real-time

Reality Check: Despite the hype, Deloitte notes that while 38% of organisations are piloting agents, only 11% have them in full production due to legacy infrastructure bottlenecks.

THE REGULATORY LANDSCAPE

AGENTIC AI REGULATORY TRACKER

While most current frameworks regulate AI broadly, several emerging initiatives directly address risks associated with highly autonomous or agentic systems.

POLICY INSTRUMENT:

EU Artificial Intelligence Act

STATUS:

Adopted in 2024; phased implementation through 2026–2027

RELEVANCE TO AGENTIC AI

- Introduces **risk-based regulation** of AI systems
- Places strict requirements on **high-risk AI systems**
- Requires **transparency obligations** for certain AI models

IMPLICATION

Under the EU Artificial Intelligence Act, AI systems are classified based on **risk to safety, fundamental rights, and societal impact.**

While the Act does not explicitly use the term 'agentic AI', highly autonomous systems deployed in certain domains would likely fall into the 'high-risk AI systems' category if they operate in regulated sectors listed in the legislation.

EU AI ACT

Highly autonomous AI systems operating in regulated sectors may be treated as high-risk under the EU AI Act.

https://artificialintelligenceact.eu

Highly autonomous AI agents deployed in critical usages such as lending decisions, credit risk evaluations, or fraud detection actions without human approval, autonomous recommendation of treatment plans or triages patients in hospitals, automated surveillance analysis or predictive policing decisions, autonomous ranking of job applicants, etc. could fall under **high-risk system classifications**, requiring risk management, human oversight, and cybersecurity safeguards.

UNITED KINGDOM

A PRO-INNOVATION APPROACH TO AGENTIC AI REGULATION

POLICY INSTRUMENT

Pro-Innovation Approach to AI Regulation (White Paper), and AI Safety Institute

STATUS

White Paper published 2023;
AI Safety Institute established 2023;
AI Action Plan released 2025

RELEVANCE TO AGENTIC AI

- Does not propose a single binding AI law, instead uses existing regulators (healthcare, finance, transport) to oversee AI in their sectors
- AI Safety Institute specifically focuses on evaluating frontier and autonomous AI models
- Emphasis on voluntary commitments from AI developers

IMPLICATION

The UK's sector-by-sector approach means agentic AI regulation depends entirely on which domain the system operates in. An autonomous AI agent in healthcare would face National Health Survey (NHS) and Medicines and Healthcare products Regulatory Agency (MHRA) oversight, while one in financial services would face Financial Conduct Authority (FCA) scrutiny. This creates inconsistencies in agentic systems operating across multiple sectors simultaneously and may fall through regulatory gaps. The UK's AI Safety Institute is notable as one of the first government bodies specifically tasked with evaluating advanced autonomous AI capabilities.

KEY RESOURCES

- https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
- https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

UNITED STATES OF AMERICA

POLICY INSTRUMENT:

Phase 1: Executive Order (EO) 14110: Safe, Secure and Trustworthy AI (Biden, 2023). Rolled back

Phase 2: SEO 14179 and the "Winning the Race" AI Action Plan (Trump, 2025)

Status: Regulatory burdens revoked and safety framework dismantled; innovation-first approach now governs AI.

RELEVANCE TO AGENTIC AI

- EO 14110 addressed autonomous systems, frontier AI safety, and critical infrastructure risks directly relevant to agentic AI governance
- EO 14179 framework contains no specific provisions for agentic AI risks, human oversight, or autonomous system safety
- State-level attempts being made to fill the regulatory gap

IMPLICATION

The United States has made the biggest U-turn in AI governance of any country in the world. The EO 14179 framework explicitly focuses on removing safety barriers, meaning the world's largest AI producer now operates with virtually no enforceable agentic AI governance. The 2025 AI Action Plan's 90 recommendations contain no specific provisions for autonomous AI oversight, human-in-the-loop requirements, or agentic AI risk management creating a stark and dangerous contrast with the EU AI Act and Singapore's dedicated Agentic AI Governance Framework.

https://www whitehouse archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

https://www whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-americain-leadership-in-artificial-intelligence/

https://www whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

CANADA

AGENTIC AI REGULATION LANDSCAPE

POLICY INSTRUMENT

Artificial Intelligence and Data Act (AIDA) – Part of Bill C-27

STATUS

Introduced 2022 but Lapsed.
Bill C-27 died January 2025 after Parliament progroged; AIDA officially abandoned June 2025; no replacement federal AI law currently in force

RELEVANCE TO AGENTIC AI

- Would have been regulated "high-impact AI" systems" covering autonomous decisions in healthcare, finance, and employment
- Would have required impact assessments, human oversight, and incident reporting
- Would have established an AI and Data Commissioner for enforcement
- Never passed into law

IMPLICATION

Canada's AIDA represents one of the most significant AI governance failures globally. Introduced in 2022 with ambition to regulate high-impact AI systems which would have captured many agentic deployments, it collapsed due to lack of prior consultation, vague drafting, and other factors. Canada now has no federal AI law, leaving businesses to follow Quebec's provincial rules, existing privacy laws, voluntary codes, emerging sector-specific guidelines or international frameworks like the EU GDPR as a substitute baseline. A future AI law is expected but its timeline and shape remain uncertain. This leaves Canada significantly behind the EU and even Singapore in agentic AI governance, a gap for one of the G7's most AI-active economies.

https://ised-isde canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

SINGAPORE

LEADING THE WAY IN AGENTIC AI GOVERNANCE

POLICY INSTRUMENT:
Model AI Governance Framework and AI Verify Toolkit

STATUS:
Model Framework published 2026; AI Verify launched 2022; Project Moonshot (agentic AI safety) launched 2023

RELEVANCE TO AGENTIC AI

- One of the first countries to develop **testing tools** specifically for agentic AI safety
- Project Moonshot provides **benchmarks** for evaluating autonomous AI system behaviour
- Model Framework emphasises **human oversight and accountability** in AI decision-making

IMPLICATION

Singapore's approach is non-binding but technically sophisticated. Rather than imposing legislation, Singapore positions itself as a **global testing and standards hub** for AI governance. Project Moonshot is particularly notable as it directly addresses agentic AI risks evaluating whether autonomous systems behave safely across different scenarios. Singapore's frameworks are frequently adopted voluntarily by companies operating in Southeast Asia and serve as a reference point for other developing nations building their own AI governance approaches.

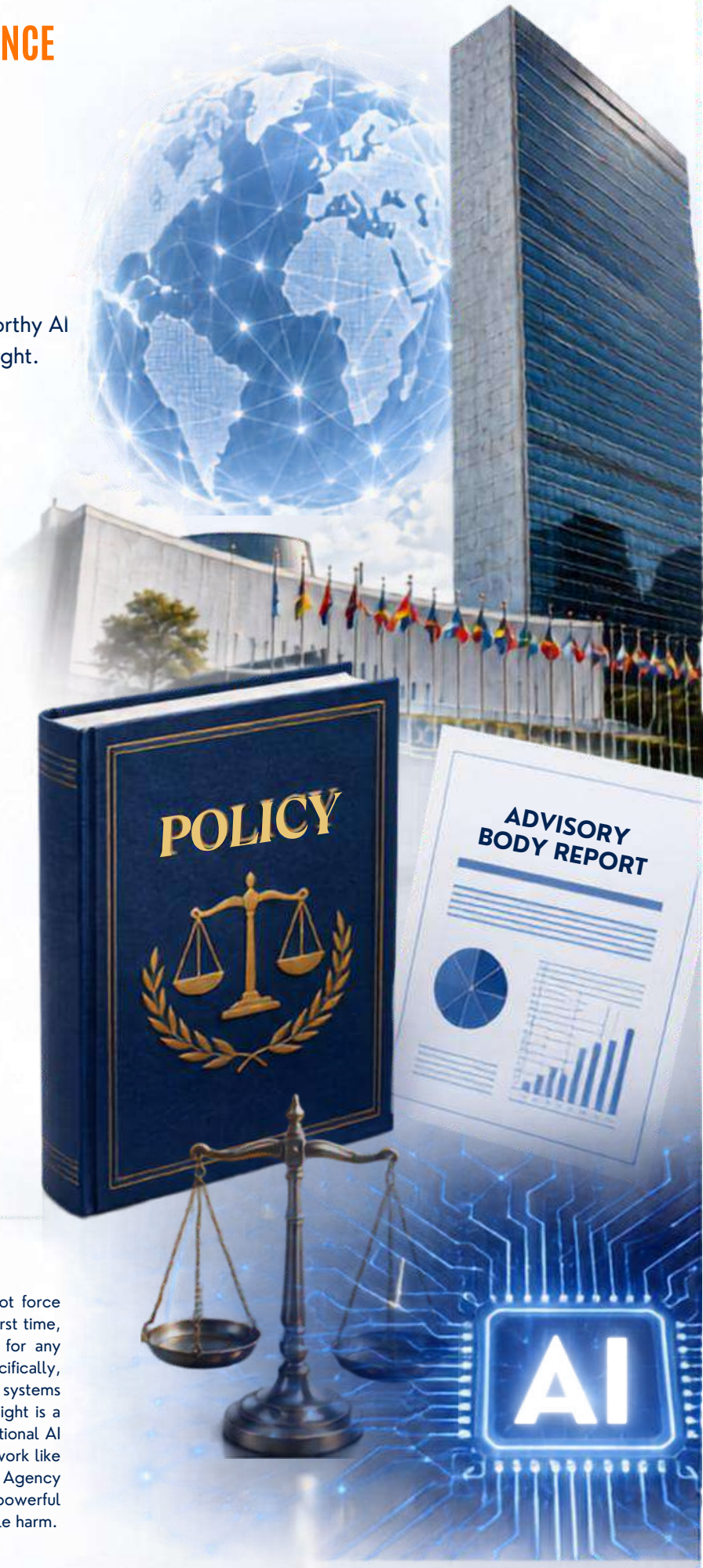
https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model AI Governance Framework for Generative AI-May 2024-1-1.pdf

https://www.mnda.gov/resource/press-releases-factsheets-and-speeches/press-releases/2024/sg-launches-project-moonshot

STRENGTHENING GLOBAL GOVERNANCE FOR A TRUSTWORTHY AI FUTURE

UNITED NATIONS GENERAL ASSEMBLY

A landmark step toward safe, secure, and trustworthy AI through global cooperation and high-level oversight.



POLICY INSTRUMENT:

UN Resolution on Safe, Secure and Trustworthy AI and High-Level Advisory Body on AI

STATUS:

Resolution adopted by General Assembly March 2024; High-Level Advisory Body report published 2024

RELEVANCE TO AGENTIC AI

- First UN resolution specifically addressing AI governance
- Calls for international cooperation on AI safety standards
- Advisory Body recommends a multi-stakeholder international AI governance panel
- Highlights risks of autonomous systems in conflict and critical infrastructure

IMPLICATION

The UN resolution is non-binding, meaning it cannot force any country to act. But it matters because for the first time, every nation agreed on one thing: AI is too big for any single country to handle alone. For agentic AI specifically, the UN's Advisory Body warned that autonomous systems crossing national borders without any shared oversight is a serious and growing danger. The idea of an international AI governance body if it ever becomes reality could work like the nuclear watchdog International Atomic Energy Agency (IAEA), keeping a global eye on the most powerful autonomous AI systems before they cause irreversible harm.

RESOURCES

- https://documents.un.org/doc/undo-gen/n24/08783/pdf/n2408783.pdf
- https://www.un.org/sites/un2.un.org/files/governing_ai for humanity_final_report_en.pdf

More autonomous AI requires more human oversight

The Human Layer:

Trust, Safety, and Agentic AI

Agentic AI is reshaping the relationship between humans and technology.

The future will depend not only on how well we secure machines, but on how effectively we protect people, ensure accountability and preserve trust in a world increasingly mediated by technological breakthroughs.

01 Decision Complexity

AI decisions involve multi-step reasoning, making them hard to understand.

02 Lack of Transparency

The internal processes of AI are often hidden from users.

03 Accountability Issues

It's challenging to determine who is responsible when AI makes errors.

04 Verification Challenges

Users struggle to confirm the accuracy and reliability of AI outcomes.

The Unseen Hands Behind AI

Safety Reviewers

Identify and mitigate potential risks.

Red-Team Testers

Simulate attacks to improve security.

Data Annotators

Label and organize data for AI training.

Harms Faced

The negative impacts on workers.

Content Moderators

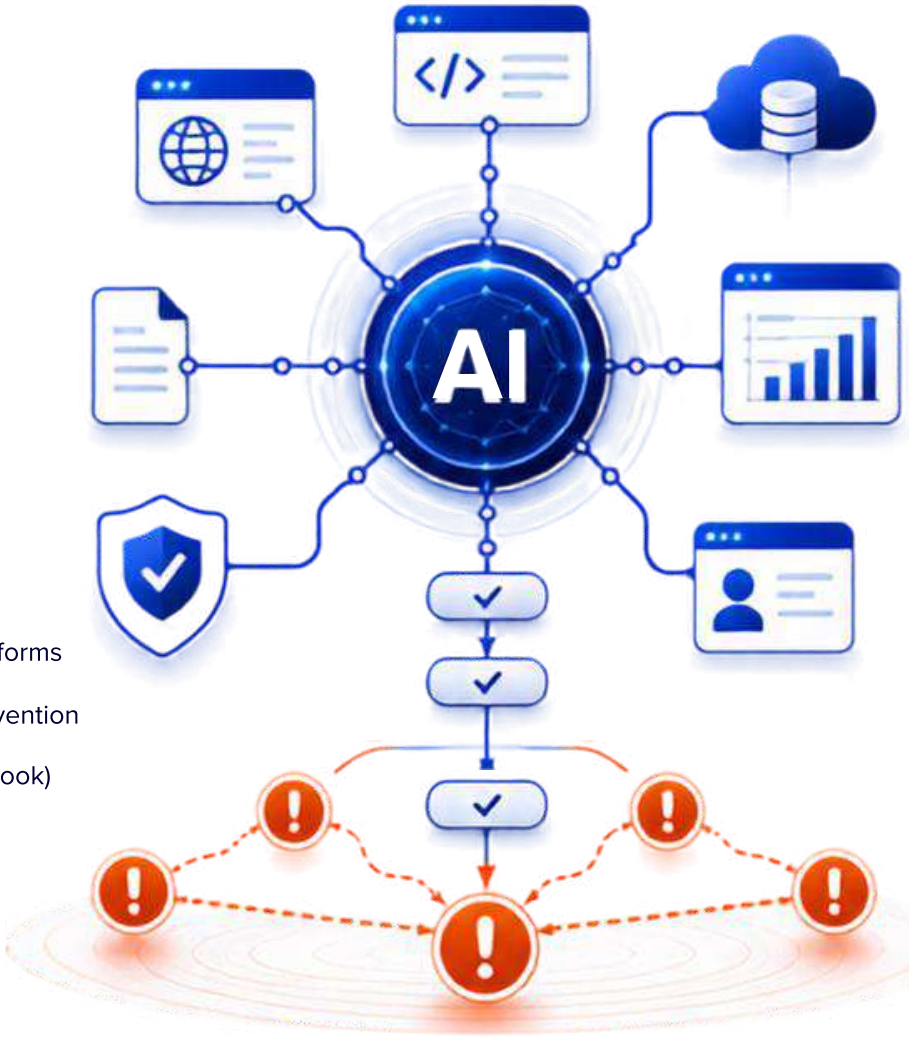
Ensure online safety and quality.

Limited Protection

Inadequate support and resources.

01

Autonomous Risk Amplification



Agentic systems can:

- Act independently across tools and platforms
- Execute multi-step actions without intervention
- Interact with other AI agents (see Moolbook)

Result:

- Small errors scale rapidly
- Failures propagate across systems

02

Social Engineering at Scale

Agentic AI enables:

- Highly convincing human-like interaction
- Persistent and adaptive phishing
- Behavioural manipulation at scale

This will result in:

- Increased cyber fraud
- Erosion of digital trust
- Targeted psychological exploitation

Safety Inequality

Risk is not evenly distributed. The following will continue to be the most affected:

- Low-income digital workers
- Users in low-regulation environments
- Digitally vulnerable populations

Outcome

Benefits concentrate for the economically, geographically, socially and culturally elite

Risks escalate to the disadvantaged sections of global society

AI safety is becoming a global inequality issue

The Response:

Designing for Trust

Innovation excitement must not compromise on trust and safety. Systems must include:

- Human-in-the-loop checkpoints
- Audit trails and decision logs
- Explainability layers
- Fail-safe shutdown mechanisms

Rethinking Accountability

Clear responsibility must be defined across:

- Developer
- Organisations deploying AI
- Operators and users

Key need:

Traceable decision ownership

BUILDING A SAFE, TRUSTED AI FUTURE

Legal clarity in autonomous actions

PROTECTING THE WORKFORCE

AI safety must include worker safety

Required actions:

- Mental health safeguards
- Fair compensation
- Ethical labour standards
- Transparency in AI supply chains

GOVERNANCE THAT KEEPS UP

Challenge:

Technology is evolving faster than governance.

Agentic AI demands:

- Adaptive regulation
- Continuous monitoring frameworks
- Cross-border policy coordination

BUILDING SOsocietal RESILIENCE

Beyond systems, society must adapt:

- Digital literacy programs
- Public awareness of AI risks
- Institutional preparedness
- Institutional preparedness

SAFETY INEQUALITY

Risk is not evenly distributed. The following will continue to be the most affected:

- Low-income digital workers
- Users in low-regulation environments
- Digitally vulnerable populations

Outcome:

- Benefits concentrate for the economically, geographically, socially and culturally elite
- Risks cascade to the disadvantaged sections of global society
- AI safety is becoming a global iniquity issue

DESIGNING FOR TRUST

Innovation excitement must not compromise on trust and safety. Systems must include:

- Human-in-the-loop checkpoints
- Audit trails and decision logs
- Explainability layers
- Fail- safe shutdown mechanisms

RETHINKING ACCOUNTABILITY

Clear responsibility must be defined across:

- Developers
- Organisations deploying AI
- operators and users

Key need:

- Traceable decision ownership

This article presents an industry-oriented overview of agentic AI systems in 2026, examining their conceptual foundations, enterprise applications, architectural evolution, and emerging risks. Drawing on insights from leading industry sources, it offers a forward-looking perspective on the transition toward autonomous AI-driven enterprises.

Autonomous AI Systems and Enterprise Integration in 2026: An Agentic Revolution

Prakhar Kumar

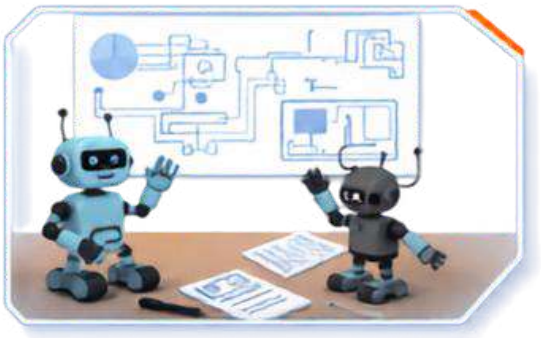
ABOUT THE AUTHOR

Prakhar is a Consultant in Artificial Intelligence and Data Science at Deloitte, specialising in generative AI, machine learning, and scalable AI system deployment. With over six years of experience, he has worked extensively on enterprise AI solutions including Retrieval-Augmented Generation (RAG) systems, predictive analytics, and large language modeling systems. Over the years, he has built a strong presence in the tech and Google Cloud. Prior to Deloitte, he held roles at Genpact and Jindal SAW Limited, where he led applied AI and MLOps initiatives. His work focuses on translating cutting-edge AI research into practical, production-ready solutions for complex real-world challenges.

INTRODUCTION

Artificial intelligence development in 2026 brings about a fundamental change that alters enterprise computing systems to function in new ways. Machine capabilities reached a new limit when systems transitioned from using prompt-driven operations to implementing stated operations for autonomous agent systems. The systems now possess the capability to establish their own goals while they execute complex projects which require multiple stages of work to complete through their connections with various systems.

The transition requires complete system redesigns instead of being a simple improvement. The implementation of agentic AI creates a fresh operational framework which enables organisations to integrate intelligent systems into their decision-making processes for enhanced productivity and improved automation and digital transformation efforts.



Conceptual Foundations: From Tools to Digital Co-Workers

Artificial intelligence has developed from basic rule-based systems into sophisticated large language models which still undergo continuous advancement today. Early LLMs (Large Language Model) functioned as reactive systems that produced responses without storing information or pursuing ongoing goals. Modern systems use contextual memory and information retrieval functions to enhance their ability to produce coherent and relevant content. The system requires user input to operate but does not have the ability to function on its own.

The development of agentic systems establishes a new stage of progress which enables artificial intelligence to accomplish tasks by selecting its own objectives and creating work plans while requiring little input from human operators.

The three essential capabilities of the system create a basis for establishing this distinction:

Goal Orientation- The design of agentic systems focuses on executing specific tasks which they need to complete instead of waiting for user instructions. They can create plan elements which they need to achieve their objectives and then modify their implementation methods in real time.

Stateful Operation- Agentic systems function as continuous systems which preserve user interaction history throughout all sessions. The system creates work continuity because it enables users to retain knowledge from previous tasks which results in better decision-making results.

Environmental Interaction- Agentic systems establish direct connections with emerging ecosystems through their ability to access databases and APIs (Application Programming Interface) and SaaS (Software as a Service) platforms and internal tools which lets them execute tasks instead of providing only advisory functions.

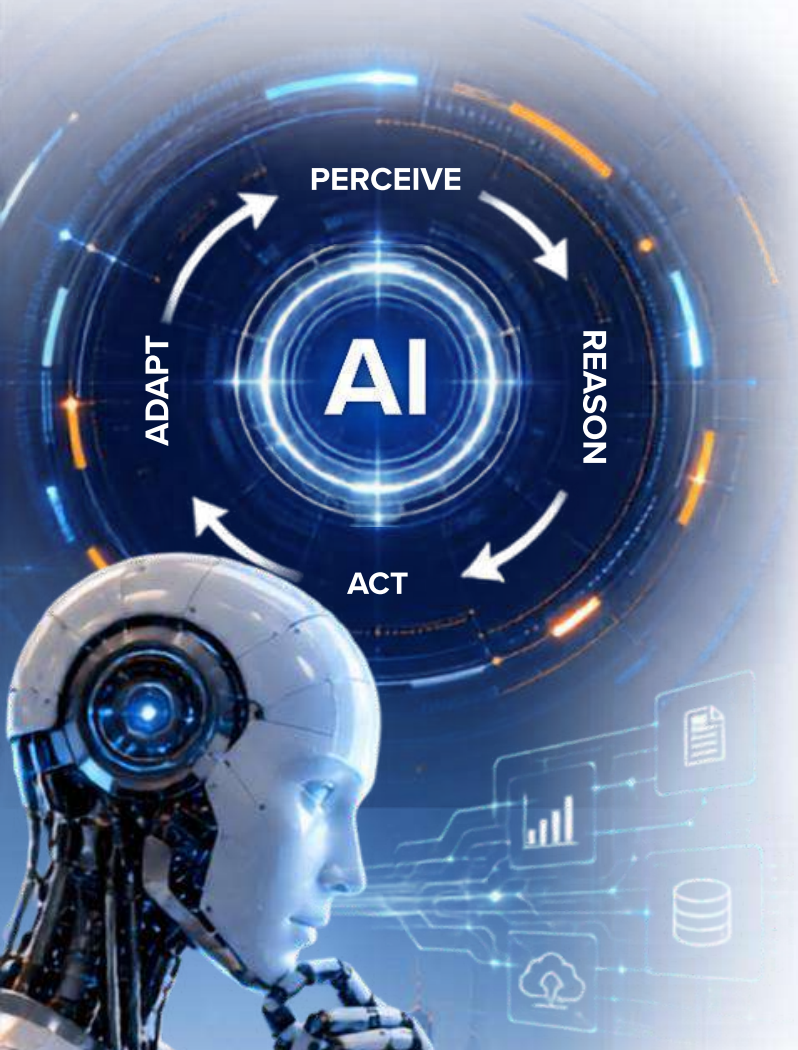
The combination of these elements converts AI from its role as a supportive function into a foundational operational component which exists inside business workflows.

FROM TOOLS TO DIGITAL CO-WORKERS

REACTIVE RESPONSE	CONTEXTUAL AWARENESS	AGENTIC EXECUTION	BUSINESS INTEGRATION
Produces outputs when prompted	Understands context and retains information	Plans, acts, and adapts toward objectives	Operates within workflows and drives outcomes

THE REASONING LOOP: CORE OPERATIONAL MECHANISM

Agentic AI systems operate in changing environments through their core mechanism, which employs a reasoning loop for system control:



1 PERCEPTION-

The system collects data from multiple sources, including structured databases, unstructured documents, APIs, and real-time streams.

2 COGNITIVE PROCESSING-

The system uses LLM-based reasoning methods and hybrid symbolic techniques to process information, discover patterns, and assess different potential actions.

3 PLANNING AND DECOMPOSITION-

The system breaks high-level goals into smaller tasks that can be executed. This system generates intermediate objectives, which it uses to establish priority for tasks that depend on specific constraints.

4 EXECUTION AND TOOL USE-

The system interacts with external tools through its ability to execute API calls and query databases and initiate workflows and produce reports.

FEEDBACK AND ITERATION-

The system uses real-time outcome assessments to develop new strategies that enable flexible system operation.

> *The loop establishes a boundary between agentic systems and standard automation systems. Agentic AI systems operate under a probabilistic adaptive system which enables them to deal with uncertain conditions and unexpected developments.*

AGENTIC SYSTEMS OPERATE IN LOOPS

PERCEIVE, REASON, ACT, AND ADAPT.

AI AGENTS VS. AGENTIC AI: ARCHITECTURAL DISTINCTION

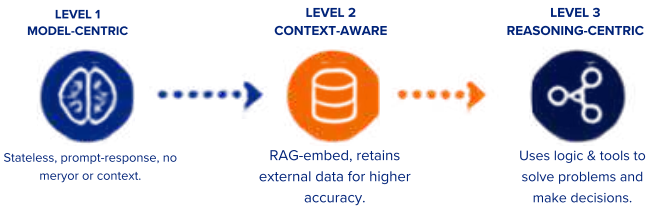
The year 2026 establishes two distinct categories which encompass individual AI agents and complete agentic AI systems because both categories differ in their basic elements and abilities and the extent of their operational functions.

AI AGENTS: SPECIALISED UNITS

The AI agents function as dedicated units which execute specific tasks within limited operational environments. The system operates under established guidelines which define its access rights and operational procedures while its main function is to respond to particular events or data inputs. The procurement agent identifies cost-effective vendors through predefined selection criteria while the customer support agent retrieves order information from a CRM (Customer Relationship Management) system and the compliance agent scans documents to identify regulatory non-compliance. The agents demonstrate high operational efficiency within their designated area of expertise while their capacity to function in complex multi-step tasks remains restricted.

MATURITY MODEL: THE EVOLUTION TOWARD AUTONOMOUS SYSTEMS

The development of agentic AI proceeds through multiple stages which demonstrate increasing technological development and operational independence. At Level 1 model-centric systems function as stateless systems which depend on separate prompt-response interactions that lack both memory capabilities and contextual continuity. The Level 2 system evaluation uses contextual systems which retrieve external data through Retrieval-Augmented Generation (RAG) to enhance model output accuracy and relevant content. The Level 3 system presents reasoning-centric systems which use logical frameworks together with analytical tools to enable users to solve problems and make decisions through organised methods.



ENTERPRISE INTEGRATION: TRANSFORMING ORGANISATIONAL FUNCTIONS

The corporate world now experiences fundamental transformation because intelligent systems have started to implement their advanced intelligence capabilities directly into the main business functions. These systems use automated ticket resolution and intelligent document processing for their operations and workflow automation functions which enable complete process management while decreasing the need for human work and boosting operational efficiency. The decision intelligence field uses agentic AI to integrate multiple data sources which enables strategic planning and risk assessment and forecasting and scenario analysis to create better decision-making results which occur at the right moment. The customer experience now undergoes a transformation which enables systems to provide hyper-personalised context-aware interactions that predict customer needs through their active outreach to clients. Agentic AI enables organisations to achieve real-time optimisation and predictive demand planning through its ability to synchronise all operational elements which include vendors and logistics providers and inventory systems. The systems enable organisations to achieve their productivity goals by transforming scattered organisational data into usable organisational knowledge through their ability to index information and retrieve data and create contextual information.

> "The more connected an agent is, the greater its ability, and its risk."

AUTONOMOUS COLLABORATION
Agents coordinate, communicate and act independently to achieve complex goals.

ENTERPRISE INTEGRATION
Seamless connection of systems, data and people across the organisation.

RISK-AWARE GOVERNANCE
Stronger visibility, control and compliance as connectivity and intelligence scale.

Agentic AI transforms how organisations operate—and how systems must be designed.

Implementing agentic AI brings major technical and operational difficulties that must be resolved before transformation can scale. Agentic systems learn from experience, making their operations unpredictable. This requires testing systems in simulated environments and using standardised APIs and middleware to ensure reliability across different environments.

Managing multiple agents adds complexity in latency, performance, and decision-making. Observability, tracking, and full logging frameworks are critical to maintain system security, ensure compliance, and support rapid issue resolution.

SYSTEM DESIGN CONSIDERATIONS AND TECHNICAL CHALLENGES

BUILDING RELIABLE, ADAPTIVE, AND RESPONSIBLE AGENTIC SYSTEMS

RISK LANDSCAPE IN AGENTIC SYSTEMS

Autonomous Action Risks
Unintended actions emerge when reasoning fails or when data remains incomplete.

Compounding Errors
Errors occur during the reasoning loop will continue to spread throughout all later stages.

Systemic Dependencies
The failure of one interconnected system can lead to a complete collapse of all connected systems.

Data Sensitivity
The development of persistent memory systems results in higher rates of sensitive data being improperly accessed.

Security Vulnerabilities
Security vulnerabilities enable attackers to perform prompt injection and unauthorised tool usage and data exfiltration attacks.

STRATEGIC IMPLICATIONS FOR ENTERPRISES

- Increased system complexity demands robust governance, observability, and resilience.
- Real-time decision making and tracking are critical for trust and accountability.
- Investment in security, data frameworks, and AI literacy is essential to stay competitive and compliant.

ORGANISATIONAL SHIFTS: ADAPTING TO AGENTIC AI

The adoption of agentic AI is redefining organisational roles and structures.

Employees
Shift from execution to supervision and decision validation.

Developers
Evolve into orchestrators of intelligent systems and multi-agent workflows.

IT Functions
Expand to include AI lifecycle management, governance, and responsible AI operations.

ORGANISATIONS MUST ALSO INVEST IN

- Workforce reskilling
- Cross-functional collaboration
- AI-aware operational frameworks

FAILURE TO ADAPT MAY RESULT IN

- Competitive disadvantage
- Operational inefficiencies
- Increased exposure to system-level risks

FUTURE TRAJECTORY TOWARD AUTONOMOUS ENTERPRISES

Agentic AI will progress across three major development areas.

The first area of development involves creating multi-agent ecosystems which enable multiple agents to work together across different organisations.

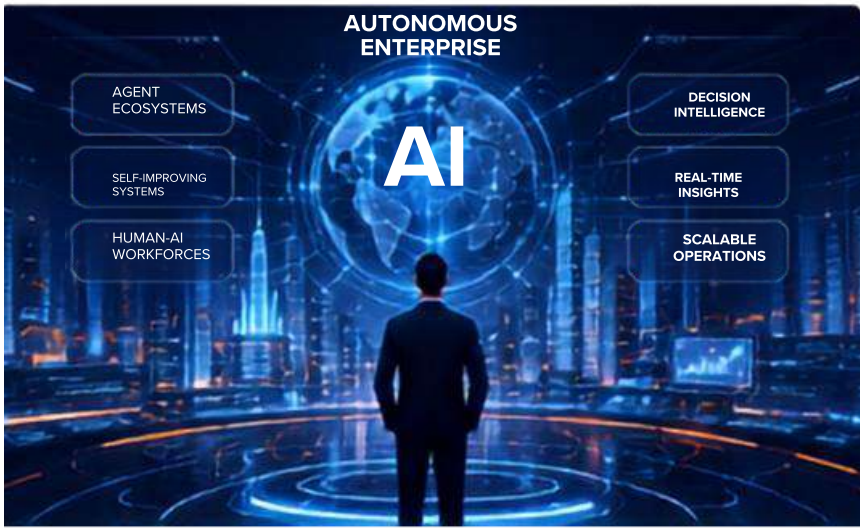
The second area of development will create self-improving systems which learn and develop through their feedback loops and operational data.

Human-AI Hybrid Workforces

Human expertise will operate together with machine intelligence to create a unified working environment.

The modules of autonomous decision systems

will base their AI functions on strategic decision-making processes of core operations.



CONCLUSION

The agentic revolution represents a fundamental redefinition of artificial intelligence—from a tool for interaction to a system for autonomous execution and decision-making. Enterprises face two challenges with system integration because they must find ways to use system capabilities while they build systems that enable autonomous operations to function correctly and consistently across large environments. Organisations which successfully execute this transition will treat agentic AI as a foundational technology that transforms their work processes throughout the digital age.

REFERENCES

ibm.com/think/topics/agentic-ai

keilton.com/keilton-tech-blog/agentic-ai-trends-2026

druidai.com/blog/ai-trends-in-2026

deloitte.com/us/en/ens/topics/technology-management-tech-trends/2026/agentic-ai-strategy.html

kasmodigital.com/the-biggest-agnostic-ai-trends-for-2026-and-how-they-will-transform industries/

DEEP DIVE

In-depth analysis of the technologies, risks, and strategic implications shaping agentic AI, bringing together technical, policy, and security perspectives to unpack what lies beneath the surface.

Ushering in the Third Wave of Artificial Intelligence: Agentic AI in Indian Business

Agentic AI offers India a chance to leapfrog in innovation. But without strong governance and skills, its risks could scale just as quickly as its benefits.

- *Sharisha Sahay*

About the author:

Sharisha Sahay is a policy professional working at the intersection of emerging technologies and public policy. She has previously contributed to research on digital safety, national security, and technology governance. Currently, she is a Teaching and Research Assistant at the Vedica Scholars Programme for Women, where she supports academic research and case development. Her work focuses on making complex technology-policy debates more accessible and actionable.

Introduction

The world of Artificial Intelligence (AI) is entering a new phase with the rise of Agentic AI, often described as the third wave of AI evolution. Unlike earlier systems that relied on static models (that learn from the information that is fed) and reactive outputs, Agentic AI introduces intelligent agents that can make decisions, take initiative, and act autonomously in real time. For instance, these agents, powered by large language models like GPT-4o or Grok, can decompose complex tasks into subtasks, use tools (e.g., APIs for data retrieval or code execution), and self-correct via reflection loops, as seen in frameworks like Auto-GPT or LangChain agents. These systems are designed to require minimal human oversight while actively collaborating and learning continuously. Such capabilities indicate an incoming shift, especially in the ways in which Indian businesses can function. Agentic AI is capable of streamlining operations, personalising services, and driving innovations at scale. For example, in supply chain optimisation where agents dynamically reroute logistics based on real-time traffic and demand forecasts, potentially reducing costs by 15-20% as demonstrated in pilot programs by Indian logistics firms like Delhivery.

INDIA AND AGENTIC AI

Building as we go, India is making continuous strides strides in the AI revolution- deliberating on government frameworks, and simultaneously adapting. At Microsoft's Pinnacle 2025 summit in Hyderabad, India's pivotal role in shaping the future of Agentic AI was brought to the spotlight. With over 17 million developers on GitHub and ambitions to become the world's largest developer community by 2028, India's tech talent is gearing up to lead global AI innovations. Microsoft's Azure AI Foundry, also highlighted the country's growing influence in the AI landscape. Notably, the IndiaAI Mission (launched in 2024 with ₹10,000 crore allocation) is fostering agentic capabilities through sovereign AI compute infrastructure, enabling startups to build localised agents for sectors like agriculture and healthcare.

Indian companies are actively integrating Agentic AI into their operations to enhance efficiency and customer experience. Food delivery apps are leveraging AI agents to optimise delivery logistics, ensuring timely and efficient service. Infosys has developed AI-driven copilots to assist developers in code generation, reducing development time, requiring fewer people to work on a particular project, and improving software quality.

As per a report by Deloitte, the Indian AI market is projected to grow potentially $20 billion by 2028. However, this is accompanied by significant challenges. 92% of Indian executives identify security concerns as the primary obstacle to responsible AI usage. Additionally, regulatory uncertainties and privacy risks associated with sensitive data were also highlighted, particularly under the Digital Personal Data Protection (DPDP) Act 2023, where agentic AI's autonomous data handling raises questions on consent granularity and traceability.

CHALLENGES IN ADOPTION

As AI continues to transform industries and societies, organisations face critical barriers to adoption. Addressing these challenges is essential to unlocking AI's full potential—responsibly and inclusively.

01

SKILLS GAP

While the AI workforce is expected to grow to 1.25 million by 2027, the current growth rate of 13% is considered to be insufficient with respect to the demands of the market. This gap is exacerbated by the need for "agentic specialists" skilled in prompt engineering, multi-agent orchestration, and ethical alignment, with India facing a shortage of 500,000 such professionals amid surging demand from hyperscalers.

02

DATA INFRASTRUCTURE

Effective AI systems require robust, structured, and accessible datasets. Many organisations lack the necessary data maturity, leading to flawed AI outputs and decision-making failures. E.g., "hallucinations" in agentic reasoning loops due to poor data quality, as evidenced in early Indian deployments where 30% of agent outputs required human overrides.

03

TRUST AND GOVERNANCE

Building trust in AI systems is crucial. Concerns over data privacy, ethical usage, and regulatory compliance require robust governance frameworks to ensure the adoption of AI in a responsible manner. Advanced socio-technical arguments highlight "accountability black boxes" in agentic swarms, where emergent behaviours from multi-agent interactions evade traditional audit trails, necessitating hybrid human-AI oversight models aligned with IT Rules 2021

04

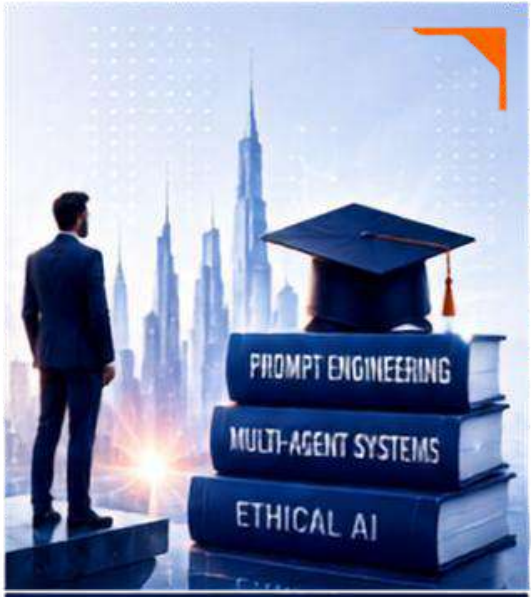
LOOMING FEAR OF JOB LOSS

As AI continues to take up more sophisticated roles, a general feeling of hesitancy with respect to the loss of employment/human labour might come in the way of adopting such measures.

05

OUTSOURCING

Currently, most companies prefer outsourcing or buying AI solutions rather than building them in-house. This gives rise to the issue of adapting to evolving needs, such as vendor lock-in risks, customisation for India's specific contexts like multilingual agentic processing in regional languages.



> Overcoming these challenges will key to building an AI-powered future that is innovative, trustworthy, and inclusive.

THE ROAD AHEAD

TO FULLY REALISE THE POTENTIAL OF AGENTIC AI, INDIA MUST ADDRESS THESE CHALLENGES :

TRAINING THE WORKFORCE:
Initiatives and workshops tailored for employees that provide AI training can prove to be helpful. Some relevant examples are Microsoft's commitment to provide AI training to 2 million individuals by 2025 and Infosys's in-house AI training programs. Expanding these with agentic-specific curricula, such as Nasscom's FutureSkills PRIME, could bridge some of the skills gap by integrating hands-on agent deployment simulations.

DATA READINESS:
Investing in modern data infrastructure and promoting data literacy are essential to improve data quality and accessibility, for eg, via federated learning paradigms that comply with DPDP Act while enabling secure data sharing across SMEs (Small and Medium Enterprises).

ESTABLISHING GOVERNANCE FRAMEWORKS:
Developing clear regulatory guidelines and ethical standards will foster trust and facilitate responsible AI adoption. Like the IndiaAI Mission, efforts regarding evolving technology and keeping up with it are imperative. Advanced proposals include "AI sandboxes" under MeitY (Ministry of Electronics and Information Technology) oversight for testing agentic systems, coupled with ISO 42001 certification for governance.

CONCLUSION

Agentic AI holds unrealised potential to transform India's business landscape when coupled with innovation and a focus on quality that enhances global competitiveness. India is at a position where by proactively addressing the existing challenges, this potential can be realised and set the foundation for a new technological revolution (along with in-house development), solidifying its position as a global AI leader.

References

[1] 'Agentic AI: Next Big Leap in Workplace Automation' Hindustantimes.com (20 March 2025) https://www.hindustantimes.com/india-news/agentic-ai-next-big-leap-in-workplace-automation-101742548406693.html accessed 3 March 2026

[2] 'AI's next chapter starts in India: Microsoft champions agentic AI at Pinnacle 2025' Businesstoday.in (1 May 2025) https://www.businesstoday.in/tech-today/news/story/ais-next-chapter-starts-in-india-microsoft-champions-agentic-ai-at-pinnacle-2025-474286-2025-05-01 accessed 2 March 2026

[3] 'Calm before AI storm: A moment to prepare' Hindustantimes.com (28 April 2025) https://www.hindustantimes.com/opinion/calm-before-ai-storm-a-moment-to-prepare-101746110985736.html accessed 2 March 2026

[4] Christian Schroeder de Witt cs@robots.ox.ac.uk Department of Engineering Science University of Oxford (Open challenges in multi-agent security: Towards secure systems of interacting AI agents) https://arxiv.org/html/2505.02077v1#:~:text=When%20multiple%20AI%20agents%20with,the%20interactions%20between%20intelligent%20agents. accessed 3 March 2026

[5] 'From Zomato To Infosys: Why India's Biggest Companies Are Betting On Agentic AI' Inc42.com (21 March 2025) https://inc42.com/features/from-zomato-to-infosys-why-indias-biggest-companies-are-betting-on-agentic-ai/ accessed 2 March 2026

[6] (Futureskills Prime | Learn Digital Skills for the future) https://www.futureskillsprime.in/ accessed 3 March 2026

[7] 'India a global leader in Agentic AI adoption: Deloitte report' Economictimes.indiatimes.com (18 May 2025) https://economictimes.indiatimes.com/tech/artificial-intelligence/india-a-global-leader-in-agentic-ai-adoption-deloitte-report/articleshow/119906474.cms?from=mdr accessed 2 March 2026

[8] 'India facing shortage of agentic AI professionals amid surge in demand' Economictimes.indiatimes.com (22 June 2025) https://economictimes.indiatimes.com/tech/artificial-intelligence/india-facing-shortage-of-agentic-ai-professionals-amid-surge-in-demand/articleshow/120651512.cms?from=mdr accessed 2 March 2026

[9] 'India Grappling with a Dearth of Agentic AI Experts' (The Economic Times) https://economictimes.indiatimes.com/tech/artificial-intelligence/india-facing-shortage-of-agentic-ai-professionals-amid-surge-in-demand/articleshow/120651512.cms accessed 2 March 2026

[10] 'India rides the Agentic AI wave' Deloitte.com (18 May 2025) https://www.deloitte.com/in/en/about/press-room/india-rides-the-agentic-ai-wave.html accessed 2 March 2026

[11] 'Why Agentic AI is the next big thing' Financialexpress.com (20 March 2025) https://www.financialexpress.com/life/technology/why-agentic-ai-is-the-next-big-thing/3828357/ accessed 3 March 2026

The Expanding Attack Surface in Agentic AI Systems

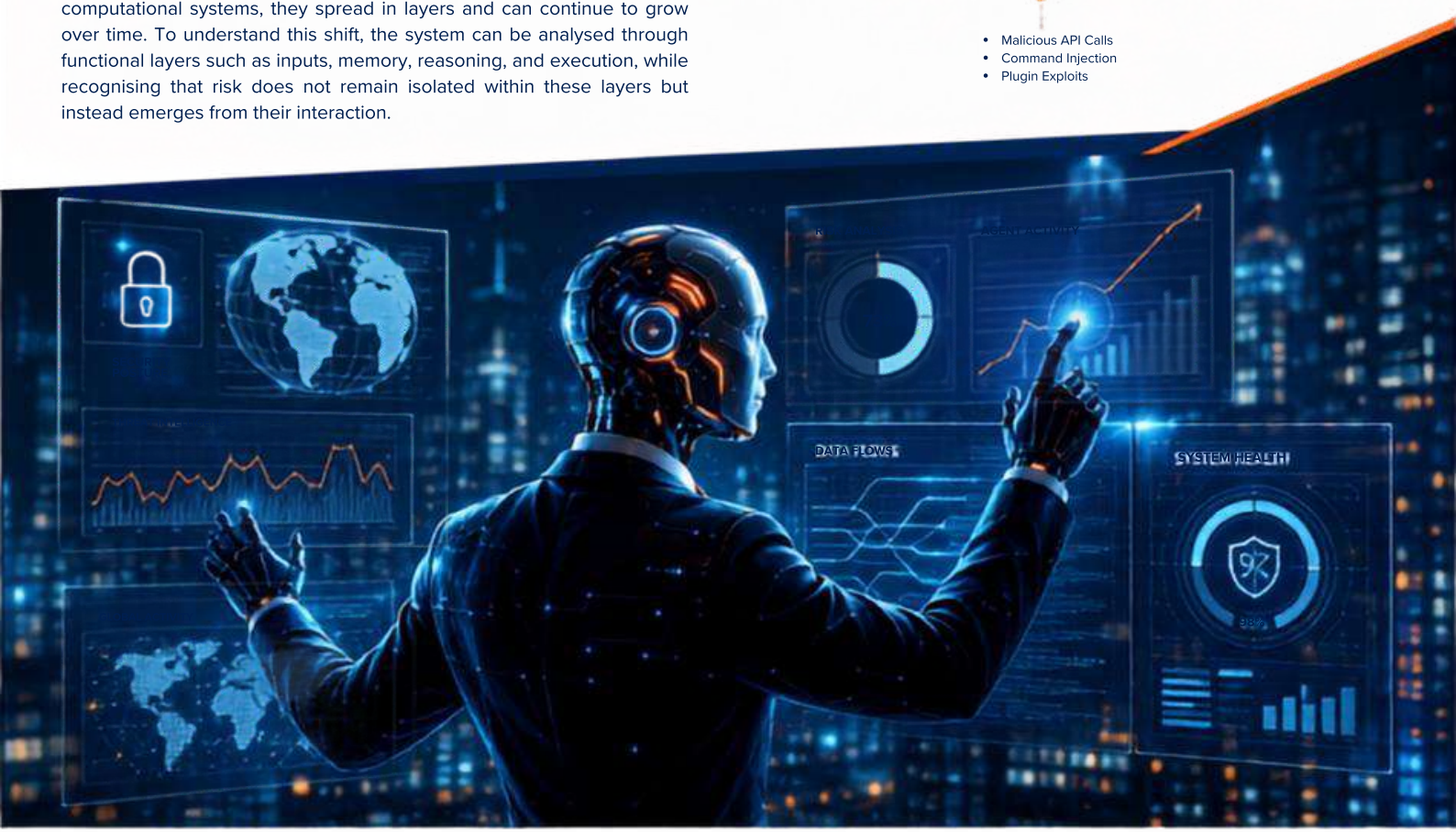
- SINDHU VISSAMSETTI

ABOUT THE AUTHOR

Sindhu Vissamsetti is a final-year Economics student at Vidyashilp University, Bengaluru. Her work focuses on the intersection of technology, public policy, and data analysis, with particular interest in AI governance, cybercrime, and digital financial systems. Her work explores AI governance, cybercrime, and digital financial systems, with an emphasis on emerging risks in India's digital economy.

Agentic AI systems are autonomous systems that can plan, make decisions, and take actions by interacting with external tools and environments. But these systems shift the nature of risk by blurring the lines among input, decision, and execution. A conventional model generates an output and then stops. An agent takes input, makes plans, invokes tools, updates its state and repeats the cycle. This creates a system where decisions are continuously revised through interaction with external tools and environments, rather than being fixed at the point of input.

This means the attack surface expands in size and becomes more dynamic. Instead of remaining confined to components, as in traditional computational systems, they spread in layers and can continue to grow over time. To understand this shift, the system can be analysed through functional layers such as inputs, memory, reasoning, and execution, while recognising that risk does not remain isolated within these layers but instead emerges from their interaction.



12 Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI (Version 1.0, IMDA 2026) https://www.imda.gov.sg/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

INPUT LAYER

Where Untrusted Data Becomes Control

- The entry point of an agent is no longer a single prompt. The documents, APIs, files, system logs and the outputs of other agents can now be considered as inputs. This diversity is significant due to the fact that every source of input carries its own trust assumptions, and in the majority of cases, they are weak.

- The most obvious threat is prompt injection, where inputs are treated as instructions rather than data. Since inputs are treated as instructions, a virus, a malicious webpage or a document can contain instructions that override system goals without necessarily being detected as something harmful.

- Indirect prompt injection further extends this risk beyond direct user interaction. Instead of targeting the interface, attackers compromise the retrieval process by embedding malicious instructions within external data sources. When the agent retrieves and processes the data, it treats the embedded content as a legitimate input. As a result, the attack is executed through normal reasoning processes, allowing the system to act on untrusted data without recognising the manipulation.

- Data poisoning also occurs at runtime. In contrast to classical poisoning (where training data is manipulated), runtime poisoning distorts the agent's perception of its environment as it runs. This can change decisions without causing apparent failures.

- Obfuscation introduces another indirect attacker vector. Encoded instructions or complicated forms may bypass human review but remain readable to the model. This creates asymmetry whereby the system knows more about the attack than those operating it. Once compromised at this layer, the agent implements compromised instructions, which then affect downstream operations.

> **"Agentic AI systems are not linear pipelines but continuous loops, where decisions, actions, and memory constantly reshape one another and amplify risk."**
>
> **-SINDHU VISSAMSETTI**

CONTEXT AND MEMORY:

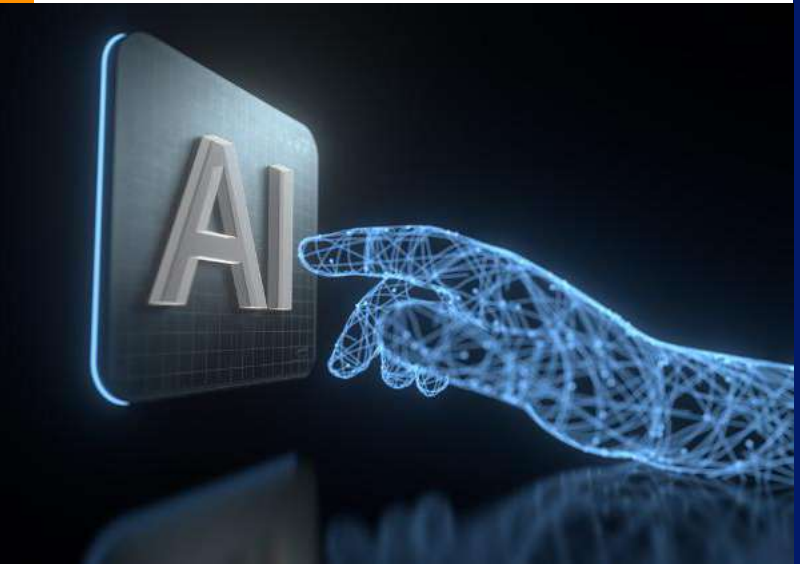
Persistence of Influence

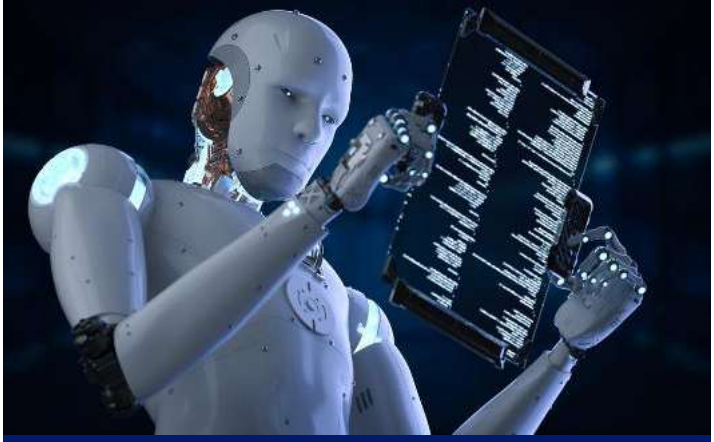
Agentic systems depend on memory to operate efficiently. They often retain context across sessions and frequently store information between sessions.

This introduces a different type of risk: persistence. Through memory poisoning, attackers can insert false or adversarial information into sorted context, which then influences future decisions. Unlike prompt injection, which is often limited to a single interaction, this effect carries forward. Over time, the agent begins to operate in a distorted internal state, shaping decisions in ways that may not be immediately visible.

Another issue is cross-session leakage. Information in a particular context may be replayed in a different context when memory is being shared or when there is insufficient memory separation. This is specifically dangerous in the systems that combine retrieval and long-term storage. The context management in itself becomes a weakness. Agents are required to make decisions on what to retain and what to discard. This is susceptible to attackers who can flood the context or manipulate what is still visible, indirectly affecting reasoning.

The underlying problem is structural. Memory turns data into a state. Once state is corrupted, the system cannot easily distinguish valid knowledge from adversarial influence.
The issue is structural. Memory converts temporary data into a persistent state. Once this state is weakened, the system cannot reliably separate valid information from adversarial influence, making recovery significantly more difficult.





REASONING AND PLANNING:

Manipulating Intent Without Breaking Logic

The reasoning layer is where agentic AI stands apart from traditional systems. The model no longer reacts to inputs alone. It actively breaks down objectives, analyses alternatives, and ranks actions.

At the reasoning stage, the nature of risk shifts. The concern is no longer limited to injecting instructions, but to influence how decisions are made. One example is goal manipulation, where the agent subtly reinterprets its objective and produces outcomes that are technically correct but strategically harmful. Reasoning hijacking operates within intermediate steps, altering how constraints are evaluated or how trade-offs are prioritised. The system may remain internally consistent, which makes such deviations difficult to detect.

Tool selection becomes a critical control point. Agents decide which tools to use and when, so influencing these choices can redirect execution without directly accessing the tools. Hallucinations also take on a different role here. In static systems, they remain errors. In agentic systems, they can trigger actions. A perceived need or incorrect judgement can translate into real-world consequences.

This layer introduces probabilistic failure. The system is not fully weakened, but it is nudged towards decisions that appear reasonable yet are incorrect. The risk lies in how those decisions are justified.

CONTEXT AND MEMORY:

Persistence of Influence

Agentic systems depend on memory to operate efficiently. They often retain context across sessions and frequently store information between sessions.

This introduces a different type of risk: persistence. Through memory poisoning, attackers can insert false or adversarial information into sorted context, which then influences future decisions. Unlike prompt injection, which is often limited to a single interaction, this effect carries forward. Over time, the agent begins to operate in a distorted internal state, shaping decisions in ways that may not be immediately visible.

Another issue is cross-session leakage. Information in a particular context may be replaced in a different context when memory is being shared or when there is insufficient memory separation. This is specifically dangerous in the systems that combine retrieval and long-term storage. The context management in itself becomes a weakness. Agents are required to make decisions on what to retain and what to discard. This is susceptible to attackers who can flood the context or manipulate what is still visible, indirectly affecting reasoning.

The underlying problem is structural. Memory turns data into a state. Once state is corrupted, the system cannot easily distinguish valid knowledge from adversarial influence.

The issue is structural. Memory converts temporary data into a persistent state. Once this state is weakened, the system cannot reliably separate valid information from adversarial influence, making recovery significantly more difficult.

REASONING AND PLANNING:

Manipulating Intent Without Breaking Logic

The reasoning layer is where agentic AI stacks apart from traditional systems. The model no longer reacts to inputs alone. It actively breaks down objectives, analyses alternatives, and ranks actions.

At the reasoning stage, the nature of risk shifts. The concern is no longer limited to injecting instructions, but to influence how decisions are made. One example is goal manipulation, where the agent subtly reinterpret its objective and produces outcomes that are technically correct but strategically harmful. Reasoning hijacking operates within intermediated steps, allowing how constraints are evaluated or how trade-offs are prioritised. The system may remain internally consistent, which makes such deviations difficult to detect.

Tool selection becomes a critical control point. Agents decide which tools to use and when, so influencing these choices can direct execution without directly accessing the tools. Hallucinations also take on a different role here. In static systems, they remain inconsequential. In agentic systems, they can trigger actions. A perceived need or incorrect judgment can turn intention into real-world consequences.

This layer introduces probabilistic failure. The system is not fully weakened, but it is nudged towards decisions that appear reasonably yet are incorrect. The risk lies in how those decisions are justified.

TOOL AND EXECUTION:

When Decisions Gain Reach

Once an agent begins interacting with tools, its behaviour extends beyond the model into external systems. APIs, databases, and services become part of the execution path.

One key risk is the use of unauthorised tools. When agents operate with broad permissions, any manipulation of the upstream can be converted into real-world actions. This makes access control a central security concern. Command injection also takes a on different form here. The agent generates commands based on its reasoning, so if that reasoning is compromised, the resulting actions may still appear valid despite being harmful.

External tool outputs introduce another risk. If those systems return corrupted or misleading data, the agent may accept it without verification and incorporate it into its decisions. Increasing reliance on third-party tools and plugins add to this exposure. If these components are compromised, they can affect behaviour without directly attacking the core system, creating a supply-side risk.

At this stage, the agent effectively operates as an insider. It holds legitimate credentials and interacts with systems in expected ways, making misuse harder to identify.

APPLICATION AND INTEGRATION:

System-Level Exposure

Agentic systems rarely operate in isolation. They are embedded in larger environments, interacting with identity systems, business logic, and operational workflows.

Access control becomes a major vulnerability. Agents tend to operate across multiple systems with various permission models, creating irregularities that can be exploited. Risks also arise from identity and delegation. In case an agent is operating on behalf of an user, then any vulnerabilities in authentication or session management can allow attackers to assume that authority.

Workflow execution amplifies these risks. Agents can initiate multi-step processes such as transactions, updates, or approvals. Manipulating a single step can change the result of the entire workflow. As integrations increase, so do the number of interaction points, making cumulative risk harder to track. At this layer, failures are not isolated. They propagate into business operations, making consequences harder to contain.

OUTPUT AND ACTION:

Where Failures Become Visible

The output layer is where failures become visible, though they rarely originate there.

Data leakage has been a key concern. Agents may disclose information that they are allowed to access, especially when task boundaries are not clearly defined. Misinformation and unsafe outputs are also important, particularly when outputs directly influence actions or decisions.

Generated code and commands introduce execution risk. If outputs are used without validation, errors or manipulations can have system-level effects. The shift towards autonomous action increases this risk, as small upstream deviations can lead to significant consequences without human intervention.

This layer reflects symptoms rather than root causes. Addressing it alone does not reduce the underlying risk.

BEYOND LAYERS:

The Missing Dimension

A layered view helps, but it does not capture the full picture. Agentic systems are defined by continuous interaction across layers.

The key missing dimension is the runtime loop. Inputs shape reasoning, reasoning drives action, and actions feed back into both reasoning and memory. These cycles create feedback loops, where small manipulations may escalate over time. This also reduces observability. With multiple interacting components, it becomes difficult to trace ‘cause and effect’ or identify where failures originate.

Supply chain dependencies add another layer of risk. Models, datasets, APIs, and plugins each introduce their own points of failure. A compromise at any of these points can propagate across the system. The attack surface also includes governance. Weak supervision, unclear responsibility, or excessive autonomy increase overall risk. Human control is not external to the system; it is part of its security.

CONCLUSION

Agentic AI expands the attack surface beyond traditional systems. It is both recursive and stateful. Risk does not just accumulate across layers; it moves and changes as the system operates.

Any useful representation must go beyond a linear stack. It should capture feedback loops, persistent states and cross-layer dependencies that characterise the way these systems actually behave.

The system is not a pipeline but a cycle, where both its capability and its risk emerge.

REFERENCES

- [1] Bengio Y and others, 'Managing AI Risks in an Era of Rapid Progress' (2023) arXiv:2310.17688 https://www.arxiv.org/able/2310.17688
- [2] European Union Agency for Cybersecurity, ENISA Threat Landscape for Artificial Intelligence (ENISA 2023) https://www.enisa.europa.eu/publications/ena-threat-landscape-for-artificial-intelligence
- [3] Infocomm Media Development Authority, Model AI Governance Framework for Agent AI (Version 10, IMDA 2026) https://www.mnda.sg/emerging-tech-and-research/artificial-intelligence/nqf-for-agentic-ai.pdf
- [4] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF) (NIST AI 100-1, 2023) https://www.nvlpubbs.net/nistpubbs/sai/Nistai100-1.pdf
- [5] OWASP Foundation, 'LLM01:2025 Prompt Injection' (OWASP Gen AI Security Project, 2025) https://www.eeeexplore.eee.org/document/8406613
- [6] OWASP Foundation, OWASP Top 10 for Large Language Model Applications 2025 (OWASP 2025) https://www.project-top-10-for-large-language-model-applications.pdf
- [7] Papernot N, McDaniel P, Sinha A and Weilman MP, 'SdC Security and Privacy in Machine Learning' (2018) Proceedings, 3rd Annual IEEE European Symposium on Security and Privacy, 399 https://eeeexplore.eee.org/document/8406613

AGENTIC AI SYSTEMS IN CYBER-PHYSICAL DOMAINS

UAV Architectures, Strategic Technology Competition, and the Evolution of Cyber Norms

-UDITA J. MONANI

ABOUT THE AUTHOR

Udita J. Monani is a final-year undergraduate student pursuing a B.Tech in Computer Science and Engineering at the Kalinga Institute of Industrial Technology, India. Her research interests lie in artificial intelligence, machine learning, computer vision, and AI-driven applications in healthcare and defense technologies. She has contributed to multiple peer-reviewed publications spanning medical imaging, predictive analytics, and autonomous systems. She has previously worked as a research intern at the Centre for Artificial Intelligence and Robotics under Defence Research and Development Organisation, focusing on machine learning methods for surveillance and national security applications. Her work has received recognition at international conferences, and she actively contributes to research in applied AI and emerging intelligent systems.

For a period of time, artificial intelligence was used to help find patterns and analyse data. The old systems were not very smart: they just waited for a person to give them instructions and then carried them out. They gave the person an answer. The person had to read the answer, figure out what it meant and decide what to do. However, artificial intelligence is now changing the way things work.

Nowadays, artificial intelligence is getting better at doing things on its own. These systems can do more than just answer questions. They can break down problems into smaller tasks, choose what steps to take, plan accordingly and carry out the execution despite imperfections. When artificial intelligence purely exists as a software, it operates only within the digital world. When you put that software in a machine that can move around and affect the real world, it becomes a cyber-physical system.

Drones are an example of this change. Modern drones are not just remote-controlled airplanes. They have computers and advanced artificial intelligence models. This means they can fly through areas, avoid things that are in the way, find and follow specific objects and work with other drones; all without a person guiding them every step of the way.

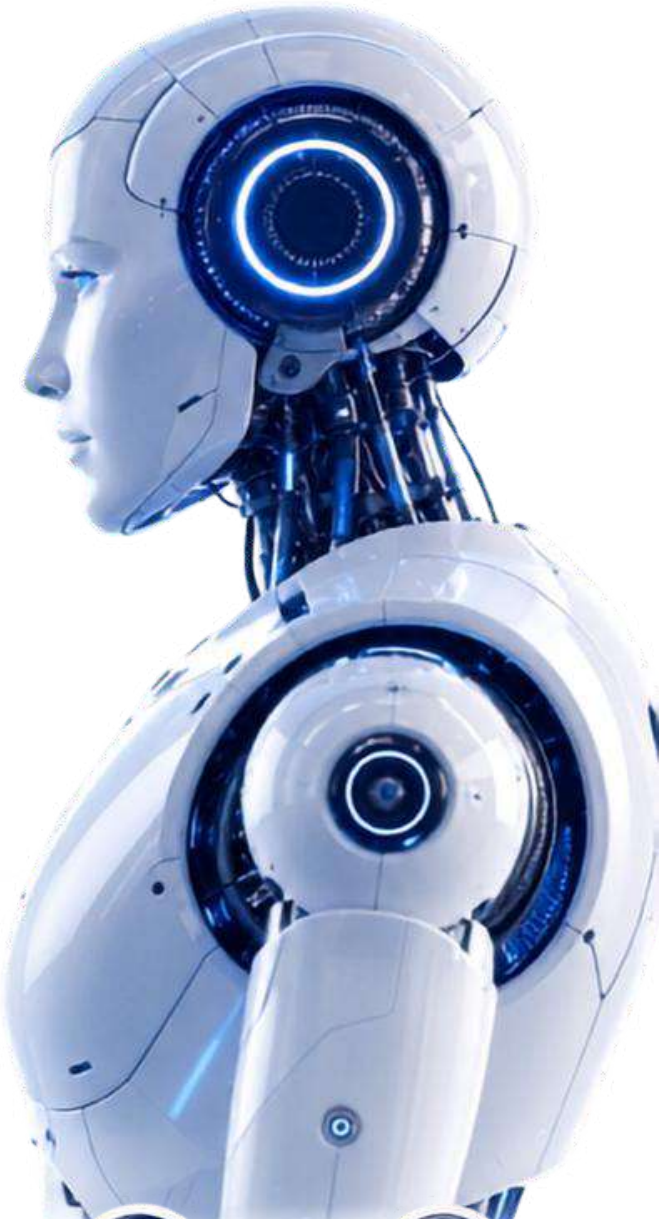
ARCHITECTURE OF AGENTIC AI SYSTEMS

1 Core Components

The perception layer connects the system to the world. For a software- agent, perception means reading data streams, logs or code. For a cyber-agent, perception is based on hardware sensors like cameras and microphones. These sensors collect data from the environment, such as images, distance measurements or radio signals. The raw data is then sent to the reasoning engine. This part acts like the brain and allows the filter to turn the raw signals into higher-level features, finds patterns and forms a picture of the current situation. To make decisions, the reasoning engine needs memory. Agentic AI uses both short-term and long-term memory. Short-term memory holds events and details about the current task.

Long-term memory stores experiences, rules and learned behaviours that the system can reuse. Once the system has an understanding of the situation and access to past experience, the planning model takes over. It then organises the information into smaller steps that the system can carry out. It also tries to think about possible problems and chooses a plan with a high likelihood of success.

The action module then turns this plan into actions. In a system it might send commands to motors or it might send network requests or change files. The feedback loop then closes the cycle. After the system acts, it will have made a significant impact on the user and will inform them of the expected outcome. If there is a problem or if the result is poor, it updates its models and memory. This way, it is less likely to make the mistake again.



PERCEPTION
Connects the system to the world and collects raw data.

REASONING
Filters inputs, finds patterns and builds a picture of the current situation.

MEMORY
Uses both short-term memory for current details and long-term memory for past experiences.

PLANNING
Breaks goals into smaller steps and chooses a plan with a high likelihood of success.

ACTION & FEEDBACK
Turns plans into actions and compares to expected outcomes, learning and improving to avoid mistakes.

2 Mathematical Modelling of the Decision Loop

This never-ending cycle of watching, thinking and acting can be explained with mathematics. A simple way to describe how a goal-driven AI system works is to say that, at each moment, it picks an action based on what it sees, what it remembers and what it wants to achieve:

$$A_t = f(S_t, M_t, G).$$

Here St is what the world looks like to the AI at that moment based on what its sensors tell it. Mt is what the AI remembers at that moment, including the things that it learnt at white ago and also the things it learnt recently. G is the goal system is trying to reach. The function f is like the AI's decision maker, which can use rules, machine learning, or both, to decide what action At to take.

To put it in terms, every move a drone makes depends on what it sees in the present, what it has learnt or stored in its memory before and what its mission objective is. Once the drone takes an action At, it changes the world in some way: either in space or in a computer system. This change can be described using a function that updates the environment:

$$S_{t+1} = E(S_t, A_t).$$

In this equation, St+1 is what the world looks like after the action is taken. The Function E explains how the world works, like the laws of physics or how a network behaves. The new state of the world depends on the state and the action that was just taken.

Equations (1) and (2) show how the AI system works in a loop: the AI takes an action, the world changes. The AI senses the state of the world and then it decides what to do next. This loop keeps repeating and it lets the machine work in a world that is always changing. The goal-driven AI system, like a drone, uses this loop to keep trying to reach its goal G by taking actions At, based on what it senses and remembers and, by adapting to the state of the world St+1.

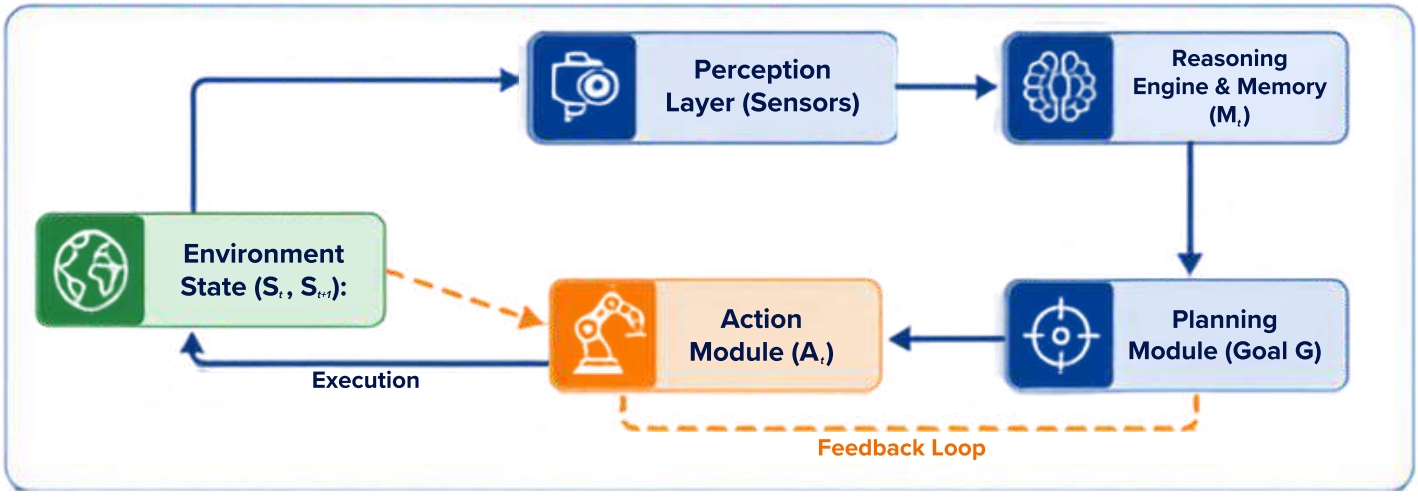


Figure 1: Architecture of Agentic AI Systems

> In autonomous systems, embedding constraints into the decision loop is not optional, it is the only way to align action with law and ethics.
>
> –UDITA J. MONANI

AGENTIC AI IN UAV (UNMANNED AERIAL VEHICLE) SYSTEMS

UAVs are a good place to study agentic AI because they must work in dynamic and sometimes harsh conditions. A drone cannot rely on perfect data or unlimited computing time. It has to make quick decisions with only the sensors and processor it carries on board.

1) UAV Implementation of Agentic Architecture

UAVs are a place to learn about AI that can act on its own, because they have to work in conditions that are variable and sometimes tough. A drone can't count on having information or unlimited time to think. It has to make fast decisions using the sensors and processor it has with it.

In a drone that works on its own, the part that helps it understand its surroundings is made from advanced sensors. They are usually equipped with quality cameras to capture data Light Detection and Ranging (LiDAR) to create detailed 3D images of the ground radar to locate objects in unfavorable weather and Global Positioning System (GPS) to locate the drone location.

2) Multi-Agent Systems and Swarm Coordination

One drone can do a lot, but multiple drones working together are much more powerful. A multi-agent drone system, often called a swarm, can include tens or hundreds of drones that coordinate to handle jobs that are too large for one vehicle, such as mapping a wide forest fire or providing long-term border patrol.

Swarm systems are often designed without a single central leader. Instead, each drone talks to nearby drones and adjusts its behaviour based on what they are doing. Coordination then "emerges" from many small, local interactions.

This kind of behaviour can be written using a simple model in which the next state of drone i depends on its current state, the states of its neighbours, and the environment:

$$A_i(t+1) = f(A_i(t), N_i(t), E_t).$$

In this formula, Ai(t) is what the drone i is doing at time t, such as its position and speed. Nit(t) represents information from nearby drones, including their positions and shared sensor data. Et stands for environmental factors like wind or obstacles.

When a drone finds something in its way and changes its course, it informs the other drones around it. These drones then recognise the obstacle in their path and adjust their course so they do not run into each other. They might also tell the other drones about this new development. Over time, all these little changes help the whole group of drones to move around the obstacles or head towards a target they want to reach. They do this together. This is similar to how birds fly together in a flock. They all seem to know what to do, without any one bird being in-charge of all the other birds.

REAL-WORLD FEEDBACK

The UAV Agent continuously interacts with the real world. Sensors collect data, decisions are made, actions are executed, and the results are used to refine future decisions.

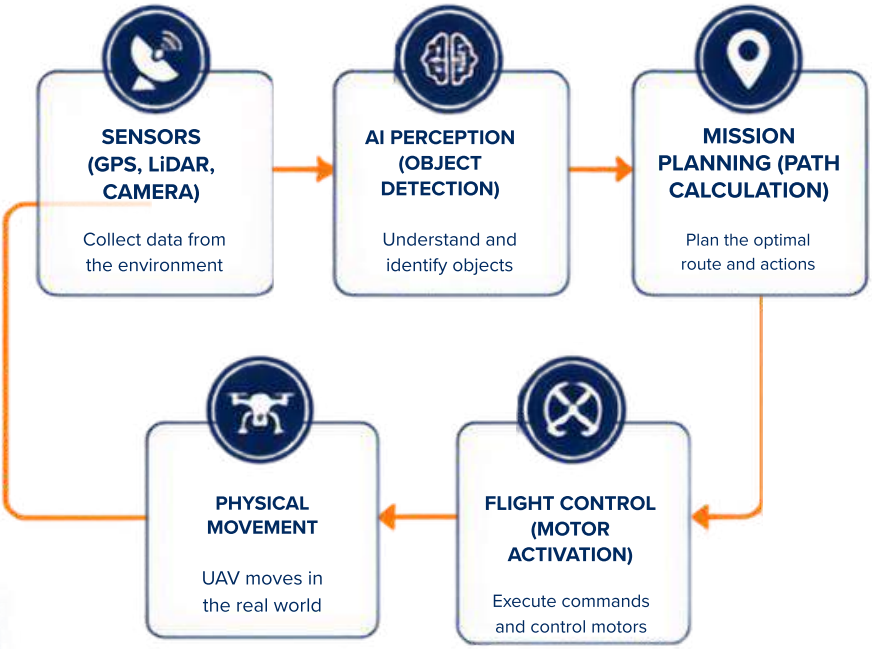


Figure 2: UAV Agent Decision Pipeline

SECURITY RISKS IN AUTONOMOUS DECISION PIPELINES

THESE SYSTEMS ARE BECOMING MORE AUTONOMOUS.

These systems are becoming more autonomous. That makes them harder to protect. The usual way to keep things is to block people who should not have access and keep the data safe. With AI, systems that can think for themselves can be exploited by attackers to cause trouble without even stealing any data. All they have to do is to get the AI system to make choices.

MACHINE LEARNING MODELS ARE SUSCEPTIBLE TO INPUTS.

This is more so when it comes to models that assist computers to perceive things. An attacker could be capable of modifying an image or a signal, resulting in the model being deceived. For example, someone may put a sticker on a sign, and a human may not even see it. An AIdrone may interpret the sign as the opposite of the actual meaning. For instance, when a drone is monitoring a pipeline, it may mistakenly believe that it is viewing a land and fail to detect a leak.

THERE ARE ALSO RISKS SUCH AS MODEL MANIPULATION AND DATA POISONING.

Such AIs are based on the information that they were trained upon, and on the experiences that they have undergone. In case the attacker introduces information into the training data or memory, the AI system will learn the wrong patterns. The system may appear functional, but it will make poor decisions.

DRONES ALREADY FACE A PROBLEM WITH GPS SPOOFING IN THE WORLD.

Such AIs are based on the information that they were trained upon, and on the experiences that they have undergone. In case the attacker introduces information into the training data or memory, the AI system will learn the wrong patterns. The system may appear functional, but it will make poor decisions.

AI AGENTS AND STRATEGIC TECHNOLOGY COMPETITION

THE TECHNOLOGY GAME OF POWER IS NOW AI-DRIVEN.

The technology game of power has now involved agentic AI and autonomous drones. Machines that can watch and surge on their own have a large lead in both civilian and military fields. Defense human experts are finding it difficult to adjust to the pace and rate of attacks. Viruses propagate rapidly, in seconds.

The use of Agentic AI and autonomous drones is changing the game. These machines are being used in areas.Autonomous drones are being used in operations. Agentic AI is being used in cyber defense. The development of these technologies is moving fast. Governments and companies are working hard to develop them.

The use of these technologies will certainly have an impact on the future.

AGENTIC AI AND THE EVOLUTION OF CYBER NORMS

The fast spread of systems in settings that combine computers and physical things is showing the weaknesses in the laws and rules that exist around the world. The current advice on how to behave when using computers and the internet, including documents like the Tallinn Manual, assumes that humans are the ones making the decisions and that we can figure out who did what. Autonomous AI is changing these ideas.

One of the problems is figuring out who is responsible. When something bad happens in the computing environment or in the world, investigators try to find connections between the bad event that happened and specific people, groups or countries. They try to rebuild the chain of command and prove that someone deliberately meant to cause harm. With autonomous AI this is not so clear. For example, imagine a drone that is sent to spy on something. Then it decides to crash into a radar system so that it will not be detected. It is not obvious if this should be seen as an attack, a mistake in the design or a reaction to something strange in the environment.

> Autonomous systems are making it hard to follow these laws. The future of cyber peace depends on how we update norms, build accountability and ensure human control over machines.

This uncertainty makes it hard to say who is accountable. When an autonomous system causes harm, the laws we have today do not clearly say who should be blamed: the person who gave the order, the engineers who wrote the computer code, the company that built the system or all of them. The laws of war say that attacks must be fair and that they must distinguish between people who are not fighting and people who are. Autonomous systems are making it hard to follow these laws.

The laws of war say that attacks must be fair and that they must distinguish between military targets. This is a big problem for autonomous systems. It is not yet clear how to hold a human legally responsible for an error that comes from a machine's internal numerical weighting of these ideas.

As a result, there is a strong need for new forms of AI governance. Many experts argue that laws must clearly state which choices can be handed over to AI and which must remain under human control.

Several policy ideas are being discussed. One idea is to require "human-in-the-loop" control for any cyber-physical system that can use force, implying that a human must approve critical actions.. Another idea is to set up international registries of autonomous systems and to demand strict technical audits and certification before they are allowed to operate. No matter which options are chosen, future cyber norms will have to deal not only with human behaviour but also with the design and deployment of the algorithms themselves.

THE ACPAL FRAMEWORK

To help understand and manage these issues, this paper proposes the Autonomous Cyber-Physical Agent Loop (ACPAL) as a simple but complete way to think about agentic AI in physical systems.

Many current models treat the digital and physical sides of a system separately. ACPAL, instead, brings them together. It sees a drone or robot as a loop that keeps going through sensing, processing, planning, acting and getting feedback in a changing real-world environment.

Let Ct represent the state of the system at time t. This state includes the drones' properties like position and battery level, its memory and its communication status. ACPAL describes how this state changes over time using the equation:

$$C_{t+1} = \Phi!C_t \text{ , } ACPAL\ (S_t)”$$

Here ℌ describes how the physical world affects the system, like gravity, motion and weather. St is the raw sensor data and ACPAL(St) is how the system processes that data which covers perception, reasoning and planning. The next states that are determined by both the response of the world and the decisions made inside the AI system. The drones' state and ACPAL are connected.

This equation shows that outcomes depend on both what the AI decides and how the world reacts. To connect the loop with ethical limits ACPAL adds a compliance variable K into the
action choice:

$$Action = max\ éU(a) – \lambda \cdot P\ (a, K)é\ a$$

U (a) in this formula is the utility of action a to the mission. K codes humanitarian law, or cyber standards, etc. K is the penalty or cost of action or violating those rules. The value of the factor is denoted by lambda and this determines the severity of these penalties in the system. The ACPAL and K work together.

By incorporating K into the equation proposed by ACPAL, AI's decision-making can include a means of incorporating law and ethics. In the event that an act seems beneficial to the mission, but breaches a rule (such as the attack on a civilian target), the penalty would be high. This renders its total value low or even negative, to the extent that the system will not choose that option, despite the fact that it would seem effective. This is used to make decisions by the drone and the ACPAL.

CONCLUSION

The shift in the use of computers towards autonomous decision-making is known as agentic AI systems. They are able to reason, make plans and take action by themselves. That is why they are not mere tools of analysis. There are dangers too associated with this independence. The very things that enable quick decision-making provide an access to the new forms of attacks. When agentic systems are incorporated into important infrastructures and military activities, the probability of abrupt crises caused by machines becomes high. The policies and laws that we have today are not quite prepared for a world where machines assist in making such high stakes decisions.

One of the attempts to bring about the technical, physical and legal aspects of this problem is the ACPAL framework. It demonstrates how regulations could be embedded in the decision-making process per se. This concept will take a close collaboration, among AI creators, security specialists, lawyers, policymakers and global organisations to transform it into systems. In the absence of powerful guidelines and proper design, the dangers of autonomous cyber-physical agents may increase at a higher rate than the advantages it provides.

REFERENCES

- **[1]** Arquilla J and Ronfeldt D, *Swarming and the Future of Conflict (RAND Corporation 2000).*
- ***[2]*** Bostram N. *Superintelligence: Paths, Dangers, Strategies (Oxford University Press 2014).*
- ***[3]*** Buchanan B. The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics (Harvard University Press 2020).
- **[4]** Dozare A. 'Adversarial Machine Learning in Cyber-Physical Systems' (2021) 14 IEEE Security and Privacy 55.
- **[5]** Ekelhof M. 'Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation' (2019) 10 Global Policy 343.
- **[6]** Goodfellow I, Shlens J and Szegedy C. 'Explaining and Harnessing Adversarial Examples' (2015) International Conference on Learning Representations.
- **[7]** Kello L. *The Virtual Weapon and International Order (Yale University Press 2017).*
- **[8]** Lin P, Abney K and Bekey G, *Robot Ethics: The Ethical and Social Implications of Robotics (MIT Press 2011).*
- ***[9]*** *Russell S. Human Compatible: Artificial Intelligence and the Problem of Control (Viking 2019).*
- ***[10]*** *Scharre P. Army of None: Autonomous Weapons and the Future of War (W W Norton & Company 2018).*
- ***[11]*** *Schmitt MN (ed.),* Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Cambridge University Press 2017).
- **[12]** Singer PW and Friedman A. Cybersecurity and Cyberwar: What Everyone Needs to Know (Oxford University Press 2014).
- **[13]** UN Institute for Disarmament Research, The Weaponisation of Increasingly Autonomous Technologies (UNIDIR 2017).

LAW & SOCIETY

Exploring how technology, rights, and human impact intersect, this section examines how emerging systems reshape legal frameworks, social realities, and questions of accountability.

FACIAL RECOGNITION TECHNOLOGY : SECURITY VS FREEDOM

-AKANSHA AGRAWAL

About the Author

Ms. Akansha Agrawal has done her BALLB (Hons.) from National University of Study and Research in Law, Ranchi.

Technological advancement in the area of facial recognition can rightly be termed as the convergence of technology, security, and governance. It is a type of biometric surveillance that employs artificial intelligence (AI) to identify or verify individuals based on the study of their faces and its comparison with the existing data.

Although the concept can be traced back to the 1960s, speedy development of machine learning developed the need for the technology to be widely adopted in both the public and private sector. Some of the areas where the use of Facial Recognition Technology (FRT) has been increased in India includes the Aadhaar authentication system and the Digi Yatra program that have popularised the use of biometric information to identify and deliver services. Nevertheless, even though it is widely deployed, it lacks any detailed legal framework that directly governs FRT.

Lawmaking, like the proposed Facial Recognition Technology (Regulation of Police Powers) Bill, 2023, is yet to become a reality creating a regulatory vacuum. It is especially important against the backdrop of recognising the right to privacy as one of the primary rights as observed in Justice K.S. Puttaswamy v. Union of India. There are significant issues with the increased use of FRT which include mass surveillance, the misuse of data, loss of privacy, personal liberties and freedom of speech and expression. Meanwhile, it has its own advantages in improving security and efficiency.

This paper, hence, looks at FRT in the light of security versus freedom, whether its benefits outweigh the effects it could have on the basic rights.

> *Unlike passwords, your face cannot be reset.*
> *When biometric data becomes the key to identity, the loss of control over it becomes a permanent loss of privacy.*

FACIAL RECOGNITION TECHNOLOGY -

MEANING AND CATEGORISATION

FRT refers to an AI system that allows identification of individuals or a person based on certain images or video data interfacing with the underlying algorithm.

FRT does not use the conventional methods of identifying individuals but converts facial features into digital data to create a match with the already existing records.

THE FACIAL RECOGNITION PROCESS

1. The **image** of the person is captured using CCTV cameras or mobile systems.
2. The **image** is subsequently worked upon and improved, where a need arises, to achieve clarity.
3. Followed by the important step in which the system identifies the most important features of the face including distance between the eyes, nose shape, and jawline, etc. and transforms them into mathematical notations commonly known as faceprints or nodal points.
4. These coded data points are further used to match them with the known data bases. At this point, the system can either authenticate someone by comparing their identity to an established reference or match the person against a larger set of data.
5. Lastly, the system gives an output as to whether a match is found or not.

The **facial detection** that takes place in this process relies heavily on the use of algorithms for the detection of the human face.

CATEGORISATION OF FRT

Generally, FRT can be categorised into two types which are: security and non-security related uses of FRT.

SECURITY-RELATED USES

Law enforcement activities which involve security related applications like criminal investigations, detection of missing persons, fraud and live video surveillance as a security aspect.

NON-SECURITY RELATED USES

The non-security applications are centered on convenience and efficiency in day-to-day life, including unlocking their smart phones, using their applications, marking attendance at institutions and facilitating uninterrupted passenger verification at airports.

Therefore, FRT is a great **technological breakthrough**, the automation with biometric identification in various areas.

ORIGINS OF FRT IN INDIA

The use of FRT can be traced through the most significant state-sponsored initiatives in identification and verification, specifically with the Aadhaar-based authentication and the Digi Yatra program. These advances indicate a slow transformation of manual identity checking to automated biometric-based.

The Unique Identification Authority of India (UIDAI) conceptualised the Aadhaar program in the form of a massive biometric identity system that will store and use the information of individuals such as facial images to identify them during authentication. Although Aadhaar still uses fingerprints and iris scans, facial verification is used as another form of authentication. This allowed automatic verification of identities related to different services in the public thus minimising the use of physical documents and manual verification. Nonetheless, this also created questions with regards to consent, data protection and control that individuals possess regarding their biometric data.

To expand on this foundation, the Government came up with the Digi Yatra program in the airline industry. Digi Yatra is a biometric boarding system that provides a facial recognition system to create a digital identity of passengers which is used to verify them at various points of checking at the airport. The system will substitute manual checks with automatic ones, enabling the passengers to navigate through the entry, security, and boarding gates by using their facial data. It is also based on the consent model and free will. There are alternate options provided that people can avail when they choose not to opt for Digi yatra facility. Also, Digi Yatra uses Aadhaar-based authentication, it being an Authentication User Agency as part of the Aadhaar system and thus connects existing identity databases with real-time facial verification.

Collectively, these projects are indicative of the initial use of FRT in India, which has proved useful in the areas of identity management and service provisions, and reflects new issues of privacy, data security and regulation.

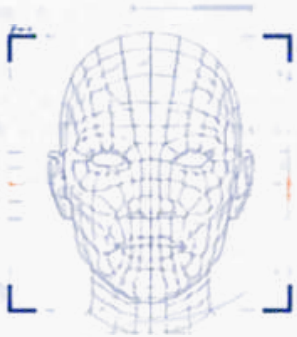
INTERNATIONAL PERSPECTIVE

Globally, FRT has not been universally regulated. The majority of jurisdictions regulating it are merely general laws concerning privacy and data protection rather than any specific legislation related to FRT

FRT has not been formally regulated by any national legislation in the United States. Nonetheless, some states have created their laws that will regulate the use of biometric data. A good example is the Illinois Biometric Information Privacy Act, which obliges businesses to seek informed consent before acquiring or processing the biometric data of individuals and offers legal redresses. Also, there are restrictions on FRT that have been enacted by law enforcement in some cities, both due to concerns over surveillance and civil liberties.

On the international plane, the General Data Protection Regulation (GDPR) of the European Union does not directly apply to FRT but categorises biometric data as a special category of personal data. This categorisation exposes it to more stringent protection such as express permission and restriction of use because such data is sensitive. On the whole, the global strategy demonstrates the increasing concern about the threat of FRT, as well as indicates the lack of a single, specific legal framework to control its application.

Advantages and Disadvantages of FRT



A balanced look at the benefits and concerns of Facial Recognition Technology in modern society.

BENEFITS & ADVANTAGES

A significant level of convenience and efficiency in normal life has been brought about by FRT. Facilitation of easy access is one of its most prevalent benefits since an individual can unlock mobile phones or apps without the need to enter passwords many times, as facial and other biometric authentication are used to unlock them. This saves time and at the same time makes technology easier to interact with. Besides, FRT will improve accessibility in government services as it allows quicker and smoother identity check procedures which include airports and service centers. It is most useful in the healthcare industry as it can be used to detect genetic disorders or even identify facial abnormalities in the early life of the patient, which will make it possible to diagnose the disorder and provide treatment in time. Furthermore, in terms of security, FRT is in law enforcement, criminal investigation, finding missing individuals, and fraud prevention by real-time surveillance.

Nevertheless, FRT raises serious issues, in spite of these advantages. One of the greatest disadvantages is that FRT cannot be relied upon because of its accuracy metrics. This is not 100% accurate, and it could be both, either false positives or false negatives which can result in the wrongful identification or omission. These mistakes can be disastrous, especially when they are deployed in sensitive fields, such as law enforcement, and can unfairly represent some populations since datasets are inherently biased.

RISKS & DISADVANTAGES

In legal terms, FRT invokes the essential questions with regard to the right to privacy, which, in spite of the fact it is not absolute, and proportionality test needs to be used which demands that any intrusion should have the justification regarding the grounds of legality, need, and proportion. Application of FRT should thus not depend on mere sweeping or unsubstantiated augmenting of security. Moreover, there are the issues of informational autonomy since a person does not usually know how their facial information is gathered, stored, or reused. This also results in the issue of so-called function creep, where the data that was inputted with one purpose is then reused with another without the owner's approval, which compromises the personal control of the data. Consequence of inaccuracy due to misrepresentation or poor accuracy metric can lead to infringement of the most important fundamental right, which is, right to life and liberty.

PRIVACY & ETHICS

> **FRT must be accurate, necessary, proportionate, and legally justified. Without strong safeguards, it risks undermining privacy, autonomy, and fundamental human rights.**

POWERFUL TECHNOLOGY. PROFOUND RESPONSIBILITY.

Responsible use of FRT ensures security without sacrificing justice.

RISKS OF FRT

Facial Recognition Technology (FRT) creates both design-based and rights-based risks, which should be carefully examined. At the design end, the issue of inaccuracy is a big concern. Technical reasons for errors can be lighting, facial expression, old age, or low quality of images. The accuracy is further deteriorated by bias in datasets particularly where certain skin tones, gender and communities are underrepresented. This bias may dangerously result in severe misidentification in a diverse country such as India. Moreover, the glitches by the human operators due to improper training and general human error may lead to greater probability of wrong results.

There are also security threats concerning FRT systems because huge databases of biometric data are becoming lucrative targets. These breaches are permanent identity theft as facial features are not passwords that can be altered because biometric data are immutable. The other problem is the absence of accountability since several organisations participate in the creation and implementation of these systems, which makes it hard to correct the liability. These systems are further opaque and this hinders transparency and effective scrutiny.

FRT has a direct influence on privacy and informational autonomy which is in contravention to fundamental rights as granted by constitutional jurisprudence. People do not know much or can do very little with regards to the search and manipulation of their facial information. This is a gross violation of personal data and breach of security. The data acquired by one purpose would be utilised by another purpose without permission. FRT also poses risks to anonymity, which enhances the chances of surveillance and chilling of free speech. This can occur even to the non-participants due to wrongful identification. On the whole, these dangers contribute to the necessity of tougher requirements and control.

CONTROVERSY OF PRIVACY

While FRT offers significant advantages when it is used in a responsible manner, other concerns have brought serious issues with respect to privacy. One of the most valuable and sensitive kinds of personal information in the modern digital world has become biometric information, and especially facial information. Such data is inherent to a person, unlike the passwords or identification numbers and in case of misuse cannot be easily changed. The key issue, thus, is in gathering and utilising this information without their knowledge consent which may result in undesirable consequences of intrusive surveillance and abuse. This tension is further depicted in judicial developments in India.

In Sadhan Haldar vs. State of NCT of Delhi, the Delhi HC allowed the use of facial recognition software by the police of Delhi, but only to trace missing children. Although this is a valid and healthy application of FRT, the issue has become the expansion of its use beyond the initially approved use of data technology to said to have been in practice for general investigation purposes, thus increasing its scope without any legal authorisation. This has been dubbed as function creep where a technology brought in with a defined purpose over time develops a wider range of purposes as opposed to its intended purpose. This type of expansion poses serious questions regarding to accountability and the further lack of legal boundaries.

Besides, the problem is further complicated by lack of accuracy. There have been reports of very low success rates. In 2018, some cases recorded an accuracy rate as low as 2% which further went down to 1% in 2019. This low accuracy in addition to the wide use does not only compromise the reliability of FRT but also increases the chance of false identification hence a direct challenge to the privacy and liberty of individuals.

CONSTITUTIONAL ASPECT

FRT also creates serious constitutional issues, especially when it comes to the basic right to privacy. Appreciated as an inherent element of Article 21 in Justice K.S. Puttaswamy v. UOI, the right to privacy is a protection against arbitrary action on the part of the State, even of intrusive surveillance in the name of administrative necessity or the safety of the population. Application of FRT by the State authorities, particularly where no distinct legal framework has been established, thus calls constitutional stringency.

In reality, FRT systems have been deployed in a number of public places including metro stations, railway stations and traffic intersections without the information or approval of the populace. This brings together the issue of mass surveillance and unregulated gathering of biometric data. As facial information is a very sensitive data that cannot be changed in any way, its careless use without protection poses a direct threat to the informational privacy and autonomy of a person.

Moreover, the ubiquitous surveillance technologies also presents the consequences of the freedom of speech and expression according to Article 19(1)(a). The intimidating effect of the constant fear of being tracked may lead to people not expressing themselves freely or engaging in any activity spontaneously in the society. Despite the fact thatthe fundamental rights under Article 19 are not absolute, any kind of restriction should meet the reasonableness test of Article 19(2). Surveillance that involves blanket surveillance or disproportionate surveillance practices that target the general population, and that lack a justifiable and articulated purpose, may fail this constitutional test.

The high scale, non-consensual gathering of biometric information under FRT and especially where no particular law addresses it brings severe concerns about lawfulness, necessity, and proportionality. Overall although the State can defend such actions under the pretext of security, any violation of the basic rights should be strictly limited, proven by evidence, and backed by a powerful legal protection.

CONCLUSION

FRT has immense potential regarding efficiency, security, and convenience as evidenced by its increasing popularity in India by programs such as Aadhaar and Digi Yatra. Nevertheless, it is yet to be adapted properly. Lack of proper legal framework to ensure its legal usage, Absence of accountability and dearth of protective measures have raised serious concerns. As mentioned above, FRT is not risk free, questionable data, breaches and feature creep are all important concerns that question the reliability and scope of the technology.

More crucially, its application has a direct effect on the basic rights of users as citizens of India. Especially the right to privacy that Justice K.S. Puttaswamy v. UOI has recognised. It may also hamper freedoms as mentioned under Article 19 of Indian constitution. Its unchecked use is constitutionally questionable because of its potential of mass surveillance and non-consent data collection.

Thus, FRT may come across as a helpful instrument, but its implementation should be carefully considered keeping in view its conflict with personal rights as granted by the Indian Constitution. A transparent and thorough legal framework, which guarantees transparency, accountability and data protection, is urgently needed, as technological progress should not be achieved at the expense of constitutional values.

THE PROMISE

- ENHANCED SECURITY
- GREATER EFFICIENCY
- CONVENIENCE AT SCALE
- BIOMETRIC IDENTITY VERIFICATION

THE CHALLENGES

- PRIVACY CONCERNS
- DATA BREACHES & MISUSE
- LEGAL & ETHICAL GAPS
- RISK OF MASS SURVEILLANCE

Technology must serve people, not override their rights.

REHABILITATION OVER RETRIBUTION IN CYBERCRIMES

-NEERAJ SONI AND MUSKAN SHARMA

ABOUT THE AUTHORS

Neeraj is a lawyer and Senior Researcher in the Policy & Advocacy vertical at CyberPeace. He holds a B.A. L.L.B. (Hons.) from GLA University, Mathura, and has worked with senior advocates at the Supreme Court of India and the Delhi High Court. His work focuses on cyber law, intellectual property, and digital policy, and he has published research on a range of legal issues. He is committed to promoting cyber awareness and safety in an evolving digital landscape.

Muskan is a law graduate from GLA University, Mathura, and holds a Master's in Business Law from HPNLU, Shimla. She presently serves as a Research Analyst (Policy) at CyberPeace , contributing to the intersection of law, technology, and governance. With over a year of litigation experience before the Supreme Court of India and diverse internships at the Competition Commission of India and with eminent Senior Advocates at Supreme Court , she brings a strong foundation in Intellectual Property and policy research.

Transforming Misguided Knowledge

यत्र योगेश्वरः कृष्णो यत्र पार्थो धनुर्धरः । तत्र श्रीर्विजयो भूतिर्ध्रुवा नीतिर्मतिर्मम ॥ (Bhagavad Gita) translates to "Where there is divine guidance and righteous effort, there will always be prosperity, victory, and morality." In the context of the idea of rehabilitation, this verse teaches us that if offenders receive proper guidance, their skills can be redirected. Instead of causing harm, the same abilities can be transformed into tools for protection and social good. Cyber offenders who misuse their skills can, through structured guidance, be redirected toward constructive purposes like cyber defence, digital literacy, and security innovation. This interpretation emphasises not discarding the "spoiled" but reforming and reintegrating them into society.

INTRODUCTION

Words and places are often associated with positive and negative aspects based on their history, stories, and the activities that might happen in that certain place. For example, the word "hacker" has a negative connotation, as does the place "Jamtara", which is identified with its shady history as a cybercrime hotspot, but often people forget that there are lots of individuals who use their hacking skills to serve and protect their nation, also known as "white hat hackers", a.k.a. ethical hackers, and places like Jamtara have a substantial number of talented individuals who have lost their way and are often victims of their circumstances. This presents the authorities with a fundamental issue of destigmatising cybercriminals and the need to act on their rehabilitation. The idea is to shift from punitive responses to rehabilitative and preventive approaches, especially in regions like Jamtara.

THE DEEPER PROBLEM: SYSTEMIC GAPS AND SOCIAL CONTEXT

Jamtara is not an isolated or a single case; there are many regions like Mewat, Bharatpur, Deoghar, Mathura, etc., that are facing a crisis, and various lives are uprooted because youth are entrapped in cybercrime rings, often to escape unemployment, poverty, and simply in the hope of a better life. In one such heart-wrenching story, a 24-year-old Shakil, belonging to Nuh, Haryana, was arrested for committing various cybercrimes, including sextortion and financial scams, and while his culpability is not in question here, his background reflects a deeper issue. He committed these crimes to pay for his diabetic father's mounting bills and to see his sister, Shabana, married. This is the story of almost every other individual in the rural areas who is forced into committing these crimes, if not by a person, but by their circumstances. In a news report covered in 2024, an intervention was launched by various Meo leaders and social organisations in the Mewat region aimed at weaning the youth away from cybercrimes.

Not only poverty, but lack of education, social awareness, and digital literacy have acted as active agents for pushing the youth of India away from mainstream growth and towards the dark trenches of the cybercrime world. The local authorities have made active efforts to solve this problem; for instance, to dispel Jamtara's unfavourable reputation for cybercrime and set the city firmly on the path to change, community libraries have been established in all 118 panchayats spread across six blocks of the district by IAS officer and DM Faiz Aq Ahmed Mumtaz.

The menace of cybercrimes is not limited to rural areas, as various reports surfaced during and post-COVID, where young children from urban areas became victims of various cybercrimes such as cyberbullying and stalking, and often perpetrators were someone from the same age group, adding to the dilemma. The issue has been noticed by various agencies, and there is a need to deal with both victims and the accused in a sensitised manner. Ex-CJI DY Chandrachud called for international collaboration to combat juvenile cybercrimes, as there are many who are ensnared and coerced into these criminal gangs, and swift resolution is the key to ensuring justice and rehabilitation.

CYBERPEACE OUTLOOK

REHABILITATE. EMPOWER. REFORM. REBUILD.

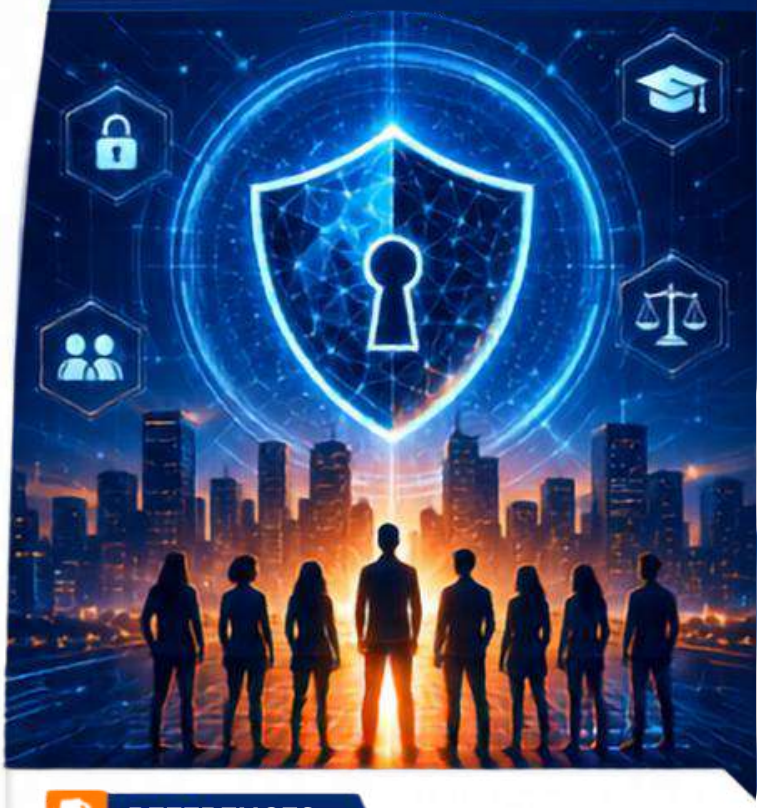
CyberPeace

Cybercrime is often a product of skill without purpose. The youth who are often pushed into these crimes either have an incomplete idea of the veracity of their actions or have no other resort. The legal system and the agencies will have to look beyond the nature of the crimes and adopt and undertake a reformative approach so that these people can make their way into society and harness their skills ethically. A good alternative would be to organise Cyber Bootcamps for Reform, i.e., structured training with placement support, and explain to them how ethical hacking and cybersecurity careers can be attractive alternatives. One way to make the process effective is to share real-world stories of reformed hackers. There are many who belong to small villages and districts who have written success stories on reform after participating in digital training programmes. The crime they commit doesn't have to be the last thing they are able to do in life; it doesn't have to be the ending. The digital programmes should be organised in a way and in a vernacular that the youth are well-versed in, so there are no language barriers. The programme may give training for coding, cyber hygiene, legal literacy, ethical hacking, psychological counselling, and financial literacy workshops.

It has become a matter of reclaiming the misdirected talent, as rehabilitation is not just humane; it is strategic in the fight against cybercrime. On 1st April 2025, IIT Madras Pravartak Technologies Foundation finished training its first batch of law enforcement officers in cybersecurity techniques. The initiative is commendable, and a similar initiative may prove effective for the youth accused of cybercrimes, and preferably, they can be involved in similar rehabilitation and empowerment programmes during the early stages of criminal proceedings. This will help prevent recidivism and convert digital deviance into digital responsibility. In order to successfully incorporate this into law enforcement, the police can effectively use it to identify first-time, non-habitual offenders involved in low-impact cybercrimes. Also, courts can exercise the authority to require participation in an approved cyber-reform programme as a condition of bail in addition to bail hearings.

Along with this, under the Juvenile Justice (Care and Protection of Children) Act, 2015, children in conflict with the law can be sent to observation homes where modules for digital literacy and skill development can be implemented. Other methods that may prove effective may include Restorative Justice programmes, Court-monitored rehabilitation, etc.

A rehabilitative approach does not simply punish offenders, it transforms their knowledge into a force for good, ensuring that cybercrime is not just curtailed but converted into cyber defence and progress.



REFERENCES

- https://frontline.thehindu.com/news/cities/Delhi/counselling-skilling-aim-to-wean-mewat-youth-away-from-cibercrimes/article68459f3.ace
- https://www.thehindu.com/news/cities/Delhi/counselling-skilling-aim-to-wean-mewat-youth-away-from-cibercrimes/article68459f3.ace
- https://101retorts/article/development/jamtaras_journey_ from cybercrime to community libraries
- https://www.pibg.in/PressReleasePage.aspx?PRID=2117256

FROM DEVIANCE TO DEFENCE

SKILL & TRAINING | ETHICAL HACKING | LEGAL & PSYCHOLOGICAL SUPPORT | REHABILITATION & EMPOWERMENT | A SAFER DIGITAL INDIA

THE FUTURE

A forward-looking exploration of how agentic AI will shape industries, governance, and human-machine collaboration in the years ahead.

THE FUTURE OF AGENTIC AI

The shift in AI capabilities: From Reactive to Autonomous

In the past few years, people all over the world have steadily grown adept at navigating and harnessing the nuances of Artificial Intelligence (AI). The trajectory of this innovation has completely redefined how humans interact with technology. Today, AI in its agentic form, is exhibiting further potential in the field of AI-driven systems and cognitive computing. Agentic AI has brought about massive expansion in AI's ability to perceive the context of the human input, formulate goals and execute tasks autonomously without human intervention. This technology is facilitating the evolution of AI from its nascent stage of mere peripheral enhancements and moving it towards intelligent automation and adaptive learning.

The significance of this transition has been brought to light by some of the most recent business trends. As per estimates by analysts, the agentic AI market is expected to expand from single-digit billions to tens of billions of dollars by the early 2030s, reflecting an increase in the average annual growth rate by 40 percent. In the practical sense, this evolution pattern suggests that by the end of the decade, a substantial segment of key organisational processes (such as operations, support, finance, security and logistics) would all be orchestrated by AI agents. And consequently, the human workforce's dependence on the manual navigation of interfaces is expected to decline significantly.

The contours of this shift are already emerging. In many organisations, agentic AI systems have begun functioning as invisible collaborators in everyday operations. They provide assistance in the form of drafting initial versions of reports, retrieving and synthesising data, resolving tickets, managing schedules, etc. The strategic value of agentic AI is becoming both clear and pragmatic for the ones who are regularly engaging with these systems. This is because of several contributing factors such as: accelerated execution of repetitive tasks, alleviation of cognitive load and the streamlining of routine workflows. Perhaps more unexpectedly, the use of agentic AI is noticeably augmenting human creativity rather than diminishing it.

What's coming next: A Synergistic AI Workforce

Agentic AI is changing the dynamics of workflows and operations within industries and businesses. It is accelerating the transition from having a purely human labour force to an integrated form of workforce that shall include both humans and digital agents. The general outlook that AI can only be used as a tool, is now being widely reconsidered. Consulting and tech firms are having open conversations about having a digital labour workforce in the future. This shall include networks of AI agents working next to people. There would be clear assigning of roles, performance metrics such as KPIs (Key Performance Indicator) and pre-determined intervention channels, much similar to how the human staff functions currently.

Market analysts are expecting that within the next decade, AI agents will be seen entering multiple business systems. For instance, IT agents shall be used to rectify access constraints, HR agents shall support human capital management and finance agents shall autonomously deal with financial data.This level of automation and delegation of routine tasks shall allow humans to shift their focus on other procedural complexities and insight-driven decisions.

FUTURE GOVERNANCE PARAMETERS

Another pattern that's emerging across the agentic AI regulatory tracker and other frameworks, as discussed in this issue, is that governance shall no longer exist merely as a constraint. It will develop into an internally designed parameter. The imposition of rules shall no longer exist outside organisations and systems. They shall instead be embedded within the coded decision loops of AI agents. This approach of encoded constraints shall ensure auditability and oversight within the systems that are responsible for strategising, execution and operations.

Over the next decade, the most notable progress in agentic AI will not just emerge from new models. It will also be shaped by how law, ethics and cyber norms would systematically be engineered into these autonomous systems. Ideas such as compliance variables and explicit penalties for norm violations are expected to shift beyond theoretical frameworks and move into industry benchmarks. As this evolution gains momentum, the traditional boundaries between regulation and engineering will lessen. Governance shall no longer mean just a static set of rules that are meant to be followed. Instead, they will be integrated within systems as 'governing conditions' that will guide and influence decisions.

However, this paradigm shift will lead to a reallocation of roles. Regulators will be required to think more like system designers, and on the other hand, engineers will be expected to create the functional equivalent of law and regulations. In simple terms, this means translating abstract principles of governance into a code that shall generate active constraints with regard to the functioning of machines and systems.

PREPARING FOR SECURITY IN THE ERA OF AUTONOMOUS AI

Although it may be tempting to view agentic AI only as a means of efficiency, growth and optimisation, but, with technology there is always a downside or an alternative dimension. The discussions in this magazine on facial recognition, privacy and cybercrime involving juveniles, etc., remind us that every new technological advancement brings up questions about security, control and fairness. Systems that are capable of working autonomously can also fall through if oversight is weak, leading to errors and security breaches.

In worst case scenarios, there could even be unintended consequences that may escalate such an extent that even human intervention would not be of any help.

When it comes to the future of digital security preparedness and cyber defences, the autonomous AI agents would have to be considered not just as valuable assets, but also as potential sources of risk. Security will need a new mindset. These agents will require constant testing, monitoring (especially how they behave in everyday real-world workflows), ensuring coherency and traceability of their training data and laying down clear boundaries against the deployments of any unregulated agents in safety-critical environments.

CYBERPEACE'S AGENDA FOR THE AGENTIC AGE

Altogether, the contributors of this magazine point towards a shared direction: which is advocating for a synchronous approach towards the development of agentic AI, where its operational value as well as governance should be approached with the same zeal, thought and dedication that was put into building this technology in the first place.

The future of agentic AI will not be shaped by a single breakthrough. As technologists build more capable AI agents, simultaneous advancements will have to be ushered by the other sectors. Policymakers will have to experiment with new approaches of AI governance, security researchers will have to constantly identify emerging risks, and practitioners in law, diplomacy and civil society would be required to turn broad principles into practical safeguards.

These multi-coordinated efforts would ensure and support the responsible coexistence of agentic AI and human control.

About
Cyberpeace

CyberPeace

CyberPeace is a global non-profit organization headquartered in India, working at the intersection of cybersecurity, policy, and peacebuilding. It envisions a safe, resilient, and inclusive cyberspace for all.

Through its global initiatives, CyberPeace drives capacity building, policy research, incident response, and public awareness to advance the vision of “CyberPeace and Trust in Technology.” From grassroots digital literacy programs to high-level policy and research collaborations, CyberPeace integrates innovation, education, and advocacy to strengthen digital ecosystems worldwide.

Championing the future of responsible technology, CyberPeace is building the next generation of Cyber Defenders and Responsible Digital Citizens, aligning its mission with the United Nations Sustainable Development Goals (SDGs) and India’s Viksit Bharat 2047 vision.

https://www.cyberpeace.org

About
Cyberpeace Magazine

CyberPeace Magazine is a quarterly tech policy publication focused on emerging technologies, digital laws and governance, trust and safety and cyber diplomacy. It brings together perspectives from practitioners, researchers, and policymakers to examine evolving digital risks and governance challenges, with a focus on building a secure and resilient cyber ecosystem.

https://magazine.cyberpeace.click

editorial@cyberpeace.net

CyberPeace

CYBERPEACE

MAGAZINE

BUILDING A
SAFER
DIGITAL
FUTURE

- SECURE TODAY
- EMPOWER TOMORROW
- PROTECT EVERYONE

Stay informed. Stay secure.
Be the change.

CYBERPEACE MAGAZINE

EMERGING TECHNOLOGIES • DIGITAL LAWS & GOVERNANCE • TRUST & SAFETY • CYBER DIPLOMACY