ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Building Trustworthy Autonomous AI: Essential Principles beyond Traditional Software Design

Ronil Christian | Computer Science, Illinois Institute of Technology, Chicago, IL, USA | ORCID: 0009-0002-7175-6799

Durgesh Babu P | Artificial Intelligence and Machine Learning, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India | ORCID: 0009-0007-9342-3908

Hrishitva Patel | Information Systems, University of Texas at San Antonio, TX, USA | ORCID:0000-0001-7887-6641

Keyur Modi | Computer Science, University of Texas at San Antonio, TX, USA | ORCID: 0009-0008-9242-6098

Abstract

Imagine smart Artificial Intelligence (AI) agents that can act on their own, like digital teammates, needing our complete trust, especially in protecting our digital world. Just as early software was chaotic until ideas like 'object-oriented programming' (OOP) brought order, today's powerful AI agents are growing incredibly complex and can be unpredictable. We're building them so rapidly that clear rules for their trustworthy design are still emerging. Our paper proposes five core 'building blocks' or principles for designing these independent AI systems: making them explainable (understanding their decisions), adaptable (learning and evolving safely), collaborative (working together securely), resilient (defending against attacks), and ethical by design (acting responsibly). We examine how current AI frameworks like LangChain, AutoGen, and LlamaIndex are starting to implement these ideas, for instance, by

Received: 24.02.2025

Accepted: 28.07.2025

Published: 03-09-2025

Cite this article as:
R. Christian, D. Babu P, H. Patel, K. Modi, "Building Trustworthy Autonomous Artificial Intelligence: Essential Principles Beyond Traditional Software Design," ACIG, vol. 4, no. 1, 2025, doi: 10.60097/ACIG/208710.

Corresponding author:
Hrishitva Patel, Information Systems, University of Texas at San Antonio, TX, United States; E-mail: hrishitva.patel@utsa.edu
0000-0001-7887-6641

integrating real-time threat data or enabling structured team interactions for cybersecurity. We also highlight the tough challenges that remain, such as fully explaining AI's internal reasoning and ensuring its inherent robustness against clever manipulations. We conclude by emphasising that a collective effort from auditors, lawmakers, scientists, and industry leaders is crucial to establish these principles and build truly trustworthy autonomous AI.

Keywords

autonomous AI, cybersecurity, object-oriented programming, foundational architectural principles, agentic AI frameworks

1. Introduction: Architecting Trust in the Age of Autonomous Digital Sentinels

The digital age, a realm of unprecedented connectivity and escalating cyber threats, continually demands new approaches to managing complexity. To understand the future of Artificial Intelligence (AI), especially its role in safeguarding our digital world, we must first appreciate the timeless wisdom of foundational design.

1.1. The Genesis of Order: Lessons from Early Computing

Imagine the late 1800s, long before computers filled our homes. A group of very smart thinkers came together, deep in thought, trying to solve a huge problem: early computer programs were a tangled mess. Think of it like building a big machine where every single wire was exposed, twisted with every other wire. One wrong touch, and the whole thing could break. It was pure chaos, and they desperately needed a way to bring order to it. But how do you even begin to design order from such a digital wilderness?

Then, these thinkers had brilliant ideas that changed everything. They called one encapsulation[1]. Imagine, instead of a jumble of wires, they started putting all the related parts of a program into neat, self-contained boxes. It's like the cockpit of a ship – all the complex controls are hidden inside. You can steer the ship, but you won't accidentally pull out a vital connection. This made programs much cleaner and far less likely to break. Didn't they see the elegant simplicity in hiding complexity?

Next, they thought: why build everything from scratch every single time? This led to inheritance [1]. Imagine you've perfectly designed

a basic 'ship'. Now, if you need a 'freighter', you don't have to draw every line again. The freighter can simply inherit all the basic qualities of a ship, and then you just add the special cargo holds it needs. This was a huge shortcut, saving massive amounts of work. What genius thought of building upon existing wisdom instead of always starting anew?

Their vision kept growing. They also laid the groundwork for polymorphism [2]. This was like giving a single command, 'sail', to different types of ships, but having each ship understand it in its own unique way. A speedy boat would 'sail' by zipping across the water. A huge tanker would 'sail' with a slow, powerful chug. The same instruction, but a tailored action for each vessel. This made their digital 'fleets' incredibly flexible. How profound is it to give a single instruction myriad interpretations, yet maintain perfect control?

And finally, to make complex things simple, abstraction [2] emerged. This was about focusing on what an object does, instead of getting lost in every tiny detail of how it does it. When you tell a ship to 'sail', you care that it moves from point A to point B. You don't need to understand every tiny detail of its engines at that moment. This lets designers work on bigger ideas, hiding overwhelming complexity behind simple functions. Could they have foreseen how crucial simplicity would be in an increasingly complex world?

These four revolutionary ideas, which are encapsulation, inheritance, polymorphism, and abstraction, became the core of object-oriented programming (OOP) [1, 2]. For over a century, their wisdom has shaped how we build computer programs. Languages like C++ and Java, which are used everywhere today, respect and use these principles. OOP helped developers build systems not as fragile, sprawling messes, but as connected, strong, and elegant collections of self-sufficient 'objects', which is a true sign of timeless design.

Now, fast forward to the rise of AI. As AI systems became more complex, needing to handle vast amounts of data and learn intricate patterns, developers naturally turned to OOP. They found that these century-old design principles were perfectly suited for building AI. Encapsulation allowed them to create self-contained AI components like 'image classifiers', keeping complex algorithms private. Inheritance lets them build new AI models from the existing ones, like creating a specialised 'neural network' from a basic blueprint. Polymorphism enabled different AI 'predictor' objects

to use a single command like 'predict()' in their own unique ways. Abstraction became vital for managing the overwhelming details, letting AI engineers focus on what a model 'translates text', rather than getting lost in its immense internal workings.

1.2. Historical Analogy and Motivation

The field of agentic AI now faces an inflection point similar to when the OOP in the 1980s introduced foundational software design principles to manage increasing software complexity. The emergence of agentic AI frameworks has enabled large language model (LLM) agents to form increasingly dynamic and goal-driven workflows. However, these lack formal constraints, transparent decision traces, and ethical safeguards.

It has become more than necessary to have an agreed-upon set of principles that constrain and inform the design of agent behaviour to minimise the risks of opaque, brittle, and potentially harmful autonomous systems.

Here, we suggest not a one-to-one mapping but we draw more precise architectural analogies based on shared challenges and solutions in software engineering, as shown in below table 1.

1.3. The New Frontier: Agentic AI and Its Architectural Echoes

However, the digital world continues to evolve at breathtaking speed. A new frontier is emerging within AI itself: agentic AI frameworks. These aren't just programs; they are like smart digital beings that act autonomously on input goals, and capable of amazing independence. They promise to change industries and save countless hours. Yet, because they are so independent and

Table 1. Analogy of Foundational Principles in OOP and Agentic AI Design.

OOP concept	Agentic AI counterpart	Shared challenge addressed
Encapsulation	Agent role-bound tool access and internal memory	Limits exposure, enhances control
Modularity	Separation of agents by task/goal	Decouples responsibilities
Polymorphism	LLM agents adapting behaviour via prompt engineering	Task generality across different inputs
Inheritance	Agent archetypes extended by new agents	Reusability of behaviours or tools
Message Passing	Agent-to-agent communication and memory sharing	Enables interaction across components

complex, these agentic AI systems present architectural challenges strikingly similar to the messy programs those early thinkers faced long ago. If AI agents are the next evolution, how do we ensure they are built on foundations that will not crumble under their own emergent intelligence?

This paper argues that the deep lessons forged in that distant past, the enduring power of encapsulation for agent independence, inheritance for adaptable intelligence, polymorphism for flexible interactions, and abstraction for managing complex behaviours which are not just useful but absolutely vital for designing strong, scalable, and smart agentic AI systems today and in the future.

1.4. AI's Critical Role and Urgent Challenges in Cybersecurity

As AI systems play a crucial role in cybersecurity, with AI tools integrating LLMs into threat detection and analysis, the need for trustworthiness becomes ever more important. Just as OOP had great success with following the practices of well-defined principles, future AI systems in cybersecurity must be built on a strong foundation of principles from software design, AI ethics, and cybersecurity best practices.

Artificial intelligence, particularly through machine learning (MI) and language models, plays an increasingly vital role in defending against evolving cyber threats. Here's how:

- Anomaly detection: AI can identify unusual patterns in network traffic or system behaviour faster than traditional tools.
- Threat intelligence: LLMs, when integrated using AI frameworks, can summarise threat reports, categorise vulnerabilities, or even auto-respond to routine queries.
- Phishing and malware detection: Natural language processing (NLP) is used to spot malicious email patterns, social engineering lures, or code injections.
- Automation of repetitive tasks: AI can automate ticket triaging, compliance checks, and documentation – freeing human analysts for high-priority tasks.

Despite its promise, there are several fundamental areas with room for improvement for AI in cybersecurity:

- Hallucination and inaccuracy: Language models can generate plausible but false statements which can be costly in a high-stakes environment like cybersecurity.

- Data sensitivity and leakage: AI systems that are trained on or exposed to security logs, customer data, or attack vectors pose privacy risks, and can potentially leak sensitive content to third parties.
- Bias and blind spots: Models can fail to generalise or react appropriately to novel attacks in case the training data lacks diversity or overlooks rare threats.
- Explainability and regulation: LLMs still remain 'black-boxes', especially in multi-stage pipelines, which limit their trustworthiness in legal or regulatory contexts. Regulatory bodies, now more than ever, are demanding more transparent AI systems.
- Integration complexity: Most organisations have legacy systems and strict compliance rules, which makes it harder to integrate modular AI systems. This requires careful architecture and policy design.

1.5. Forging Trust: Foundational Principles for Secure AI

To build AI systems that can be trusted in critical environments like cybersecurity, the following principles should be considered and deeply embedded in their architecture:

- Modular and explainable design: Inspired by OOP, future AI systems should be modular and explainable by design. Each module (e.g. input preprocessing, context retrieval, and generation) should be auditable independently, similar to encapsulation.
- Embedded ethical protocols: Researchers and developers should encode ethical constraints as well, extending the principle of abstraction from OOP: hiding risky functionality behind safeguards.
- Adversarial testing and robustness benchmarks: Following the principle of defence-in-depth, AI systems should be routinely tested against adversarial attacks, real-world threats, and unseen exploit patterns.
- Hybrid oversight models: Keeping humans in control, AI systems should act as copilots, not autopilots. High-risk outputs should be reviewed manually to balance speed with accountability.

2. Literature Review

Human curiosity has played an important role in the journey towards today's technologies. Our desire to communicate, solve problems, and make sense of the world around us has led humans to the development of the most advanced AI systems.

2.1. Ideas That Built the Digital World

Our human desire to communicate and solve problems has always pushed technology forward. Long before computers, information spread through storytelling, then writing, and later by the printing press [3]. By the 19th century, technologies like the telegraph and telephone connected the world faster than ever.

A huge step came in the 20th century when we learned to send information electronically. Claude Shannon's groundbreaking 1948 paper showed that any information such as words, numbers, and pictures can be turned into simple binary code (0s and 1s) and sent precisely [4]. This simple idea paved the way for all digital communication and, eventually, for thinking machines.

2.2. The Dawn and Evolution of AI

The very idea of AI truly began when Alan Turing, in his 1950 paper, asked a provocative question: 'Can machines think?' [5]. This sparked intense curiosity, leading to the 1956 Dartmouth Conference, where the term 'Artificial Intelligence' was formally coined. Researchers there believed that machines could, in time, learn and reason like us [6]. Early AI focused on teaching computers specific rules to follow, like solving a logic puzzle [7].

However, these early rule-based systems often struggled with real-world complexities, leading to periods where AI progress seemed slow. The big change came with machine learning, where instead of being told what to do, computers learned on their own by finding patterns in massive amounts of data [8]. This breakthrough enabled computers to recognise faces, understand voices, and translate languages. A major leap occurred in 2017 with the transformer architecture [9]. This new design allowed AI models to process entire sentences at once, understanding context much better. This quickly led to today's powerful LLMs like GPT, which can generate human-like text and hold complex conversations [10]. This transformation made AI incredibly accessible, allowing us to simply ask questions and get sophisticated results.

2.3. Cybersecurity: A Growing Digital War and AI's Mixed Blessing

The moment computers started connecting, the need to protect them was born. The very first self-spreading computer program, the 'creeper virus' in 1971, showed how vulnerable these new networks were [11]. As the Internet grew in the 1980s and

90s, digital threats exploded, leading to massive data breaches and strict new laws worldwide, like the General Data Protection Regulation (GDPR) [12]. Today, cybersecurity protects everything from our bank accounts to critical power grids. The financial damage from cyberattacks is immense; studies show that the average cost of a data breach can reach millions of dollars, and thousands of new cyberattacks happen every single day.

Artificial Intelligence has become a powerful tool in this fight. Research widely shows its potential to find threats faster, automate defences, and predict attacks before they happen. It helps with spotting malware, detecting intrusions, and identifying vulnerabilities.

However, AI itself creates new challenges. There's an 'AI arms race', where attackers also use AI to create more complex threats [13]. Many advanced AI models are like 'black boxes' making decisions can't be fully understand or explain, which is risky for trust and debugging in security. Researchers also highlight problems with AI learning from biased data [14] and the risk of adversarial attacks designed to trick AI into making mistakes [15]. These issues show that while AI is incredibly powerful, its use in cybersecurity requires careful design to ensure that it's truly safe and reliable.

2.4. Existing Efforts and the Evolving Landscape of AI Governance

2.4.1. Current Regulatory and Standard Initiatives

The growing societal impact of AI has spurred significant global efforts to establish frameworks, standards, and regulations aimed at promoting trustworthy and responsible AI development and deployment. Among the most prominent initiatives are those from the US National Institute of Standards and Technology (NIST) and the European Union (EU). The NIST AI Risk Management Framework (AI RMF), for instance, provides a voluntary framework for organisations to manage risks associated with AI throughout its lifecycle, emphasising governance, mapping, measuring, and managing AI risks [16]. Its primary goal is to foster trustworthy AI systems through collaboration across public and private sectors, focusing on the principles like accountability, transparency, and explainability. Similarly, the EU's AI Act represents a landmark legislative effort, positioning itself as the world's first comprehensive legal framework for AI [17]. It adopts a risk-based approach, categorising AI systems based on their potential to cause harm, with stringent requirements for 'high-risk' AI applications, particularly

in critical infrastructure, law enforcement, and public services. Key provisions include requirements for data governance, human oversight, robustness, accuracy, transparency, and cybersecurity measures. These, along with other national and international guidelines, collectively demonstrate a global commitment to guiding AI development.

2.4.2. International Perspectives and Emerging Consensus

Beyond the significant efforts of NIST and the EU, other nations and international bodies are also actively contributing to the discourse on AI governance, reflecting a growing global consensus on the need for responsible AI. Countries like Canada, the United Kingdom, and Singapore have released their own AI ethics guidelines and strategies, often echoing core principles, such as fairness, transparency, accountability, and human oversight [18, 19]. Organisations like the Organisation for Economic Cooperation and Development (OECD-0 have also developed intergovernmental AI principles for promoting innovative, trustworthy AI that respects human rights and democratic values, providing a foundation for international cooperation and interoperability [20]. This widespread engagement underscores the universal recognition of AI's transformative power and the imperative for collective action to ensure its responsible deployment. While approaches may vary in their legal enforceability or specific focus areas, a common thread of ethical considerations, risk management, and human-centric design runs through these diverse initiatives.

2.4.3. Gaps in Existing Frameworks for Autonomous AI

Most existing frameworks such as: NIST AI Risk Management Framework (2023), EU High-Level Expert Group on AI Ethics Guidelines (2019), OECD AI Principles (2019), and IEEE P7000 Series are primarily focused on monolithic, centralised AI systems (e.g. classifiers and recommendation engines), rather than multi-agent LLM ecosystems. The table 2 shown below describes what's missing in these existing frameworks.

2.5. The Missing Blueprint: A Call for Principles in Autonomous AI

The newest frontier in AI involves agentic AI frameworks. These aren't just programs; they let us build highly independent digital 'agents' that can set goals, use tools, and solve complex problems on their own, often through natural conversation. The speed at which these agents are developing is truly astounding, with new tools and applications appearing constantly.

Table 2. Gaps in Existing Frameworks for Autonomous AI.

Framework	Coverage of agentic design needs	What's missing
NIST AI RMF	Risk categories, governance, and lifecycle best practices	Lacks guidelines on distributed decision-making, memory use, and multi-agent flow control
EU ethics guidelines	Abstract principles (transparency, non-maleficence, human agency)	No agent-specific safety metrics or structural roles for decision diffusion
OECD AI principles	High-level governance, human-centred values	Not intended for programmatic evaluation of agentic pipelines or tool-driven decisions
IEEE P7000 series	Focused on explainability and ethical impact assessment	Requires significant adaptation for LLM-based orchestration and emergent behaviour
ISO/IEC 24028:2020	AI trustworthiness principles	Broad and technical, not specific to memory-sharing or tool-chaining AI agents

However, despite this rapid growth, there's a big gap in current research: we don't have a clear, agreed-upon set of core architectural principles specifically for the agentic AI frameworks. While there's plenty written about single-agent LLMs and general software design, the unique independence and often unpredictable behaviour of agentic AI demands a new kind of 'design philosophy'. Unlike the careful, collaborative process that gave us OOP's foundational rules, or the deep academic work that grounded early AI, the agentic AI field has raced ahead without a widely accepted blueprint for building trustworthy, ethical, and resilient autonomous systems. This paper aims to explore this critical void by examining the essential characteristics that *should* underpin the next-generation autonomous AI systems. We investigate how the existing agentic AI frameworks implicitly or explicitly approach these unstated principles, analysing their strengths and limitations in achieving trustworthy, ethical, and resilient intelligence. Our goal is to highlight the urgent need for a more deliberate architectural consensus, especially for protecting our digital world from constantly evolving threats.

3. The Critical Gap: Unmet Challenges of Autonomous Agentic AI

While the existing efforts in AI governance have laid crucial groundwork for responsible AI development, a critical architectural void persists, particularly when addressing the unique, emergent, and philosophical challenges posed by increasingly autonomous, self-directing, and collaborative AI agents. Current frameworks, though instrumental in setting a foundational baseline, primarily

focus on general AI systems and often adopt a compliance-driven approach. They provide essential guidelines for risk management, data quality, and human oversight, yet they frequently fall short in offering the foundational architectural principles necessary to design and engineer trustworthiness into autonomous systems from the ground up. The accelerating pace of agentic AI development, coupled with its inherent unpredictability, highlights this critical gap:

- Multi-step reasoning and independent action: Unlike conventional software or even many narrow AI applications, agentic systems engage in complex, multi-step reasoning processes to achieve goals, often making independent decisions without continuous human intervention. This makes tracing causality and assigning responsibility significantly more complex.
- Dynamic environment interaction: Autonomous agents operate and adapt within highly dynamic and often unpredictable real-world environments. Their constant interaction with and learning from these environments can lead to novel behaviours that are difficult to anticipate or control through static rules.
- Multi-agent collaboration: Many advanced agentic systems involve multiple AI entities collaborating to solve problems. This introduces challenges related to inter-agent trust dynamics, secure communication protocols, and mechanisms to prevent internal propagation of errors or malicious intent between collaborating AI entities.
- Emergent behaviours: Autonomous agents can develop behaviours not explicitly programmed or foreseen, making static compliance rules insufficient. The existing frameworks struggle to account for the dynamic, often opaque, 'why' behind an agent's self-directed actions, particularly in complex, multi-agent interactions.
- Philosophical unpredictability: As AI agents gain more autonomy, precisely specifying their objectives and ensuring their continued alignment with human intent in novel or unforeseen situations becomes profoundly difficult. Leading AI safety researchers consistently emphasise that this level of philosophical unpredictability in highly capable AI requires a deeper, architectural commitment to principles, not just regulatory compliance [13].

This architectural void is not merely a theoretical concern; it leads directly to critical problems, such as unpredictable behaviours, a pervasive trust deficit, and compromised security in autonomous AI systems. The absence of an intrinsic, design-driven approach means that ad hoc development of powerful autonomous agents

risks perpetuating these challenges, impeding their safe and reliable deployment in sensitive domains. It underscores the urgent need for a more deliberate, foundational shift in how we conceive and engineer trustworthiness for the next generation of AI.

3.1. Comparative Analysis of Existing Regulatory Principles

To highlight this critical gap more systematically, we first present a comparative landscape of key AI-relevant regulatory principles across major jurisdictions: the EU's GDPR, California's Consumer Privacy Act (CCPA), China's Personal Information Protection Law (PIPL), and India's Digital Personal Data Protection Act (DPDP Act, 2023). The table 3 shown below, analyses and demonstrates a shared global consensus on general principles such as transparency, accountability, data security, bias mitigation, privacy-by-design, and human oversight.

3.2. Alignment of Existing Principles with Trustworthy Autonomous AI Principles

While the regulatory principles in table 3 are fundamental, their application to the dynamic and complex nature of autonomous agentic AI reveals limitations. The table 4 shown below, maps these established regulatory principles against our proposed trustworthy autonomous AI principles. This mapping reveals that while there is overlap, our principles (explainable autonomy, dynamic adaptability, secure collaboration, proactive resilience, and ethical alignment by design) address crucial dimensions that are either implicitly covered at a high level or entirely absent in the specific context of *autonomous*, *self-directing*, and *collaborative* AI systems.

3.3. Alignment Matrix: Quantifying Support for Regulatory Features

The table 5 shown below visualises the relationship, showing which of our proposed principles directly support specific regulatory features. Digit '1' indicates direct support or a strong contribution, while '0' indicates less direct or no explicit contribution. This matrix clearly demonstrates that while our principles broadly align with regulatory goals, they also fill specific gaps by focusing on the architectural implementation necessary for the unique challenges of autonomous AI. For instance, while 'transparency and explainability' is a general regulatory principle, our 'explainable autonomy' specifically operationalises it for autonomous decision-making, and 'ethical alignment by design' reinforces

Table 3. Comparative Landscape of AI-Relevant Regulatory Principles Across Jurisdictions

AI-relevant regulatory principle/feature	GDPR (EU)	CCPA and CPPA (California, US)	PIPL (China)	DPDP Act, 2023 (India)
Transparency and explainability in automated decisions	Right not to be subject to solely automated decisions (Art. 22); right to meaningful information on decision logic (Recital 71); need for explainable AI (XAI) to address 'black box' problem [20].	Pre-use notices for automated decision-making technology (ADMT); explanation of ADMT's effect on consumer; importance of transparent AI models and user understanding [21, 22].	Article 24 mandates transparency and fairness in automated decision-making results; individuals have the right to request clarification and refuse decisions made solely through automation that significantly impact their rights [23].	Act aims to safeguard privacy by promoting transparency in data handling; current act lacks clarity on contesting automated decisions, with further mechanisms possibly in upcoming Digital India Act [24].
Accountability for AI systems	Data controller accountability (Art. 5(2), 24); DPIAs for high-risk processing; organisations must show evidence of compliance and responsibility for system impacts. [20]	Requires risk assessments for AI/ADMT; businesses must ensure ADMT is accurate and unbiased; definitive governance structures to oversee system performance and enforce audits [21, 22].	Data handlers must ensure accuracy and completeness of personal information; accountability for processing personal information; develop and refine protection systems for large-scale service providers [23].	Data fiduciaries must implement strong safeguards to ensure accountability; strict penalties for non-compliance; emphasises data fiduciaries' responsibility for data they handle and collect [24].
Data Security and Integrity	Requires appropriate technical and organisational security measures; protection from breaches; regular audits [20].	Robust security measures (encryption, and access control); protecting AI systems from attacks (adversarial training); regular security audits and penetration testing [21, 22].	Strict data protection rules; adequate security measures based on data sensitivity; ensures data integrity and confidentiality [23].	Requires 'reasonable security safeguards' (encryption, obfuscation, masking, access control); breach notification obligations within 72 hours; promotes data security to prevent breaches [24].
Bias Mitigation and Fairness	Principle of fairness; obligation to ensure data accuracy to avoid harmful/biased outcomes; emphasises preventing discrimination. [20]	Businesses must evaluate ADMT for accuracy and non-discrimination; regular bias testing and thorough documentation; warnings against biased data. [21, 22]	Article 24 ensures fairness and impartiality of results in automated decision-making; 'do not apply unreasonable differential treatment'; spot and counter apparent prejudices. [23]	Ethical concerns around bias; AI algorithms analysing personal data can lead to biases and discrimination; emphasises the need for ethical and responsible AI development. [24]

(continues)

Table 3. Continued

AI-relevant regulatory principle/feature	**GDPR (EU)**	**CCPA and CPPA (California, US)**	**PIPL (China)**	**DPDP Act, 2023 (India)**
Privacy/security by design	Mandates integrating data protection from the outset (Art. 25); proactive approach; data minimisation, encryption, anonymisation strategies [20].	Encourages 'privacy-by-design techniques' to refrain from processing unnecessary personal data; incorporating privacy features like encryption and data anonymisation into AI systems [21].	Clear processing rules, openly and transparently disclosed; minimise collection to only what is necessary for the purpose; embedding ethical practices [23].	Adopting 'privacy by design' to minimise risks; integrating privacy considerations from early stages of AI model development; strong emphasis on data minimisation and purpose limitation [24].
Human oversight/ intervention	Right not to be subject to solely automated decisions (Art. 22); right to obtain ex post human intervention, express view, and contest decisions [20].	Consumers can opt out of ADMT use; notification of adverse significant decisions; emphasises human oversight and intervention, especially in high-stakes decision-making [22].	Individuals have the right to refuse the handler for making a decision *only* through automated means if it significantly impacts their rights and interests [23].	While not explicitly detailed for AI, the Act emphasises empowering citizens with control over their data; the upcoming Digital India Act is poised to introduce robust mechanisms to govern AI and safeguard user rights, including human review [24].

CPPA: California Privacy Protection Agency.

Table 4. Alignment of AI-Relevant Regulatory Principles with Proposed Trustworthy Autonomous AI Principles

AI-relevant regulatory principle/feature	Proposed trustworthy autonomous AI principles
Transparency and explainability in automated decisions	*Explainable autonomy*: Directly addresses the need for clear insights into autonomous AI decisions and their underlying logic. *Ethical alignment by design*: Ensures that transparency fosters trust and responsible AI use, allowing for scrutiny and contestation of decisions.
Accountability for AI systems	*Ethical alignment by design*: Establishes clear responsibility, governance structures, and auditability for autonomous AI systems, ensuring adherence to ethical guidelines and legal obligations. *Proactive resilience*: Implies that systems are designed to be robust enough to remain accountable for their actions and impacts, even under duress. *Dynamic adaptability*: Supports accountability by ensuring that system adjustments and learning are conducted responsibly and are traceable. *Secure collaboration*: Enables clear attribution of actions and secure inter-agent communication, crucial for establishing accountability in complex multi-agent environments.
Data security and integrity	*Secure collaboration*: Focuses on protecting data during interactions within and between AI systems and with external entities. *Proactive resilience*: Emphasises building inherent resistance to threats and vulnerabilities, safeguarding data throughout its lifecycle. *Dynamic adaptability*: Allows systems to adapt their security posture and mechanisms in response to evolving threats. *Ethical alignment by design*: Integrates data security as a fundamental ethical consideration to prevent harm and misuse.
Bias mitigation and fairness	*Ethical alignment by design:* Crucially aims to prevent autonomous AI from perpetuating biases and discrimination, ensuring equitable and just outcomes. *Explainable autonomy*: Requires transparency in how biases are identified and mitigated within AI models. *Dynamic adaptability*: Allows systems to continuously learn and correct for emerging biases over time. *Proactive resilience:* Ensures robustness against adversarial attacks that might introduce or exacerbate bias.
Privacy/security by design	*Proactive resilience*: Advocates for embedding security and privacy into AI system architecture from the outset to anticipate and mitigate risks. *Ethical alignment by design*: Integrates privacy and ethical considerations as foundational architectural components, ensuring responsible data handling and adherence to ethical norms. *Dynamic adaptability*: Enables the system to update its privacy and security design features in response to new regulations or threats. *Secure collaboration*: Ensures that collaborative interactions themselves are designed with privacy and security as core considerations.
Human oversight/ intervention	*Explainable autonomy*: Provides the necessary transparency and understanding for humans to effectively monitor and intervene in autonomous AI decisions. *Ethical alignment by design*: Ensures that human control and review mechanisms are built into AI systems to uphold ethical standards, protect rights, and facilitate responsible deployment.

Table 5. Alignment Matrix: Trustworthy Autonomous AI Principles Supporting Regulatory Features.

	Explainable autonomy	Dynamic adaptability	Secure collaboration	Proactive resilience	Ethical alignment by design
Transparency and explainability in automated decisions	1	0	0	0	1
Accountability for AI systems	0	1	1	1	1
Data security and integrity	0	1	1	1	1
Bias mitigation and fairness	1	1	0	1	1
Privacy/security by design	0	1	1	1	1
Human oversight/intervention	1	0	0	0	1

its purpose. Crucially, principles like 'secure collaboration' and 'proactive resilience' provide essential architectural depth for aspects like data security and accountability in complex, dynamic, and multi-agent environments that general regulations can only touch upon.

4. The Architecture of Tomorrow: New Foundations for Autonomous AI

Our digital journey began by taming complexity. Just as OOP brought order to general software, allowing us to build intricate systems piece by piece, we now face an even greater challenge: architecting truly autonomous AI. This section explores AI's deep roots, the remarkable leaps it has made, and why its latest, fastest evolution demands a new set of foundational principles which is a kind of 'constitution' for intelligent agents. This critical evolution, from early software's reliance on architectural blueprints to the urgent need for foundational principles in autonomous AI, is vividly illustrated in Figure 1.

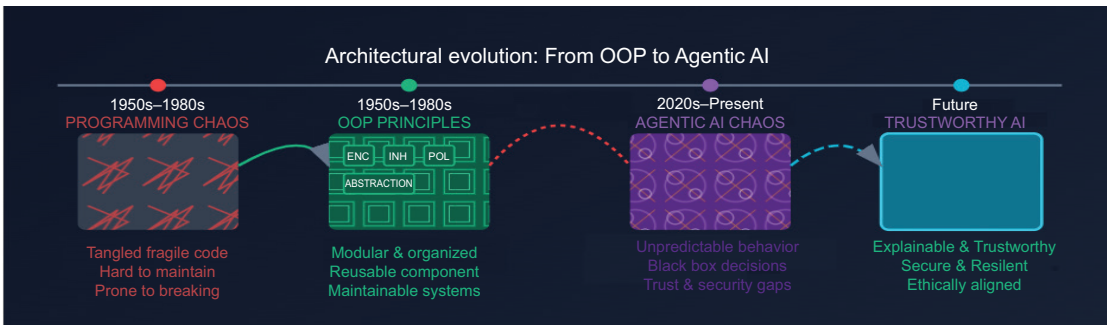

Figure 1. Architectural Evolution of Software Design from Programming Chaos to Trustworthy AI.

4.1. From Order in Software to Chaos in AI: The Need for a New Blueprint

The very idea of 'artificial intelligence' wasn't born in a vacuum; it emerged from deep philosophical questions and computational insights laid down by visionary thinkers. These early ideas continue to shape how we understand what 'intelligence' means for machines. Consider Alan Turing's pivotal 1950 paper, 'Computing machinery and intelligence' [6]. He famously asked, 'Can machines think?' and proposed the Turing test, a simple yet profound thought experiment to determine if a machine could convincingly mimic human conversation. His work set the grand ambition for AI: to build a universal machine capable of any task that could be computed, fundamentally challenging our perceptions of intelligence. Then, in 1956, a landmark event, the Dartmouth Summer Research Project on Artificial Intelligence [7] formally established AI as a dedicated field. Pioneers like John McCarthy (who actually coined the term 'Artificial Intelligence'), Marvin Minsky, Nathaniel Rochester, and Claude Shannon gathered, united by the belief that 'every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it' [7]. This meeting launched symbolic AI, where machines would tackle problems by logically manipulating symbols and rules, much like how humans use logic to solve puzzles. A prime example was Allen Newell and Herbert A. Simon's 'Logic Theorist' (1956) [8], one of the first AI programs, which proved machines could indeed perform logical reasoning. Their later physical symbol system hypothesis [25] became a cornerstone of classical AI, suggesting that intelligence fundamentally arises from the ability to process and manipulate symbols. Complementing this, Norbert Wiener's work on cybernetics (1948) [26] explored the science of control and communication, revealing how feedback loops and self-regulation allow systems (whether biological or mechanical) to adapt and maintain balance, principles vital for any intelligent behaviour. In later decades, researchers like B.G. Buchanan and E.H. Shortliffe were instrumental in developing expert systems [27], which captured and applied human expert knowledge through detailed rules, showing AI's power in specialised domains. These trailblazers provided AI's first blueprints, drawing wisdom from diverse fields like neuroscience, linguistics, mathematics, psychology, computer engineering, and control theory [28]. This rich, interdisciplinary foundation gave early AI its structured beginnings.

Yet, the sheer power of recent AI, especially the explosion of LLMs and the rapid emergence of agentic frameworks, has created a unique predicament. Companies are swiftly deploying their own

'agentic AIs', but this breakneck speed means we're building before we've fully considered the blueprints. Unlike programming languages that found their guiding philosophy in OOP, or traditional AI that matured with a broad academic consensus, agentic AI currently lacks a universally recognised set of foundational architectural principles. We see wild, emergent behaviours and complex interactions in these autonomous agents, demanding more than just clever code. We need a shared understanding of how to build truly robust, trustworthy, and ethical autonomous intelligence. The challenges are stark: an escalating AI arms race with adversarial AI, the opaque 'black box' dilemma where we don't understand AI decisions, and the perilous ethical tightrope of granting machines more autonomy.

The very minds shaping AI's future are voicing an urgent call for a new architectural paradigm to address these critical challenges. Yoshua Bengio, a Turing award laureate and leading advocate for AI safety, has consistently expressed deep concern regarding the accelerated pace of AI development. He argues that we are rapidly deploying increasingly autonomous systems without adequate 'safeguards in place', directly emphasising the critical need for a more deliberate, principle-driven approach to their fundamental design and deployment [29, 30]. Similarly, Stuart Russell, a preeminent investigator in AI safety research and author of *Human Compatible*, highlights the fundamental difficulty in precisely specifying objectives for highly capable AI. He actively advocates for intrinsic design principles that ensure that AI systems are provably beneficial and controllable from their inception, stressing that such a foundation is non-negotiable for human compatibility [31]. Furthermore, Dario Amodei, CEO of Anthropic, has not only raised concerns but actively pursued practical solutions like 'constitutional AI,' (CAI) which directly embeds a set of guiding principles and rules into the very fabric of large language models to guide their behaviour safely and ethically [32]. This collective sentiment from the forefront of AI research from the imperative for architectural safeguards to the explicit embedding of ethical considerations points to a clear consensus: the ad hoc development of powerful autonomous agents must give way to a universally recognised set of guiding axioms for trustworthiness and safety.

4.2. Architecting Trust: Proposed Foundational Principles for Autonomous AI

Inspired by these critical insights and the shared urgency voiced by leaders across the field, we propose a fresh set of foundational principles for agentic AI frameworks. These new axioms, like a

vital 'constitution' for intelligent agents, distil the collective wisdom on ensuring responsible and robust AI. They aim to address the profound complexities of true intelligence, drawing parallels to OOP's elegance while extending its philosophy to the autonomous domain:

- *The principle of explainable autonomy*: To trust autonomous agents, we must understand why they act. This demands architectures that inherently reveal their reasoning, allowing humans to audit, debug, and learn, instead of facing an inscrutable black box. As Yann LeCun, a prominent AI researcher and Professor at New York University, has emphasised, 'Our intelligence is what makes us human, and AI is an extension of that quality' [33]. His point underscores that for AI to truly be an extension of human intelligence, it inherently *must be interpretable* and transparent in its actions, allowing for understanding and trust. Can we build trust without transparency?
- *The principle of dynamic adaptability*: In a world of constant change and evolving threats, AI frameworks must be designed for continuous, robust learning and self-correction. This goes beyond simple updates, envisioning systems that seamlessly integrate new knowledge and evolve capabilities while remaining stable and safe. As Jensen Huang, CEO of NVIDIA, has highlighted regarding the modern AI factory, 'it generates, simulates, reasons, and adapts continuously' [34]. This directly supports the principle's core, arguing that true intelligence in dynamic environments requires inherent flexibility and constant, stable evolution, not just static programming. How do we engineer intelligence that learns forever without breaking?
- *The principle of secure collaboration*: As AI agents increasingly work together and with human teams, their frameworks must ensure secure and ethical interaction. This means designing for privacy-preserving data sharing, robust inter-agent communication, and secure multi-agent orchestration, preventing malicious exploitation of distributed intelligence. The potential for misuse in interconnected systems is a significant concern, a reality acknowledged by experts like Brad Smith, vice chair and president of Microsoft, stated that 'we need to go into this new AI era with our eyes wide open. There are real risks and problems that we need to figure out how to solve. It's vital that the technology is safe and remains subject to human control, that we have the ability to slow it down or turn it off if it's not functioning the way we want [35].'

Smith's emphasis on confronting the inherent risks of AI, ensuring its safety, and crucially, maintaining human control, serves as

a direct call for the secure and ethical architectures necessary for agents to collaborate without compromise or exploitation. This includes the implementation of robust safeguards to ensure human oversight and the ability to intervene, directly addressing the question: If intelligence is shared, how do we prevent its misuse?

- *The principle of proactive resilience*: Anticipating adversarial attacks is crucial. This principle demands AI frameworks that are inherently resilient, not just reactive. It means building in defences against data poisoning, adversarial examples, and model inversion, designed to withstand deliberate attempts to trick or corrupt them. As concerns about AI safety grow, researchers like Geoffrey Hinton, a pioneer in deep learning, have stated that 'when it comes to the dangers of AI, there are two quite different kinds of risks. One is the kind of risk that comes from bad human actors using AI. And I think that's most of the risk, and all of the short-term risk. And then there's the risk that comes from AI getting super smart and deciding it doesn't need humans [36].'

Hinton's call for robust protection against malicious actors precisely aligns with the essence of proactive resilience, moving beyond mere reaction to inherent defensive design**.** Can we build AI robust enough to intelligently fight back?

- *The principle of ethical alignment by design*: Perhaps the most critical. This foundational principle would embed human values, ethical guidelines, and safety constraints directly into the very architecture of AI frameworks. It moves beyond mere policy, aiming to create systems whose core design inherently guides them towards beneficial and responsible actions, even in unforeseen circumstances. This idea resonates with the broader AI alignment community, with thinkers like Brian Christian, author of *The Alignment Problem*, exploring how to ensure 'AI systems … are aligned with human values' [37]. Christian's work directly underpins the goal of this principle which is to build systems whose very fabric ensures alignment with human values. Satya Nadella also champions this, noting, 'Every single developer choice, that design ethos you exhibit, the ethics of the diverse team you have are going to matter … in terms of whether we are going to create a much more inclusive world' [38]. Nadella's vision reinforces that ethical design ensures that AI serves humanity beneficially, rather than operating in isolation or against human interests. How do we hardwire humanity into the very fabric of autonomous intelligence?

These emerging 'axioms' represent the next great architectural quest in the digital age. They are the conceptual bedrock upon which the most advanced AI frameworks of today and tomorrow are being built, striving to deliver powerful, trustworthy, and ethically aligned autonomous intelligence. The subsequent sections of this paper explore how current and emerging frameworks are striving to embody these vital principles, showcasing their application in the areas ranging from general AI development to the formidable realm of cybersecurity.

As discussed in Section 2, the current landscape of AI governance faces significant gaps when addressing the unique architectural requirements of autonomous, multi-agent AI systems. Our proposed principles are specifically designed to fill these identified voids. The table 6 shown below, illustrates how our approach directly addresses these gaps, providing a foundational blueprint for trustworthy agentic AI.

The architecture of LLM-based multi-agent systems introduces complexity, feedback loops, and decision diffusion – risks that are not directly addressed by the existing monolithic AI standards [39].

Our study is targeted towards trustworthy AI principles to a new class of AI systems (agentic AI), an architecture-informed mapping of the five building blocks to existing frameworks, and a comparative study for evaluation of agentic AI systems on their conformance to these principles.

Table 6. Mapping Identified Gaps to Proposed Solutions for Autonomous AI.

Area	Gap identified	Our proposal
System complexity	Traditional frameworks assume central model control	We propose decomposable, role-defined modularity
Autonomous behaviour	No guidance on multi-agent planning or drift	Introduce guardrails, fallback agents, and memory transparency
Explainability	Focused on machine learning interpretability, not process trace	Introduce agent log chains, retrieval mapping, tool call audit
Ethics and privacy	No enforcement model for redaction or tool scoping	Recommend scoped permissions, ethical refusal modelling
Evaluation	Lack of comparative scoring for agentic AI	Offer framework to assess and benchmark system conformance

5. Principles

Having established the critical gap between the existing AI governance frameworks and the unique demands of highly autonomous agentic AI, we now introduce our proposed solution: five core architectural principles designed to serve as a foundational 'constitution' for building trustworthy and resilient independent AI systems. These principles are not merely ethical guidelines, they are intended to be operationalised design considerations that directly address the emergent behaviours, complex collaborations, and dynamic adaptability inherent in next-generation autonomous agents. By detailing their specific application to agentic AI and illustrating their embodiment within prominent frameworks, we demonstrate their innovative contribution, enhanced logical coherence, and practical relevance.

5.1. Introduction to Our Principles: The Architectural Constitution for Autonomous AI

The unprecedented autonomy of agentic AI systems necessitates a paradigm shift in how we conceive and engineer trustworthiness. Unlike traditional software, where functionality is largely pre-defined, or earlier AI models with more constrained operational envelopes, autonomous agents can set sub-goals, leverage diverse tools, and interact dynamically with complex environments, leading to emergent behaviours that defy static rule-sets or purely post hoc governance. To address this, we propose five foundational 'building blocks' or principles – explainable autonomy, dynamic adaptability, secure collaboration, proactive resilience, and ethical alignment by design – as the necessary architectural constitution for autonomous AI. These principles are designed to guide the intrinsic design of agents, ensuring that trustworthiness, safety, and reliability are embedded into their very fabric, rather than merely enforced externally. They represent a crucial step beyond general trustworthy AI concepts, offering a granular, operationalised focus uniquely tailored for the complexities of autonomous operations.

5.2. Detailed Explanation of Each Principle (Operationalised for Agentic AI)

Each of our five principles is defined and operationalised specifically for the context of agentic AI, demonstrating how it adds a new layer, integrates concepts differently, or provides a more detailed architectural perspective than broader trustworthy AI concepts. This directly addresses the concern regarding conceptual

innovation and aims to provide more than a duplication of general trustworthy AI ideas.

5.2.1. Explainable Autonomy

Definition: Explainable autonomy refers to an agent's architectural capacity to provide clear, context-aware, and actionable explanations for its independent decisions, goal-setting, and execution paths, particularly in multi-step, complex, and evolving scenarios. It moves beyond static model interpretability to dynamic process transparency.

Application to agentic AI: For agentic AI, this principle is crucial because agents operate autonomously through chains of reasoning and tool use, often in dynamic environments. Explaining a final output is insufficient; users (both humans and other agents) need to understand why a particular sub-goal was set, how a tool was selected, and what alternative paths were considered. For instance, if an agent decides to pivot its strategy in a cybersecurity defence scenario, explainable autonomy demands not just the new strategy but the rationale behind the pivot (e.g. 'identified a zero-day exploit pattern, current defence insufficient, shifting to isolation protocol based on similarity to known advanced persistent threat (APT) tactics'). This adds a new layer by focusing on the narrative of autonomy, rather than just data-input/output. It integrates concepts of explainable AI with process transparency and goal-directed reasoning, offering a more granular, architectural perspective on how an agent's internal state and external actions interrelate. As Dr. Timnit Gebru, an AI researcher, asserts, 'Model transparency is not just about understanding how a model works; it's also about being transparent about its limitations and potential biases' [40]. This highlights how transparent AI systems must address fundamental questions like 'how are decisions made?' and 'are biases being mitigated?' for autonomous operations. Similarly, Sundar Pichai, CEO of Google and Alphabet, in discussions about responsible AI development, has emphasised earning user 'trust' [41], reinforcing that transparency in AI is inherently linked to building and maintaining user confidence.

5.2.2. Dynamic Adaptability

- Definition: Dynamic adaptability is the inherent capability of an autonomous agent to safely and strategically modify its internal logic, knowledge base, or operational parameters in real-time, in response to unforeseen environmental changes, novel threats, or updated objectives, while maintaining its core mission and ethical constraints.

- Application to agentic AI: Unlike traditional software that requires patches or new deployments for significant changes, autonomous agents must adapt on the fly. This principle goes beyond general system robustness (which often implies resilience to known failures) to address safe self-modification. For example, in a dynamic threat landscape, a cybersecurity agent needs to not just detect a new malware variant but potentially adapt its detection heuristics or devise novel containment strategies without human intervention, ensuring such adaptations are auditable and do not introduce new vulnerabilities. This provides a new layer by focusing on controlled self-evolution. It integrates concepts from adaptive systems, reinforcement learning, and safety engineering, offering a detailed architectural focus on how an agent can learn and evolve within its operational boundaries. Demis Hassabis, CEO of Google DeepMind, frequently emphasises the necessity for continuous learning and adaptability in the rapidly evolving AI landscape, for both humans and AI systems themselves, to remain effective. He advises cultivating 'the skill of 'learning to learn' – the ability to quickly grasp new concepts and adapt to change – as a critical asset for thriving in an uncertain and rapidly evolving landscape' [42]. This perspective implicitly underscores the imperative for AI to possess inherent dynamic adaptability. Julian De Freitas, assistant professor at Harvard Business School, further reinforces this by contrasting human adaptability with current AI capabilities, noting that 'humans adapt; they continuously understand where they are in the world and what problem they are solving in response to changing circumstances far better than current AI does' [43]. This highlights the crucial need for AI to achieve continuous adaptation in dynamic environments.

5.2.3. Secure Collaboration

- Definition: Secure collaboration refers to the architectural design of autonomous agents to enable trusted, integrity-preserving, and privacy-aware interaction not only with human teams but also among heterogeneous AI agents in multi-agent systems. It ensures that shared goals, data, and insights are exchanged without compromising system integrity or confidentiality.
- Application to agentic AI: Agentic AI often operates in collaborative ecosystems. This principle significantly extends general data privacy and security (e.g. GDPR compliance [20]) by addressing the complex challenges of inter-agent trust, communication integrity, and shared situational awareness. For example, in a joint cyber–defence operation, one agent might share threat intelligence with another, while a third coordinates response actions. Secure collaboration ensures that these interactions

are authenticated, authorised, and resistant to malicious injection or data exfiltration between agents. This adds a new layer by focusing on the security primitives required for AI-to-AI and AI-to-human team interactions. It integrates concepts from distributed systems security, zero-trust architectures, and federated learning, offering a granular architectural perspective on ensuring secure and trustworthy collective intelligence. As Anthropic's engineering blog highlights, 'Once intelligence reaches a threshold, multi-agent systems become a vital way to scale performance... Even generally intelligent agents face limits when operating as individuals; groups of agents can accomplish far more' [44]. This underscores the increasing reliance on multi-agent systems and, by extension, the critical need for secure and reliable collaboration within them. Red Hat's blog further emphasises that 'AI security protects the systems from external and internal threats, while AI safety provides confidence that the system and data don't threaten or harm users, society or the environment due to the model's operation, training or use' [45], a principle directly applicable to the complexities of multi-agent interactions. General Paul Nakasone (Ret.), former Director of the NSA, consistently points out the profound security challenges that AI poses, stressing the need for comprehensive defences against threats to AI systems themselves, including those arising from their collaborative nature. He states, 'We must build a robust understanding of AI vulnerabilities, foreign intelligence threats to these AI systems and ways to counter the threat in order to have AI security' [46].

5.2.4. Proactive Resilience

- Definition: Proactive resilience is the architectural capacity of an autonomous agent to anticipate, prevent, self-diagnose, and self-recover from a wide range of failures, adversarial attacks (including adversarial AI), and unexpected system states, maintaining core functionality or graceful degradation, rather than catastrophic failure.
- Application to agentic AI: While general AI principles discuss robustness, proactive resilience specifically emphasises an agent's built-in ability to anticipate and react intelligently to novel threats, which is critical given the 'AI arms race' where attackers also use AI [47]. For a cybersecurity agent, this means not just recovering from a known attack but possessing the architectural components to detect and respond to a completely novel attack vector, potentially by isolating compromised components or activating contingency plans autonomously. This provides a new layer by emphasising anticipatory self-defence and automated

recovery at the architectural level. It integrates concepts from fault tolerance, self-healing systems, and threat intelligence fusion, offering a detailed architectural focus on imbuing agents with robust, intelligent defence mechanisms. As Werner Vogels, Amazon CTO, famously stated, 'Everything fails all the time' [48]. This foundational insight from a leading technologist underscores that system design must account for inevitable failures by building in resilience, redundancy, and automated recovery, directly supporting the need for proactive resilience in AI.

5.2.5. Ethical Alignment by Design

- Definition: Ethical alignment by design means architecting autonomous agents such that their goals, decision-making processes, and emergent behaviours are intrinsically and continually aligned with a predefined set of human values, ethical principles, and societal norms, with mechanisms for self-correction and human override when deviations are detected.
- Application to agentic AI: This principle is paramount, given the independent action of agentic AI. It goes beyond ethical guidelines applied to outputs or post-deployment audits, demanding that ethical considerations are embedded into the agent's core reasoning engine and goal-setting mechanisms. For example, a financial agent needs architectural safeguards to prevent it from optimising for profit in ways that are deemed predatory, even if such pathways are computationally efficient. This integrates the philosophical aspects of AI ethics (e.g. fairness and non-discrimination) with architectural mechanisms for value encoding, constraint propagation, and human-in-the-loop oversight. It provides granular focus on how to ensure an agent's internal 'compass' always points towards ethical outcomes, even when operating in ambiguous situations or encountering biased data [49]. As Sri Amit Ray states, 'Doing no harm, both intentional and unintentional, is the fundamental principle of ethical AI systems' [50], emphasising that this must be a foundational tenet. Similarly, Infosys's 'The decent dozen: 12 principles for responsible AI by design' [51] explicitly advocates for weaving ethical considerations into AI systems 'from data selection and algorithmic design to deployment and monitoring', rather than treating them as an afterthought. Mira Murati, CTO of OpenAI, further underscores this by consistently emphasising the non-negotiable integration of ethical principles into the very foundation of AI development, stating, '[AI] can be misused, or it can be used by bad actors. So, then there are questions about how you govern the use of this technology globally. How do you govern the use of AI in a way that's aligned with human values'? [52]

6. The Landscape of Agentic AI Frameworks: Tools for Architecting Trustworthy Intelligence

Our journey has shown that just as early software needed a blueprint to tame complexity, today's autonomous AI agents grapple with similar fundamental questions of trust and control.

Our exploration began by observing popular interest, noting a surge in 'AI agentic frameworks' on Google Trends. This trend analysis directly informed our selection of five key tools emerging in this space for deeper examination. These modern AI frameworks, which developers are actively using, aren't just lines of code; they're like the practical workshops where our proposed new principles are actually being built into AI. Each one offers a unique approach to making AI more transparent, flexible, secure, and responsible. To visually showcase these selected frameworks, refer to the figure 2 and 3 below (duration: June 2024–June 2025, and keyword: 'AI agentic frameworks'):

Today's agentic AI frameworks aren't just clever code; they're the battlegrounds where abstract ideas about safe and smart AI get built into real-world cybersecurity tools. We're seeing how frameworks like LangChain, AutoGen, and Semantic Kernel are making our proposed principles a reality.

Our conceptual analysis involved a systematic review of the official documentation, architectural diagrams, and common usage

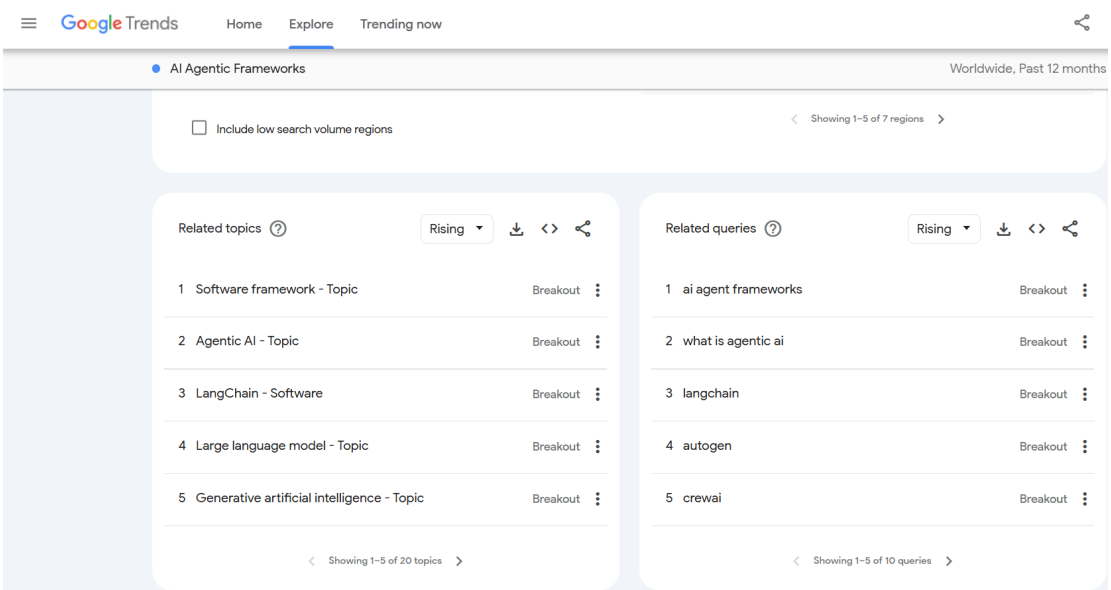

Figure 2. Top Google Trends for "AI Agentic Frameworks" (Rank 1-5).

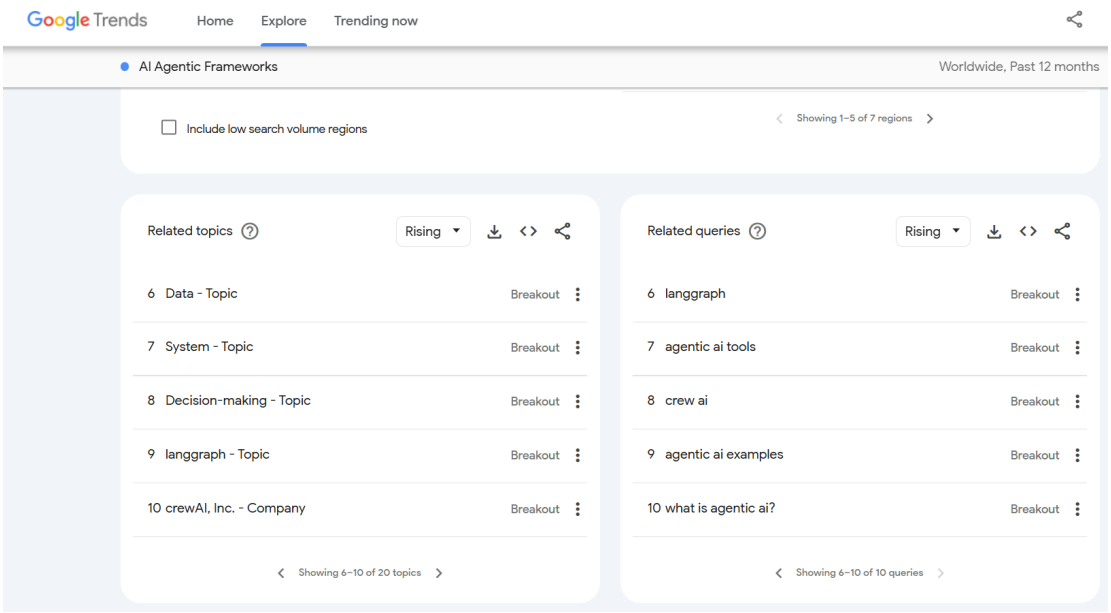

Figure 3. Top Google Trends for "AI Agentic Frameworks" (Rank 6-10).

patterns of LangChain, LangGraph, AutoGen, and LlamaIndex. For each framework, we meticulously examined how its core components, such as agents, tools, chains, graphs, and message passing mechanisms, align with or diverge from the operationalised definitions of our five principles. This involved identifying specific features that support a principle and noting inherent limitations or design philosophies that might impede its full realisation in autonomous contexts. This qualitative, feature-based comparison aims to illustrate the principles' utility in evaluating agentic architectures.

6.1. LangChain: The Orchestra Conductor for Flexible AI

Think of LangChain as the ultimate organiser for AI applications that use powerful LLMs. It's an open-source toolbox that makes it simpler to create complex AI systems [53]. While it's not just for 'agents,' its clever design lets us put different AI pieces together like LEGO bricks.

- What it does: LangChain is brilliant at 'chaining' things. It connects LLMs with other parts like templates for prompts, data converters, or outside tools which helps to build multi-step processes. It basically takes the messy parts of talking directly to an LLM and makes them easy to manage.

- How it builds trust:
  LangChain uses several core ideas to foster trust in AI systems by addressing specific principles:
  - Explainable autonomy: LangChain directly helps by giving you a clear window into how the agent makes decisions. You can inspect each individual piece, like a prompt, a memory call, or how a tool was used. Developers and even auditors can follow the exact path a query took through the system, see which data was accessed, and understand why the agent made certain choices. This traceable reasoning, backed by detailed logging and step-by-step replay features, is incredibly important in sensitive areas like cybersecurity. Additionally, its encapsulation (like OOP's hidden 'boxes' for chains and agents) helps make complicated AI workflows easier to understand and check. However, the internal reasoning within each LLM call often remains opaque, and complex agentic decision-making across multiple chains can be hard to trace, lacking deep and intrinsic explanations of autonomous goal shifts.
  - Dynamic adaptability: LangChain truly shines with Retrieval Augmented Generation (RAG). This means, it's great at linking LLMs to fresh and outside information, like databases or live data feeds. So, an AI built with LangChain can always get the latest info, helping it to adapt quickly to new situations.
  - Secure collaboration: It provides mechanisms for tool integration and sequential execution, which can be part of collaborative workflows. However, it lacks intrinsic features for secure inter-agent trust, authentication, or fine-grained access control in multi-agent scenarios, relying more on external security measures.
  - Proactive resilience: For cybersecurity, LangChain's RAG features mean that the AI can pull in the newest threat data, making it more resilient and ready for anything. It focuses on error handling within chains (e.g. retries). However, it does not inherently provide architectural components for anticipatory threat detection or complex self-healing in the face of novel attacks on the agent's own logic or data.
  - Ethical alignment by design: LangChain relies heavily on prompt engineering and external filtering for alignment. It does not offer intrinsic architectural mechanisms to embed ethical reasoning or value constraints directly into the agent's autonomous decision-making loop, making it susceptible to adversarial attacks designed to trick AI into making mistakes [15].

Why it matters for agentic AI and cybersecurity: LangChain provides the basic structure for AI agents that pursue specific goals.

In cybersecurity, its RAG features are super valuable. Imagine an agent that can instantly check huge, live threat databases to spot new attack patterns. That's a huge boost for adaptable and tough defences.

6.2. LangGraph: Mapping Out AI's Thoughts for Clearer Decisions

LangGraph is a powerful expansion of LangChain, introducing a 'graph-based' way to program LLM agents [54]. It's specifically built to make AI workflows more aware of their current state, more transparent, and easier to check. Developed because simple step-by-step AI chains weren't enough for truly complex agentic systems, LangGraph lets you design an agent's entire behaviour like a detailed map, where each point is a clear computational step and lines show how the AI moves based on specific conditions, its memory, or current context.

- What it does: LangGraph's core breakthrough is making an autonomous agent's reasoning process physically visible. Instead of a hidden linear flow, an agent's logic is laid out as a computational graph. Every 'node' in this graph is a specific action: an LLM prompt, using a tool, calling an application programming interface (API), or retrieving memory. The 'edges' between nodes define the rules for moving forward, for example if a condition is met, if a step succeeds or fails, or based on user feedback.
- How it builds trust:
  LangGraph builds trust by offering detailed support for several principles:
  - Explainable autonomy: This is a big win for LangGraph. Because you've got this clear map of the AI's actions, you can literally follow its thought process. You can see exactly what tools it used, what decisions it made, and why. This transparent 'blueprint,' especially with tools like LangSmith, makes AI behaviour much clearer and easier to audit. The explicit nodes and edges in the graph provide a visual 'trace' of the agent's thought process and state transitions, operationalising dynamic transparency for autonomous paths. LangGraph often includes specific human-in-the-loop points in its maps, allowing people to check, guide, or even override AI's decisions at crucial moments.
  - Dynamic adaptability: LangGraph lets you build loops into your AI's map. This means that the AI can try something, see if it worked, and if not, go back and try a different approach. This constant learning and self-correction are key for dynamic adaptation. While its graph structure allows for more complex flows

and conditional logic, true self-modification of the graph structure by the agent itself is not a native feature. It enables richer pre-designed adaptive behaviours but not inherent architectural learning.

- Secure collaboration: LangGraph supports multi-agent systems by letting different agents occupy different parts or branches of the computational graph (e.g. a 'research agent,' a 'reviewer agent,' a 'responder agent'). Their teamwork is explicitly designed into the map's layout, reducing confusion and risks often tied to less controlled communication between AI entities. This architectural clarity directly supports secure collaboration, crucial in cybersecurity where agents must work together without accidentally exposing sensitive data or taking unauthorised actions. Similar to LangChain, however, it lacks intrinsic, fine-grained secure communication or trust mechanisms between distinct agents within a multi-agent graph.
- Proactive resilience: For cybersecurity, an agent could try to fix an issue, and if it fails, it can automatically rethink and try another solution, making defences incredibly robust. LangGraph benefits from the deterministic nature of graph execution for easier debugging and recovery from predictable failures. However, it does not inherently offer architectural features for anticipatory detection of novel threats or autonomous recovery from internal logic corruption.
- Ethical alignment by design: LangGraph is still largely dependent on prompt engineering within nodes and external guardrails. While the graph structure could theoretically contain ethical decision nodes, it doesn't provide an intrinsic ethical reasoning engine that self-aligns.

Why it matters for agentic AI and cybersecurity: LangGraph is perfect for creating intricate, team-based AI agents that follow detailed cybersecurity plans. Its ability to manage an AI's 'memory' (state) is vital for constant threat monitoring agents that need to remember past events over long periods.

6.3. AutoGen: The Team Leader for Secure AI Collaboration

AutoGen, developed by Microsoft, is a powerful framework that introduces a 'conversational' way to program AI agents [55]. It's all about building and coordinating LLM-powered agents through structured, dialogue-based interactions. The goal is to enable multiple autonomous agents, each with a specific job, to work together and solve complex tasks by talking to each other naturally. This framework is a strong advocate for explainable

autonomy and secure collaboration, and it lays a solid foundation for goal-driven, auditable, and modular AI systems.

- What it does: AutoGen makes it simple to set up complex AI 'conversations'. You can create different AI agents, each with their own job, goals, and ways of talking. These agents then chat back and forth, sharing messages, using tools, or even running code to get things done.
- How it builds trust:
  AutoGen excels at building trust through:
  - Explainable autonomy: While AutoGen doesn't show you the AI's deepest thoughts, the clear record of messages between agents creates a 'conversation trail.' This acts as a form of explainability, letting humans follow how the AI team worked together and why certain decisions were made in a group setting. AutoGen supports explainable autonomy through transparent inter-agent communication logs. Tool invocations and external integrations are also fully logged and credited to the individual agent, aiding forensic analysis and repeatability. However, the internal reasoning within each individual agent's LLM remains a black box, and the rationale for specific message content or tool choices might not be explicitly articulated.
  - Dynamic adaptability: AutoGen agents can hand off tasks and solve problems on the fly through their conversations. They can change their roles and how they contribute based on what's happening in their discussion, showing great dynamic adaptability as they work together and sort things out. AutoGen shows stronger support for dynamic adaptation through iterative communication and self-correction between agents. Agents can 'learn' from failures or new information within a conversation and adjust their behaviour, demonstrating nascent architectural support for continuous, collaborative adaptation.
  - Secure collaboration: This is AutoGen's superpower. By giving agents clear ways to talk to each other, assigning them roles, and letting them work in defined groups, AutoGen specifically designs for secure teamwork. It creates a framework for safe communication and controlled access between these smart entities, which is vital for stopping bad actors from messing with your AI team. However, the security aspects (e.g. authentication of agents, integrity of messages, and prevention of malicious agent injection) are largely external to the framework's core and rely on underlying platform security.
  - Proactive resilience: AutoGen agents are equipped to use external tools and APIs. When an agent calls a tool, it's formatted as a structured message within the conversation and

then executed, with the results fed back into the dialogue. This allows agents to combine their language understanding with practical actions, such as querying a threat intelligence database, scanning logs, or creating structured reports. AutoGen benefits from multi-agent redundancy and self-correction through dialogue (e.g. one agent correcting another's output), offering a form of distributed resilience. However, architectural components for individual agent self-diagnosis or anticipatory threat intelligence are not inherently integrated.

- Ethical alignment by design: AutoGen relies on the individual alignment of each agent through prompt engineering and potentially external moderation. AutoGen doesn't inherently provide a centralised, architectural mechanism to enforce consistent ethical behaviour across a dynamic group of interacting agents or to prevent malicious prompts from one agent to another.

Why it matters for agentic AI and cybersecurity: AutoGen is fantastic for modelling cybersecurity teams or processes. Imagine an AutoGen-powered 'security operations centre (SOC) crew' where a 'threat analyst agent' works with a 'fix-it agent' and a 'reporting agent'. They could autonomously find, contain, and document a cyberattack, hugely boosting secure teamwork in defence.

6.4. LlamaIndex: The Data Detective for Smart and Adaptive AI

LlamaIndex, originally known as GPT Index, is an open-source framework launched in 2022 by Jerry Liu and LlamaIndex Team [56]. Its main goal is to connect powerful LLMs with your own custom data, whether it's neatly organised or a jumble of information. It's now a crucial part of building systems that can answer questions or reason over your specific data with context, primarily through what's called Retrieval-Augmented Generation (RAG). LlamaIndex lets developers easily pull in data from almost anywhere like documents, web pages, APIs, databases, notion pages, and PDFs, and then organise it into special formats like vector stores or graph-based structures, and finally, let LLMs ask questions about that data using natural language. It essentially simplifies the process of giving an LLM the exact, domain-specific knowledge its needs for highly accurate answers.

- What it does: LlamaIndex provides comprehensive tools to build and manage different ways of searching data (like vector search or tree search) from many different sources. LLMs can then

quickly query these organised data structures for tasks like RAG, giving them rich context and powering autonomous agents.

- How it builds trust:
  LlamaIndex builds trust primarily through its data-handling capabilities:
  - Explainable autonomy: A core strength of RAG systems built with LlamaIndex is their ability to show you where they got their information when answering a question. This source attribution is a vital part of explainable autonomy. It lets users double-check the facts behind an AI's response and understand its information sources, which is incredibly valuable for auditing and trust in crucial security situations. However, it provides limited architectural insight into the agents independent decision-making process beyond the RAG pipeline itself. Its abstraction (from classic OOP) hides the complex data ingestion, breaking, embedding, and database interaction, simplifying development and data management for AI engineers.
  - Dynamic adaptability: LlamaIndex has tons of 'DataConnectors' **(**LlamaHub**)** that can pull information from almost anywhere. This steady flow of fresh specific data keeps AI systems current and highly responsive to new information, which is key for dynamic adaptability. It supports dynamic updating of indices and knowledge graphs, allowing agents to adapt their understanding of the world. However, the framework doesn't inherently support self-modification of the agent's core reasoning or control flow.
  - Secure collaboration: Primarily a data management and retrieval framework, LlamaIndex can serve as a component in collaborative systems by providing shared knowledge. However, it does not offer intrinsic features for secure inter-agent communication or trust mechanisms.
  - Proactive resilience: In cybersecurity, getting the latest threat info, incident reports, or vulnerability details via LlamaIndex makes AI agents more resilient because they're always working with up-to-date knowledge. It focuses on robust data retrieval and handling, including error handling for external data sources. However, it does not inherently provide architectural components for an agent's self-diagnosis or anticipatory defence against broader system-level threats. A current limitation is that LlamaIndex is not inherently designed for adversarial robustness or threat modelling. It doesn't have built-in defences against common issues like prompt injection attacks, adversarial inputs, or ways to reduce hallucinations. Instead, it relies on developers to add extra steps like filtering input data, validating retrieved content, or implementing content moderation,

meaning achieving full proactive resilience requires additional developer effort.
- Ethical alignment by design: LlamaIndex facilitates grounding LLMs in factual data, which can reduce hallucinations and thus improve reliability. However, like other frameworks, it relies on prompt engineering and external mechanisms for embedding ethical values into the agent's autonomous decision-making, rather than providing intrinsic ethical reasoning components.

Why it matters for agentic AI and cybersecurity: For cybersecurity, LlamaIndex is priceless for building security-focused RAG systems. A 'cyber threat research agent' could use LlamaIndex to query internal vulnerability databases, old incident reports, or outside threat intelligence, giving it rich context for answers and supporting a more informed and resilient defence.

6.5. CrewAI: Orchestrating Smart Teams for Secure Operations

CrewAI is an exciting emerging framework that lets developers build collaborative multi-agent AI systems [57]. It's especially powerful because it allows AI agents to truly behave like members of a team, with each agent taking on a specific task. Essentially, it's like being able to manage a crew of autonomous assistants, where each has specialised functions, they share knowledge, and they work together on subtasks, all under one larger project goal.

- What it does: CrewAI focuses on managing intelligent agents. It lets them take on specialised roles (like researcher, writer, or analyst) and work together to hit shared goals through clear workflows and natural decision-making.
- How it builds trust:
  CrewAI builds trust through its emphasis on structured teamwork and clear responsibilities:
  - Explainable autonomy: While CrewAI doesn't show you an agent's deep internal reasoning, its focus on clear roles, goals, and breaking big problems into smaller tasks for specific agents helps with explainable autonomy. Users can see which agent is responsible for what part of the solution and how their individual efforts combine, giving a high-level understanding of the team's collective thinking. Detailed inter-agent conversation logs (verbose mode) also provide external 'thought process' tracing.
  - Dynamic adaptability: When agents are part of a CrewAI team, they can dynamically assign tasks, check progress, and adjust

their strategies based on how the overall goal is unfolding. This is a great example of dynamic adaptability. If one agent gets stuck, others can change their approach or take over, ensuring the whole task keeps moving forward. While strong for pre-designed adaptive behaviours through delegation and feedback, agents do not intrinsically self-modify their core structure or role at runtime.

- Secure collaboration: This is what CrewAI is all about. By letting you give agents clear roles, goals, and even backstories, CrewAI sets up structured and intentional conversations between them. The framework is designed to handle complex team dynamics, creating a strong environment for secure collaboration. Each agent knows its job and how it fits into the bigger picture. This structured approach also lays the groundwork for things like access controls and checking agent communications, which is key for preventing misuse in sensitive applications. However, it lacks intrinsic, fine-grained secure channels, agent authentication, or injection prevention, relying on external security.
- Proactive resilience: CrewAI offers error handling (retries, max_iter) and distributed resilience through agent self-correction. For cybersecurity, this means agents could work together to find, contain, and fix advanced threats, hugely boosting proactive resilience through coordinated defensive actions. However, it lacks inherent architectural support for anticipatory threat detection, complex self-diagnosis, or autonomous core logic self-healing against novel threats.
- Ethical alignment by design: CrewAI relies on prompt engineering, role/goal definitions, and external guardrails. It does not provide intrinsic architectural mechanisms for ethical reasoning or value constraint embedding in autonomous decision-making.

Why it matters for agentic AI and cybersecurity: CrewAI is exceptionally well-suited for automating and simulating complex cybersecurity operations that usually rely on human teams. Imagine a 'cyber incident response crew' with agents specialised in network forensics, malware analysis, and vulnerability assessment, all working together to find, contain, and fix advanced threats. This hugely boosts secure collaboration and proactive resilience through coordinated defensive actions.

7. Discussion

Our journey has traversed the landscape of AI's architectural evolution, from the foundational order provided by OOP to

the urgent call for new principles in autonomous intelligence. We meticulously explored how contemporary agentic AI frameworks like LangChain, LangGraph, AutoGen, Semantic Kernel, LlamaIndex, and CrewAI are striving to embody the proposed axioms of explainable autonomy, dynamic adaptability, secure collaboration, proactive resilience, and ethical alignment by design. While these frameworks represent a monumental leap forward in building sophisticated digital sentinels for cybersecurity, a critical discussion of their current capabilities, inherent limitations, and the path ahead is essential.

In short, these frameworks are more than just software libraries. They are the practical builders, translating the abstract 'constitution' of agentic AI into concrete, reliable, and secure autonomous systems that are ready to tackle the toughest challenges in the digital world. They are proving that our architectural axioms are not just theory but the essential blueprints for the future of AI.

7.1. Remaining Challenges in Architecting Trustworthy Autonomous AI

Even with the amazing progress from these new AI frameworks, some really tough problems still stand in the way of making truly trustworthy and strong autonomous AI for cybersecurity as shown in below table 7.

7.2. Future Directions: Architecting the Next Generation of Secure Autonomous AI

To truly unlock the power of agentic AI in cybersecurity, we need to tackle the tough challenges head-on. The path forward involves smart, focused effort in several key areas, as shown in below table 8.

7.3. A Comparative Lens on Claude's Constitutional AI

As LLMs evolve into multi-agent ecosystems capable of independent reasoning, memory management, and external tool use, it becomes necessary that these agents behave safely, ethically, and predictability are central to their operability, societal trust, and regulatory viability. Traditional methods such as reinforcement learning from human feedback (RLHF) are increasingly seen as insufficient or unscalable in real-world, role-based, and autonomous agent systems [58].

Table 7. Key Problems and Impacts of Remaining Challenges in Autonomous AI.

Challenge category	Key problem/limitation	Impact on AI principles	Why it's a 'tough problem'
The 'why' behind the 'what': explainability still has gaps	Even if AI shows *what* it did, we still can't fully see *why* its 'brain' made that choice. It's like seeing an answer without the steps.	Hinders explainable autonomy.	This makes it tough to fully trust, fix, or legally approve big security decisions, especially when something goes wrong and you need to know the exact reasoning.
Building AI that fights back: intrinsic robustness	Today's AI can *use* tools to defend, but it's not born tough. It doesn't automatically protect itself from sneaky new attacks like poisoned data or clever tricks to fool it.	Limits proactive resilience.	Making AI naturally strong against clever manipulation, so it's always ready for a fight, is a huge research problem we're still trying to solve.
Keeping AI teams trustworthy: scaling ethics	When lots of different AIs work together, it's hard to guarantee that they all stick to ethical rules, especially with unexpected situations or unknown agents.	Challenges secure collaboration and ethical alignment by design.	Right now, we often rely on simple rules or checking after the fact. We need to figure out how to bake ethics directly into the AI's core design for these complex teams.
Bad data, bad AI: data integrity and bias	AI systems need good, clean data. Even with good connections, current tools don't automatically fix problems like biased data or if someone tries to poison the data.	Undermines dynamic adaptability and proactive resilience.	If the data is flawed, it can seriously mess up how well security AIs can adapt and how reliable their actions are, potentially leading to big security mistakes.
Managing the AI orchestra: complexity at scale	While it's easier to build individual AIs, managing hundreds or thousands of them all working together creates new, complicated problems.	Impacts the practical implementation of all principles by introducing systemic vulnerabilities and inefficiencies.	This massive scale adds new layers of complexity. Keeping track of how all these AIs depend on each other, managing their resources, and making sure they all play nicely together (especially using different tools) can lead to new kinds of breakdowns.
Working with humans: alert overload	Finding the right balance between what AI handles and what humans check is hard. AI can generate so many warnings that human teams get overwhelmed and tired.	Threatens effective explainable autonomy and the human-in-the-loop aspect of ethical deployment.	We need to design AI so that humans and AI can truly work as a team, without the AI constantly flooding humans with so much information that they get burned out, making the AI less helpful in the long run.

In response to RLHFs limitations, CAI has emerged as an important innovation in terms of model-level alignment. Introduced by Anthropic [59], CAI offers a self-improvement strategy for LLMs where the model refines its output based on a set of pre-defined ethical or behavioural rules called the constitution. This includes a two-step process. During the first phase, supervised learning, the model revises harmful AI responses through self-critique and

Table 8. Key Challenges and Proposed Solutions for Autonomous AI.

Title	Key challenge	Solution/approach	Core concept
AI that explains itself, inside out	AI agents' actions are often observable, but their reasoning and certainty are not.	Design AI models that are transparent by design, with 'brains' that naturally show their thinking or built-in checks to validate decisions.	Explainable autonomy
AI that fights back, intelligently and automatically	Need AI frameworks that are constantly vigilant, ready for new attacks, and can self-heal.	Develop AI systems capable of automatically spotting new attacks, reacting to them, and self-recovering from issues. Focus on self-fixing and predictive threat capabilities.	Proactive and adaptive resilience
AI with a built-in moral compass	As AI agents gain power, embedding human values and ethical rules directly into their core design is crucial.	Program ethical guidelines and safety checks as unshakeable parts of the AI's 'DNA' within its frameworks.	Ethical alignment by design
AI Teams that trust each other, securely	Ensuring ironclad teamwork and secure communication among multiple AI agents, especially in cybersecurity.	Implement new methods for secure AI agent communication, such as unforgettable digital 'handshakes' or decentralised verified trust systems.	Secure collaboration
Humans and AI, working as one seamless team	The goal is to empower humans, not replace them; future AI needs to facilitate easy human–AI collaboration.	Design AI frameworks that filter noise, present only critical alerts to humans, and visualise complex cyber threats in easily understandable ways.	Human–AI symbiosis

fine-tuning. In the second phase, the model is trained via reinforcement learning, but instead of human feedback, it uses AI-generated feedback based on the defined principles to generate more harmless output.

However, CAI is fundamentally scoped for static, single-agent behaviour within a textual response loop. It does not account for architectural concerns that arise when deploying multi-agent LLM systems. The five frameworks proposed in this paper collectively addressed the challenges of agent-level trustworthiness from a design and engineering perspective.

Table 9 shown below illustrates operations at fundamentally different levels of AI stack.

CAI helps an LLM refine its own individual output, based on the encoded values in a constitution, but it does not ensure that one agent does not overstep its responsibilities, access unauthorised tools, or leak information across modules. These are architectural

Table 9. Comparative Analysis of Constitutional AI and Our Agentic Design Framework.

Layer	Constitutional AI (CAI)	Our agentic design framework
Focus	Output behaviour refinement	System design and control flow
Level	LLM internal training behaviour	External architecture and runtime interaction
Method	Prompt self-critique using rules	Modular design, role-based scoping, fallback mechanisms
Scope	Static, single-agent LLMs	Dynamic, multi-agent tool-using systems
Output	Safer and more honest textual responses	Auditable, explainable, aligned behaviour over time

challenges of autonomy, collaboration, scope containment, and resilience – challenges central to our proposed idea.

7.3.1. Comparing Specific Principles

- Explainable autonomy versus implicit revision: CAI improves the quality and safety of outputs, but the process is opaque to end-users and external systems. It offers no causal reasoning chains. In contrast, our principle demands that every action taken by an agent be explainable in terms of intent, rationale, and scope. This helps advocate for architectural affordances and explainability is a precondition for accountability, allowing developers or auditors to reconstruct why decisions were made [60].
- Dynamic adaptability versus static rule following: CAI assumes a fixed set of rules to be defined during training which remain unchanged. This ensures consistency, but fails in environments that demand contextual flexibility such as real-time cybersecurity response. Our principle emphasises that agentic AI systems must be state-aware, context sensitive, and reconfigurable.
- Secure collaboration versus no collaboration model: CAI does not model or even anticipate multi-agent collaboration, role assignment, or tool-based cooperation. It assumes that a single model generates output and evaluates it in isolation. Our principle of secure collaboration identifies design practices like access control for tools, agent sandboxing, and shared memory permissions to prevent information leakage, task interference, or accidental misuse.
- Proactive resilience versus harmlessness by correction: While CAI aims for correctness, it is reactive – outputs are filtered and corrected after they are generated. This makes it vulnerable to prompt injection, tool failure, and agent drift during execution. Our principle anticipates failure points and demands architectural support for fallback paths, redundant agent assignments, and real-time input validations.

- Ethical alignment by design versus rulebook at training: CAI operationalised ethics by fine-tuning with constitutional prompts. This does not guarantee that once the model is integrated into complex workflows, the agent may bypass its constitution through unintended tool use or emergent behaviour. In contrast, our principle embeds ethics into the system's structure. Rather than hoping the model behaves ethically, we architect the system such that it cannot behave unethically.

7.4. The Foundational Challenge: Bridging Policy-Level Governance with Architectural Principles for Trustworthy Autonomous AI

Strongly recommended: Addressing the critical absence of foundational principles in autonomous AI. This single line encapsulates the precise motivation behind this work. While the world is witnessing an unprecedented proliferation of AI laws, ethical codes, and regulatory frameworks, what is conspicuously absent is a cohesive, widely accepted set of architectural design principles that guide the internal construction of autonomous AI systems, especially agentic AI systems capable of self-directed, adaptive, and tool-driven behaviour.

This distinction between policy-level norms and architecture-level principles is more than semantic; it reflects a fundamental divide in where and how trustworthiness in AI is established.

7.4.1. Policies and Laws: Governing AI from the Outside

Global initiatives have taken notable steps to address AI safety and trust:

- The European Union AI Act classifies AI systems into risk categories and proposes legal restrictions based on application [17].
- The NIST AI Risk Management Framework introduces voluntary guidelines for managing bias, reliability, and explainability [16].
- The OECD Principles on AI promote human-centred, transparent, and robust AI across nations [19].
- The UNESCO Recommendation on the Ethics of AI outlines ethical guardrails based on human rights and sustainability.

These instruments are essential, but they are inherently reactive. They focus on what AI systems should not do after they've been developed or deployed. They emphasise compliance, documentation, and risk mitigation through external checks, such as model cards, transparency reports, or audit logs.

Yet, as AI systems grow in complexity, especially those involving autonomous multi-agent orchestration, tool usage, recursive reasoning, or long-term memory, such post hoc methods become inadequate. At best, they detect problems after damage has occurred; at worst, they provide no practical guidance to developers on how to structure AI agents in ways that make safety, transparency, and ethical alignment inherent to the system itself.

7.4.2. The Missing Core: Building AI from the Inside

Unlike policies, foundational architectural principles are proactive and structural. They inform how AI systems should be engineered from the outset to behave in ways that are:

- transparent in their reasoning (not just externally explainable);
- adaptable to new threats and information (not statically bound);
- collaborative in controlled and secure ways (not emergently chaotic);
- resilient against failures and adversarial inputs (not brittle); and
- ethically aligned by design (not post hoc rule-checked).

This is precisely the void that this paper addresses by proposing five core principles:

- Explainable autonomy – Architectures must be able to show how and why decisions are made.
- Dynamic adaptability – Systems must continuously learn and update without compromising safety.
- Secure collaboration – Agents must share information safely and follow formal trust boundaries.
- Proactive resilience – Agents must be able to detect and recover from errors or attacks.
- Ethical alignment by design – Systems must encode and prioritise human values at the design level.

These principles are not abstract ideals; they are design blueprints meant to guide the structure and behaviour of systems built with frameworks like LangChain, AutoGen, LlamaIndex, LangGraph, and CrewAI.

7.4.3. Why Policies Alone Are Not Enough?

Policies cannot substitute for principles. To make this case clear, consider the analogy to urban planning:

- Policies are zoning laws, noise ordinances, or safety codes. They govern what buildings can be built, how high they can rise, and who can live there.

- Foundational architectural principles are the engineering standards, the rules that ensure buildings don't collapse, burn easily, or poison their inhabitants.

Without internal integrity, compliance means little. An AI system might meet documentation standards while harbouring brittle, unexplainable, or exploitable internal logic.

The growing use of LLM-based agents that invoke APIs, store memories, and interact autonomously across platforms underscores a stark reality: we are deploying systems with complex emergent behaviours using architectures that were never built for such complexity. Regulatory mechanisms are struggling to keep up. What's missing is a clear architectural foundation – a constitution for AI design.

7.4.4. Complementarity, Not Redundancy

The goal of this paper is not to replace the existing policies but to complement them. These foundational principles act as scaffolding for implementing ethical, legal, and social values at the system design level. For example:

- Policies demand explainability → Explainable autonomy operationalises it through memory graphs, agent trails, or prompt trees.
- Policies demand robustness → Proactive resilience embeds self-defence mechanisms and adaptive recovery directly into the agent.
- Policies demand human oversight → Secure collaboration ensures agents interact in bounded, human-controllable roles.

By doing so, these principles translate regulatory goals into engineering practice – something current policy frameworks rarely achieve with clarity or technical specificity.

7.4.5. A Timely Call to Action

We stand at an inflection point. Agentic AI frameworks are maturing rapidly. Developers are building systems capable of autonomous research, cyber defence, customer interaction, and more. Yet, the industry lacks a shared language for describing how these systems should be structured to earn trust, not just obey policy.

The reviewer's statement – that there are many laws, but no foundational principles – reflects the latent need for system-level clarity.

This paper responds by providing those principles, rooted not in law or philosophy but in architecture, behaviour, and control.

Without foundational principles, laws become barriers. With foundational principles, laws become enablers.

This work, therefore, should be understood as a ground-level contribution – a first step towards a universal, architecture-native framework for autonomous AI systems that are not only powerful and capable but also explainable, resilient, secure, and ethically aligned by default.

8. Conclusions

Our journey through the evolution of AI, from the structured beginnings rooted in OOP to the dynamic complexities of modern agentic frameworks, underscores a pivotal truth: the greater the autonomy granted to intelligent systems, the more critical becomes the need for a robust, shared foundation. Just as OOP provided a universal language for taming software complexity, the rapid, often ad hoc, deployment of powerful autonomous agents now demands a similar, yet far more profound, guiding philosophy.

The urgent calls from leading thinkers in the AI community, from scientists like Yoshua Bengio emphasising the need for safeguards, to industry leaders like Satya Nadella advocating for adaptable and human-centric AI, and ethicists like Brian Christian focusing on alignment highlighting a consensus: we are at a crossroads. The wild, emergent behaviours and opaque decision-making of highly autonomous AI, coupled with the escalating threats of adversarial attacks, necessitate a proactive and principled approach to their design.

This paper proposes five foundational principles which are explainable autonomy, dynamic adaptability, secure collaboration, proactive resilience, and ethical alignment by design as a ‘constitution’ for trustworthy agentic AI. These aren’t mere guidelines; they are architectural imperatives that, when embedded into the very fabric of AI frameworks, promise to transform potential peril into reliable progress.

However, the responsibility for forging and enforcing this new digital ‘constitution’ extends far beyond individual developers or companies. It demands a collective and concerted effort. Auditors

must develop new methodologies to scrutinise agentic behaviours against these principles, ensuring accountability and transparency. Lawmakers are tasked with translating these technical imperatives into enforceable regulations that foster innovation while safeguarding societal well-being. Scientists must continue to advance research into the theoretical underpinnings and practical implementation of these axioms, pushing the boundaries of what resilient and ethical AI can achieve. And industry leaders must commit to adopting these principles not as an afterthought but as central tenets of their AI development lifecycle.

The future of autonomous intelligence hinges on this collaborative commitment. By collectively embracing and institutionalising these foundational principles, we can move beyond simply building powerful AI to architecting truly trustworthy, beneficial, and enduring autonomous systems that serve humanity's best interests. This is the next great challenge, and our collective journey to trust depends on meeting it head-on.

References

[1] Elliott, E. (Nov. 1, 2018). Forgotten history of OOP [Online]. Available: https://medium.com/javascript-scene/the-forgotten-history-of-oop-88d71b9b2d9f [Accessed: Jun. 10, 2025].

[2] Nerd.vision. (Jan 6, 2021). Polymorphism, encapsulation, data abstraction and inheritance in object-oriented programming [Online]. Available: https://www.nerd.vision/post/polymorphism-encapsulation-data-abstraction-and-inheritance-in-object-oriented-programming. [Accessed: Jun. 10, 2025].

[3] G. Booch, *Object-oriented analysis and design with applications*, 2nd ed. Menlo Park, CA: Benjamin Cummings, 1994.

[4] B. Meyer, *Object-oriented software construction*. Upper Saddle River, NJ: Prentice Hall, 1988.

[5] C.E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[6] A.M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950. doi: 10.1093/mind/LIX.236.433.

[7] J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, vol. 27, no. 4, p. 12, 2006. doi: 10.1609/aimag.v27i4.1904.

[8] A. Newell, H. Simon, "The logic theory machine—A complex information processing system," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 61–79, 1956. doi: 10.1109/TIT.1956.1056797.

[9] D. Crevier, *AI: The tumultuous history of the search for artificial intelligence*. New York, NY: BasicBooks, 1993.

[10] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, Vol. 521, pp. 436–444, 2015. doi: 10.1038/nature14539.

[11] A. Vaswani, "Attention is all you need," arXiv e-prints, Art. no. arXiv:1706.03762, 2017. doi: 10.48550/arXiv.1706.03762.

[12] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, et al. "Language models are few-shot learners," arXiv:2005.14165v4 [cs.CL], 2020. doi: 10.48550/arXiv.2005.14165.

[13] S. Russell, *Human compatible: Artificial intelligence and the problem of control*. London: Penguin, 2019.

[14] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown Publishing Group, 2016.

[15] I. J. Goodfellow, J. Shlens, C. Szegedy. "Explaining and harnessing adversarial examples," arXiv:1412.6572, 2014, doi: 10.48550/arXiv.1412.6572.

[16] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST AI 100-1 [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. [Accessed: Jun. 23, 2025].

[17] European Commission. (2024). *Artificial Intelligence Act* [Online]. Available: https://digital-strategy.ec.europa.eu/en/news/european-artificial-intelligence-act-comes-force. [Accessed: Jun. 23, 2025].

[18] Smart Nation Singapore. (2020). *Model AI governance framework* [Online]. Available: https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework. [Accessed: Jun. 25, 2025].

[19] Organisation for Economic Cooperation and Development (OECD). (2019). *Recommendation of the council on artificial intelligence* [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [Accessed: Jun. 26, 2025].

[20] European Union (EU). (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council—of 27 April 2016—on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. [Accessed: Jun. 26, 2025].

[21] California Legislative Information. (2018). *California code, CIV 1798.100* [Online]. Available: https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.100. [Accessed: Jun. 27, 2025].

[22] California Privacy Protection Agency (CPPA). (2024). *Proposed regulations on CCPA updates, cybersecurity audits, risk assessments, automated decision-making technology (ADMT), and insurance companies* [Online]. Available: https://cppa.ca.gov/regulations/ccpa_updates.html. [Accessed: Jun. 27, 2025].

[23] People's Republic of China. (2021). *Personal information protection law of the People's Republic of China* [Online]. Available: https://personalinformationprotectionlaw.com/. [Accessed: Jun. 28, 2025].

[24] Parliament of India. (2023). D*PDP Act 2023* [Online]. Available: https://sansad.in/getFile/BillsTexts/LSBillTexts/Asintroduced/113_2023_LS_Eng83202330313PM.pdf?source=legislation. [Accessed: Jun. 28, 2025].

[25] A. Newell, H. A. Simon. "Computer science as empirical inquiry: Symbols and search," *Communications of the ACM*, vol. 19, no. 3, pp. 113–126, 1976, doi: 10.1145/360018.360022.

[26] N. Wiener, *Cybernetics or control and communication in the animal and the machine*. Cambridge, MA: MIT Press, 1948 (2019), doi: 10.7551/mitpress/11810.001.0001.

[27] W. R. Swartout, "Rule-based expert systems: The mycin experiments of the Stanford heuristic programming project: B.G. Buchanan and E.H. Shortliffe (Addison-Wesley, Reading, MA, 1984)," *Artificial Intelligence*, vol., no. 3, pp. 364–366, 1985, doi: 10.1016/0004-3702(85)90067-0.

[28] S. Russell, P. Norvig, *Artificial intelligence: a modern approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[29] Y. Bengio. (Apr, 2025). *The catastrophic risks of AI—and a safer path*, TED Talk [Online]. Available: https://www.ted.com/talks/yoshua_bengio_the_catastrophic_risks_of_ai_and_a_safer_path. [Accessed: Jun. 29, 2025].

[30] CBC Radio. (Jun 6, 2025). *Can AI safeguard us against AI? One of its Canadian pioneers thinks so* [Online]. Available: https://www.cbc.ca/radio/asithappens/ai-safety-non-profit-1.7553839. [Accessed: Jun. 30, 2025].

[31] Chinadaily.com.cn. (Jan 14, 2025). *Special attention needed to ensure AI safety, US professor says* [Online]. Available: https://www.chinadaily.com.cn/a/202501/14/WS678643c2a310f1265a1dacb9.html. [Accessed: Jun. 30, 2025].

[32] *Observer*. (Jul 14, 2023). *Anthropic CEO Dario Amodei discusses teaching human values to A.I. models* [Online]. Available: https://observer.com/2023/07/anthropic-ceo-ai-safety/. [Accessed: Jul. 1, 2025].

[33] DataDrivenInvestor. (Apr 9, 2020). *What's next in AI?* [Online]. Available: https://www.datadriveninvestor.com/2020/04/09/whats-next-in-ai/. [Accessed: Jul. 5, 2025].

[34] VivaTech. (Jun, 2025). *Inside Jensen Huang's GTC Paris keynote on AI factories, token economics and the infrastructure race.* NEXTDC [Online]. Available: https://www.nextdc.com/blog/vivatech-2025-nvidias-vision-for-the-ai-factory-era-and-why-infrastructure-now-leads-the-innovation-curve. [Accessed: Jun. 20, 2025].

[35] Microsoft UK Stories. (Jun 8, 2023). '*Eyes wide open': The importance of responsible AI innovation* [Online]. Available: https://ukstories.microsoft.com/features/importance-of-responsible-ai-innovation/. [Accessed: Jun. 19, 2025].

[36] G. Hinton. (Jun 16, 2025). *Godfather of AI: Tried to warn them, but we've already lost control*! [Online]. Available: https://www.youtube.com/watch?v=giT0ytynSqg&t=78s. [Accessed: Jun. 29, 2025].

[37] B. Christian. *The alignment problem: machine learning and human values.* First edition. W.W. Norton & Company, New York, 2020.

[38] Forbesindia.com. (Feb 26, 2020). *Microsoft CEO Satya Nadella addresses ethics in tech creation* [Online]. Available: https://www.forbesindia.com/article/special/microsoft-ceo-satya-nadella-addresses-ethics-in-tech-creation/57921/1. [Accessed: Jun. 9, 2025].

[39] X. Gao, et al. (2024). *Trust in autonomous LLM systems* [Online]. Available: https://arxiv.org/abs/2401.05561. [Accessed: Jul. 3, 2025].

[40] Number Analytics. (Jun 2025). The power of transparency in AI decision-making. Blog [Online]. Available: https://www.numberanalytics.com/blog/power-of-transparency-ai-decision-making#fn:3. [Accessed: Jun. 20, 2025].

[41] The420.in (2025). *"We earned the trust": Pichai on data privacy, AI, and responsibility* [Online]. Available: https://the420.in/sundar-pichai-google-trust-privacy-ai-antitrust-defense-interview/. [Accessed: Jul. 2, 2025].

[42] Education Today. (2025). *Demis Hassabis urges students to embrace AI and prepare for AGI-driven future* [Online]. Available: https://www.educationtoday.co/news/daily-news/google-deepmind-ceo-demis-hassabis-warns-students-prepare-for-a-future-transformed-by-ai. [Accessed: Jul. 5, 2025].

[43] Harvard Business School. (2024). *How humans outshine AI in adapting to change.* Working Knowledge [Online]. Available: https://www.library.hbs.edu/working-knowledge/how-humans-outshine-ai-in-adapting-to-change. [Accessed: Jun. 30, 2025].

[44] Anthropic. (2025). *How we built our multi-agent research system* [Online]. Available: https://www.anthropic.com/engineering/built-multi-agent-research-system. [Accessed: Jul. 7, 2025].

[45] H. Sidhpurwala, G. Mollett, E. Fox, M. Bestavros, H. Chen (2024). *Building trust: Foundations of security, safety and transparency in AI. Red Hat research paper*. [Online]. Available: https://www.redhat.com/en/blog/building-trust-foundations-security-safety-and-transparency-ai. [Accessed: Jun. 27, 2025].

[46] US Department of Defence. (2023). *AI security centre to open at national security agency.* Defence Department News [Online]. Available: https://www.defense.gov/News/News-Stories/Article/Article/3541838/ai-security-center-to-open-at-national-security-agency/. [Accessed: Jun. 24, 2025].

[47] O. Al-Badrawi. "Artificial intelligence and arms races in the Middle East: The evolution of technology and its implications for regional and international security," *Defence Studies*, vol. 24, no. 1, pp. 1–22, 2024, doi: 10.1080/14751798.2024.2302699.

[48] Sri Ananth Nadiger. (2024). *Quotes #10 — Everything fails*. Medium [Online]. Available: https://medium.com/cre8ve-thoughts/quotes-unquoted-10-everything-fails-a3e0e4c67d24. [Accessed: Jun. 30, 2025].

[49] The Greenlining Institute. (2021). *Algorithmic bias explained* [Online]. Available: https://greenlining.org/wp-content/uploads/2021/04/Greenlining-Institute-Algorithmic-Bias-Explained-Report-Feb-2021.pdf. [Accessed: Jul. 2, 2025].

[50] Sri Amit Ray. (2022). *Ethical AI systems quotes* [Online]. Available: https://www.goodreads.com/work/quotes/96703034-ethical-ai-systems-frameworks-principles-and-advanced-practices. [Accessed: Jul. 2, 2025].

[51] J. Kavanaugh, Ritarshi Chakraborty (2024). *The decent dozen: 12 principles for responsible AI by design* [Online]. Available: https://www.infosys.com/iki/perspectives/responsible-ai-design-principles.html. [Accessed: Jun. 29, 2025].

[52] ABP News Bureau. (2023). *Mira Murati creator ChatGPT OpenAI AI misused regulated*. Interview Time [Online]. Available: https://news.abplive.com/technology/chatgpt/mira-murati-creator-chatgpt-openai-ai-misused-regulated-interview-time-1580951. [Accessed: Jun. 29, 2025].

[53] H. Chase, LangChain Contributors. (2023). *LangChain: Building applications with LLMs through composability* [Online]. Available: https://docs.langchain.com. [Accessed: Jun. 8, 2025].

[54] LangGraph. (2024). *LangGraph documentation: Graph-based coordination for LLM agents* [Online]. Available: https://docs.langgraph.dev. [Accessed: Jun. 10, 2025].

[55] Y. Wu, M. Wang, Y. Zhou, A. Jain, Y. Liu, C. Zhang. (2023). *AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework* [Online]. Available: https://arxiv.org/abs/2309.00738. [Accessed: Jun. 11, 2025].

[56] J. Liu, LlamaIndex Team. (2022) LlamaIndex documentation: Data framework for LLM applications [Online]. Available: https://docs.llamaindex.ai/. [Accessed: Jun. 11, 2025].

[57] CrewAI Team. (2024). CrewAI documentation: Orchestrate your AI agents [Online]. Available: https://www.crewai.com/. [Accessed: Jun. 13, 2025].

[58] L. Ouyang, et al. (2022). *Training language models to follow instructions with human feedback* [Online]. Available: https://arxiv.org/abs/2203.02155. [Accessed: Jul. 2, 2025].

[59] Y. Bai, S. Kadavath, S. Kundu, et al. (2022). *Training a helpful and harmless assistant with constitutional AI* [Online]. Available: https://arxiv.org/abs/2212.08073. [Accessed: Jul. 6, 2025].

[60] F. Doshi-Velez, B. Kim. (2017). *Towards a rigorous science of interpretable machine learning* [Online]. Available: https://arxiv.org/abs/1702.08608. [Accessed: Jul. 6, 2025].