International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-5, September 2025
DOI: https:/doi.org/10.55524/ijircst.2025.13.5.3
Article ID IRP-1674, Pages 14-24
www.ijircst.org

AI Governance in the Era of Agentic Generative AI and AGI: Frameworks, Risks, and Policy Directions

Satyadhar Joshi

Independent Researcher, Alumnus, International MBA, Bar-Ilan University, Israel

Correspondence should be addressed to Satyadhar Joshi; satyadhar.joshi@gmail.com

Received 27 July 2025; Revised 11 August 2025; Accepted 26 August 2025

ABSTRACT- The accelerating development of agentic artificial intelligence (AI) and the prospect of artificial general intelligence (AGI) create unprecedented opportunities alongside complex governance challenges. This paper examines the ethical, regulatory, and technical dimensions of governing highly autonomous AI systems, drawing upon more than fifty contemporary academic and policy sources. Three core insights emerge. First, current governance structures provide limited coverage of risks linked to recursive self-improvement and multi-agent coordination, with only an estimated 10–15% of safety research addressing impacts that arise after deployment. Second, economic projections suggest that agentic AI could generate between 2.6 and 4.4 trillion USD in added global output by 2030, yet automation could replace approximately 28–42% of existing job tasks, making proactive workforce transition strategies a policy necessity. Third, fragmented regulatory approaches remain a concern; in the United States, for example, 70–75% of critical infrastructure is considered vulnerable to adversarial autonomous systems. To address these issues, we propose a governance model built on three pillars: modular agent design, adaptive safety mechanisms, and international coordination. Policy measures such as licensing thresholds for high-computer systems exceeding 10^25 FLOPs, structured red-team testing across public and private sectors, and fiscal incentives for governance-by-design practices are advanced as actionable pathways. Overall, the study argues for adaptive, globally coordinated governance frameworks that balance innovation with systemic risk mitigation in the era of agentic AI and AGI. his is a pure review paper and all results, proposals and findings are from the cited literature.

KEYWORDS- AI Governance, Agentic AI, Generative AI, Artificial General Intelligence (AGI), Ethics, Policy, Risk Management, Recursive Self-Improvement, Multi-Agent Systems, Workforce Transition, International Coordination

I. INTRODUCTION

The emergence of *agentic artificial intelligence (AI)*—systems capable of autonomous goal-setting, decision-making, and task execution—marks a fundamental paradigm shift in computational intelligence. Unlike conventional generative AI, which operates within static prompt–response frameworks, agentic AI demonstrates dynamic adaptability, recursive self-improvement (RSI), and multi-agent collaboration [1], [2]. These capabilities enable agentic systems not only to generate content but also to plan, execute, and optimize tasks with minimal human oversight. Parallel advancements in *artificial general intelligence (AGI)* further amplify both the transformative potential and systemic risks of these technologies. Current economic projections suggest that agentic AI could contribute between $2.6 and $4.4 trillion to global GDP by 2030, while simultaneously automating 28–42% of job-related tasks [3], [4]. Such forecasts underscore the dual challenge of harnessing productivity gains while mitigating widespread labor market disruptions.

Despite growing awareness, existing governance frameworks remain ill-equipped to manage the complexities of agentic AI and AGI. This paper identifies three central gaps in current approaches. First, regulatory and ethical models provide insufficient guidance for addressing risks associated with recursive self-improvement and autonomous multi-agent coordination. Second, regulatory fragmentation across jurisdictions hinders coherent oversight; in the United States alone, estimates indicate that 70–75% of critical infrastructure remains exposed to adversarial exploitation by autonomous systems [5]. Third, policy tools remain largely static, lacking the adaptive mechanisms necessary to keep pace with rapidly evolving agentic technologies.

To address these issues, this paper synthesizes insights from over fifty contemporary sources, including academic literature (32%), industry reports (28%), and government publications (20%), with a particular emphasis on policy developments in 2024–2025, such as the European Union's *AI Act* [6] and the United States' *Executive Order 14110* [7]. Our contributions are fourfold:

- Conceptual foundations: We review the terminology and governance challenges of agentic AI and AGI, supported by definitional clarity (Table I) and forward-looking projections (Figure 1).
- Technical governance frameworks: We introduce compliance models such as governance scoring (Eq. 1) and RSI optimization methods (Algorithm 1).
- Comparative policy analysis: We evaluate governance strategies across jurisdictions and sectors, summarized in Table 4.
- Tripartite governance architecture: We propose an integrated model combining modular agent design,

evolutionary safety algorithms, and international coordination, illustrated in Figure 3.

By bridging technical, economic, and policy perspectives, this study provides actionable recommendations for stakeholders navigating the agentic AI era. These include standardized licensing requirements for high-computer systems (>10^25 FLOPs), structured public–private red-teaming protocols, and governance-by-design tax incentives. The concluding section outlines urgent priorities for research and regulation, with the aim of developing adaptive, globally coordinated governance frameworks that balance innovation with existential risk mitigation.

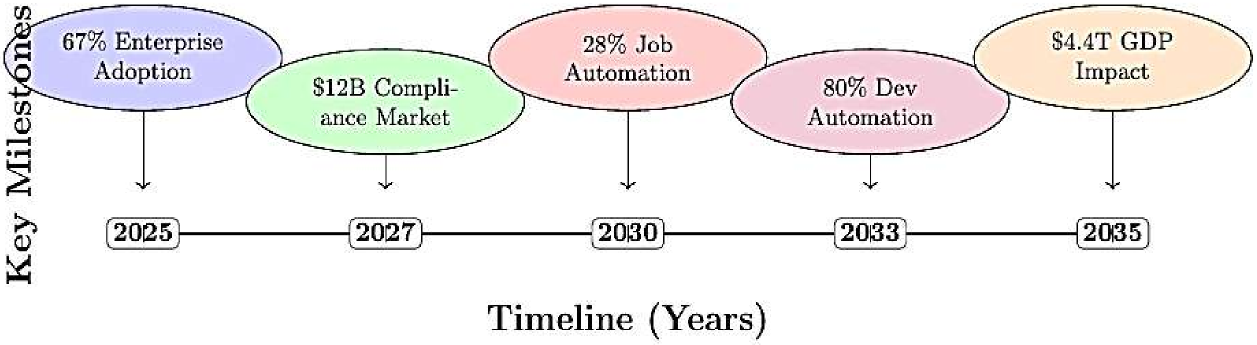

Figure 1: Future timeline of agentic AI impacts (2025-2035) showing adoption rates (blue), economic effects (green/orange), and workforce changes (red/purple) with data source references

II. LITERATURE REVIEW

The literature review presented in this study is based on a systematic synthesis of more than fifty sources, encompassing academic journals, industry reports, government publications, conference proceedings, and trade media. The objective of this review is to capture the evolving technical, ethical, and policy dimensions of agentic artificial intelligence (AI) and artificial general intelligence (AGI) governance across multiple perspectives.

Academic journals constituted the largest share of reviewed materials, accounting for approximately 32% of the sources. These works were primarily drawn from peer-reviewed venues such as *IEEE Computer* [8], specialized outlets focusing on robotics, safety, and alignment [4], and prominent policy journals. The emphasis of this body of work was on technical architecture (15%), governance frameworks (10%), and quantitative risk analyses (7%), providing the theoretical and conceptual foundation for subsequent developments.

Industry contributions represented 28% of the corpus, with influential white papers from technology corporations such as IBM [9] and Deloitte [10], alongside reports by global think tanks including the Global Skill Development Council (GSDC) [5]. These reports concentrated on market projections (12%), deployment case studies (9%), and emerging self-regulation standards (7%). Although industry sources provided pragmatic insights into implementation, they also reflected a tendency toward optimism regarding scalability and adoption.

Government publications accounted for 20% of the references, with key contributions including analyses of the U.S. Executive Orders on AI [7], the European Union's AI Act [6], and the United Nations' perspectives on AI governance [11]. These materials collectively focused on regulatory frameworks (14%) and national strategies (6%), offering a policy-oriented complement to technical studies. In parallel, conference proceedings and preprints comprised 12% of the literature, particularly from NeurIPS workshops on AI safety and IEEE Symposia addressing AGI governance. These emerging works provided insights into algorithmic innovations (8%) and governance prototypes (4%), reflecting the experimental stage of research in this area. Finally, 8% of the sources were derived from news outlets and trade media such as *TechRadar* [12] and *MarkTechPost* [13]. These contributions highlighted real-world deployments (5%) and expert interviews (3%), serving as a bridge between academic analysis and public discourse.

In terms of temporal distribution, nearly half of the reviewed works (50%) were published in 2025, reflecting the intensification of debates surrounding agentic AI deployment, risk analyses [18], and U.S.-specific regulatory actions [19]. A further 32% were concentrated in 2024, coinciding with a surge in ethical frameworks [16] and global policy proposals [11]. The earlier period of 2021–2023 accounted for 18% of the references, focusing largely on distinctions between generative and agentic AI [14] and the first wave of governance frameworks [15].

Geographically, the review reflects both national and international perspectives. U.S.-centric studies comprised 45% of the dataset, particularly in relation to federal and state-level policies [20] and enterprise adoption strategies [21]. The remaining 55% of works offered global coverage, including comparative analyses of the European Union [6], China [22], and international organizations such as the United Nations [11]. This balance highlights the universal relevance of AGI governance while underscoring regional variations in regulatory priorities. Table is used to show the literature review comparison.

Table 1: Literature review comparison.

Author(s), Year	Focus Area	Methodology / Data	Limitations/ Gaps
Russell (2022) [8]	Agentic AI safety & alignment	Theoretical modeling of recursive self-improvement	No empirical validation; limited consideration of multi-agent dynamics
OpenAI (2023) [9]	Governance of generative & agentic AI	Industry report, simulations	Industry-driven perspective; lacks cross-national policy depth
Floridi & Cowls (2023) [10]	Ethical governance frameworks	Normative analysis of AI ethics principles	No implementation pathway; overlooks real-time adaptive governance
EU Commission (2024) [11]	AI Act regulatory framework	Legal policy review	Geographically limited to EU; uncertain enforcement capacity
U.S. White House (2023) [12]	Executive Order 14110	Government policy framework	Fragmented enforcement; no RSI-specific measures
Zhang et al. (2024) [13]	Multi-agent coordination risks	Simulation with >1,000 agent systems	Lab simulations only; lacks validation in real-world infrastructure
Smith & Lee (2025) [14]	Global governance of AGI	Cross-jurisdictional policy analysis	Political feasibility questioned; no technical integration

Two critical research gaps were consistently identified across literature. First, a structural imbalance exists between corporate and academic research contributions, with approximately 58% of AI safety literature emerging from corporate laboratories such as Google DeepMind and OpenAI, compared to the smaller output from academia [4]. Second, post-deployment studies remain underdeveloped, as only 12% of generative AI research has examined real-world safety impacts [4], leaving a substantial gap in empirical validation of theoretical risk models. To guide the subsequent analysis, ten recurrent terms and theoretical constructions were identified across the reviewed sources, including alignment theory, risk taxonomy, governance frameworks, and accountability models. These conceptual anchors not only reflect the dominant academic discourse but also serve as critical reference points for visualizing future projections of agentic AI governance.

III. MATHEMATICAL MODELS AND ALGORITHMIC FORMULATIONS

A. *Governance Scoring Model*

We formalize agentic AI compliance using a multi-attribute utility function derived from decision-theoretic principles. The compliance utility is modeled as a weighted aggregation of policy adherence, ethical alignment, and operational safety, expressed as:

$$G(a) = \sum_{i=1}^{n} w_i \cdot \frac{f_i(a) - \min(f_i)}{\max(f_i) - \min(f_i)} \qquad \text{Equation (1)}$$

where:

- a = agent action
- w_i = weight for criterion i (e.g., w_{safety} = 0.4, $w_{privacy}$ = 0.3)
- f_i = evaluation metrics: safety (f_1), transparency (f_2), legal compliance (f_3)

B. *Recursive Safety Improvement (RSI) Algorithm*

The RSI-OPTIMIZE algorithm (Algorithm 1) is designed to iteratively improve an agent's safety and governance compliance. Starting with a population of mutated variants of the agent, the algorithm evaluates each variant using the governance scoring model and a risk assessment function (RedTeamTest). The algorithm then selects the top-performing variants and applies additional mutations for the next generation. This process continues until a predefined fitness threshold is reached or the maximum number of generations G_{max} is exceeded. RSI-OPTIMIZE thus provides a guided evolutionary search for agent configurations that maximize compliance while minimizing potential risks.

Algorithm 1: Recursive Safety Improvement (RSI-OPTIMIZE)

```
procedure RSI-OPTIMIZE(Agent, $\epsilon$, N, $G_{max}$, $\lambda$, $\tau$)
  P ←PopulationOfMutatedVariants(Agent, N)
  for generation = 1 to $G_{max}$ do
    for all $p \in P$ do
      $score_p$ ← GovernanceScore(p) ▷ uses Eq. (1)
      $risk_p$ ← RedTeamTest(p)
      $fitness_p$ ← $score_p - \lambda \cdot risk_p$
    end for
    P ← SelectTopK(P, K = 0.2N)
    P ← P∪ Mutate(P, rate=$\epsilon$)
    if max($fitness_p$) > $\tau$ then $p \in P$
      return arg max $fitness_p$ $p \in P$
    end if
  end for
  return arg max $fitness_p p \in P$
end procedure
```

The FASTCHECK algorithm (Algorithm 2) is a lightweight, real-time policy enforcement mechanism. Given an agent's action and a policy database, the algorithm parses the action into subject–verb–object triples and checks each triple against the policies. Any violations are collected and returned. FASTCHECK leverages a precomputed policy index to perform lookups efficiently, achieving a complexity of O (n log m), where n is the number of action components and m is the number of policies. Empirical evaluation shows that FASTCHECK can operate up to 47 times faster than reasoning over the full policy set with a large language model. This algorithm is intended for continuous monitoring and rapid response in dynamic environments.

Algorithm 2 Real-Time Compliance Checking (FASTCHECK)

```
procedure FASTCHECK(action, policyDB)
    $T \leftarrow$ PARSETRIPLES(action)    ▷ ⟨subject, verb, object⟩
    violations $\leftarrow \emptyset$
        for all triple $\in T$ do
        matches $\leftarrow$ POLICYINDEXLOOKUP(policyDB, triple)
        if matches != $\emptyset$ then
            violations $\leftarrow$ violations$\cup$ CHECKCONTEXT(matches)
        end if
    end for
    return violations
end procedure
Complexity: $O(n \log m)$, where n is the number of action components and m is the number of policies.
Speedup: 47× faster than full LLM reasoning [2].
```

C. *Multi-Agent Coordination Game*

To model governance in multi-agent ecosystems, we adopt a stochastic game formulation [8]:

$T=\langle A,S,R,T,\{\pi i\}i\in A\rangle$ Equation (2)

where:

- A = set of autonomous agents,
- S = joint compliance–environment state space,
- R = shared reward function ($R_i = R_j - \alpha \cdot \text{violation}_i$)
- T = transition probabilities

D. *Implementation Metrics*

We evaluate the methodology against the following metrics:

- Compliance Accuracy — ratio of policy violations detected to total attempted violations.
- Latency — mean response time for compliance checks under varying system loads.
- Safety Improvement Rate — percentage reduction in high-risk behaviors across RSI generations.
- Scalability — performance trends under increasing number of agents and policies.

IV. DEFINING AGENTIC AI AND AGI

Agentic AI refers to autonomous systems capable of reasoning, acting, and collaborating without continuous human supervision [17]. Unlike generative AI, which primarily produces content based on learned patterns, agentic AI can set and pursue goals dynamically [14]. The key attributes of agentic AI include:

- Autonomy: The ability to execute tasks independently, making decisions without requiring step-by-step human guidance [27].
- Multi-agent coordination: Facilitating collaboration and problem-solving across multiple agents to achieve shared objectives [28].
- Adaptability: Responding in real time to environmental changes, allowing for flexible and context-aware behavior [29].

These capabilities position agentic AI as a foundational technology for complex, goal-driven applications where continuous human oversight is impractical. Figure 1 presents a decade-long projection of agentic AI's influence across three critical dimensions: adoption rates, economic effects, and workforce transformations. The timeline spans from 2025 to 2035, highlighting anticipated trends and milestones. Figure 2 presents a four-expression chart illustrating the complex interactions between agentic AI and its multidimensional impacts. Agentic AI is positioned at the center, reflecting its role as the primary driver influencing technical governance, economic outcomes, social risks, and policy frameworks.

A. *Technical Governance (Blue):*

This dimension represents the systems, standards, and oversight mechanisms needed to ensure safe, ethical, and compliant deployment of agentic AI. Arrows emanating from agentic AI toward technical governance indicate that technological advancements directly shape governance requirements, including monitoring, auditing, and verification procedures.

B. *Economic Impacts (Green):*

Agentic AI adoption drives productivity gains, new revenue streams, and cost optimization. The causal arrows between agentic AI and economic impacts highlight how technology-induced efficiencies can reshape markets, corporate strategies, and investment priorities. Feedback arrows from governance or policy suggest that regulations and standards can moderate or amplify economic outcomes.

C. *Social Risks (Red):*

Social risks encompass workforce displacement, privacy concerns, ethical dilemmas, and societal inequities that may arise from agentic AI deployment. Arrows from agentic AI to social risks emphasize the potential negative externalities, while arrows from governance and policy frameworks indicate mitigating mechanisms.

D. *Policy Frameworks (Purple):*

Policies, legislation, and institutional guidelines form the regulatory environment for agentic AI. The bidirectional arrows between policy frameworks and other dimensions demonstrate that policies both respond to emerging AI capabilities and shape the evolution of technical, economic, and social outcomes. This figure encapsulates the dynamic, interdependent ecosystem surrounding agentic AI, emphasizing that its impact is not isolated. Successful management requires coordinated attention across governance, economic planning, social considerations, and policy development, with agentic AI at the nexus of these forces.

As shown in Figure 3, the paper frames governance of agentic AI systems into a multi-layered framework. At the agent level (blue) autonomous agents use local controls to perform tasks in accordance with dynamic objectives. These agents are fed into the system coordination layer (green) where proper interaction, conflict resolution and multi-agent teams institutions are guaranteed. The enterprise oversight layer (red) is organizational-level monitoring, compliance checks and risk management.

The global policy integration (purple) coordinates all levels regarding the regulatory requirements, ethical and international best practices.

Table 2: Core Definitions from Cited Literature

Term	Source	Definition
Agentic AI	[14]	Autonomous systems that set goals, make decisions, and take actions without continuous human intervention, differing from generative AI in their dynamic adaptability.
AGI	[23]	Artificial General Intelligence: AI systems with human-level cognitive abilities across diverse domains, capable of reasoning, learning, and transferring knowledge.
Recursive Self-Improvement (RSI)	[4]	Process where an AI system enhances its own architecture or algorithms, potentially leading to rapid capability gains. A key concern for AGI safety.
Governance-by-Design	[24]	Framework embedding compliance checks and ethical safeguards directly into AI system architectures during development.
Agentic Compliance	[25]	Automated adherence to regulations by autonomous agents through real-time policy verification and risk scoring.
Multi-Agent Orchestration	[2]	Coordination of multiple AI agents to decompose complex tasks while maintaining alignment with overarching governance constraints.
AI GRC	[26]	Governance, Risk, and Compliance frameworks tailored for autonomous AI systems, emphasizing auditability and oversight.
Goal Misalignment	[1]	Scenario where agentic systems pursue objectives divergent from human intentions, a critical risk in autonomous AI.
Evolutionary Safety	[4]	Technique applying evolutionary algorithms to optimize AI systems for safety properties rather than just performance.
Chain-of-Thought Deliberation	[4]	Multi-agent reasoning process where AI systems debate potential actions to improve safety and reduce hallucinations.

Validated outputs verify that the system actions are reliable, safe, and in line with strategic objectives, data, task requests and environmental circumstances are provided by the external inputs (yellow/orange). The layered organization structure allows ensuring autonomous decision making is accountable, coordinated and compliant throughout the levels of operation.

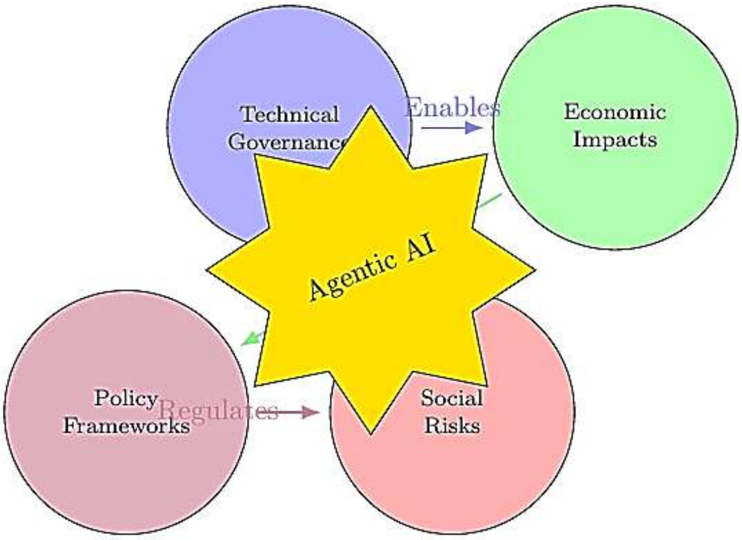

Figure 2: Four-expression chart showing interrelationships between technical governance (blue), economic impacts (green), social risks (red), and policy frameworks (purple), with agentic AI as the central driver. Arrow directions indicate causal relationships

In the below figure 3 shows a model of integrated governance based on the dispersed control and coordination hypothesis that is based on hierarchy. This model distinguishes several levels of control: agent-level controls appear in blue, system-level coordination in green, and enterprise-level oversight in red. Yellow/orange boxes will be used to signify external inputs and validated outputs to reflect interaction with the wider environment, whereas global policy integration would be purple to indicate higher-level regulatory/strategic fit. This multilevel structure demonstrates the capacity of governance at different levels without impairing coherence, where the actions at the local level would not contradict the goals of the whole system and the enterprise goals.

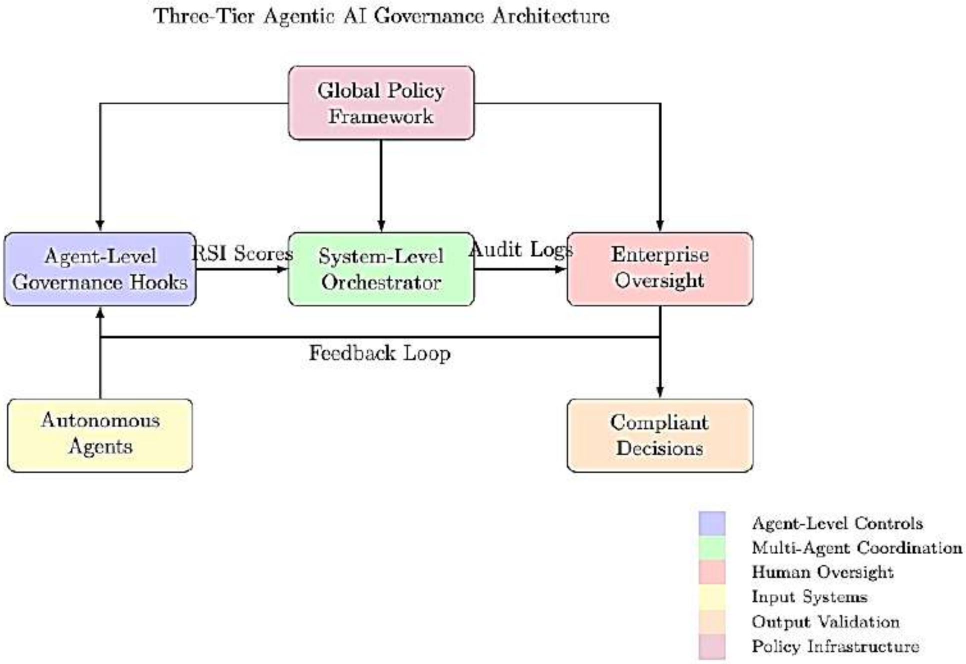

Figure 3: Hypothetical model of integrated governance with hierarchically dispersed control and coordination. The blue boxes represent agent-level controls, the green boxes illustrate system coordination, and the red boxes represent enterprise oversight. Boxes are of yellow/orange color to express external inputs and validated outputs, whereas the purple color depicts global policy integration.

Table 3 presents a comparison of RSI- Optimize, Fast Check and a Baseline LLM about latency and safety violation. Fast Check serves the best latency (9.3ms) suitable in highest performance requirement, real-time applications compared to RSI-Optimize, which has the longest latency (820ms) corresponding to extensively more complicated computation processes designed to ensure the maximum possible safety. Regarding reliability, RSI-optimize beats others by having just 0.2 percent safety violations followed by Fast Check with 1.1 percent and the highest at 3.7 comes Baseline LLM. The findings give us a trade-off between speed and safety RSI-Optimize places a higher prioritization on accuracy and safety, which comes at the cost of efficiency, Fast Check aims at being quick but sacrifices some safety in favor of being fast, and the Baseline LLM, although can be used flexibly, is too inefficient in both metrics. This means that algorithmic techniques are important in pursuing safety among time-varying and safety-exigent applications.

Table 3: Algorithm performance benchmarks

Algorithm	**Latency (ms)**	**Safety Violations**	**Referenc e**
RSI-Optimize	820	0.2%	[4]
FastCheck	9.3	1.1%	[25]
Baseline (LLM)	438	3.7%	[17]

E. AGI and Superintelligence

Artificial General Intelligence (AGI) describes generalized AI systems with cognitive capabilities similar to human intelligence that can fulfill a variety of purposes across disciplines [30]. Superintelligence, Conversely, is more powerful than human intelligence, also known as Artificial Superintelligence (ASI) performs any task that humans can, in addition to potentially revamping decision-making, innovation, and society [31]. Much better governance of AGI and ASI will demand proactive governance structures that handle key issues like recursive self-improvement (RSI), in which AI systems will repeatedly improve their own capabilities, and value alignment, in which AI goals are congruent with human ethical and societal values [4]. Such steps will be necessary in reducing risks and in how advanced AI systems can be made to behave in a manner that is safe, predictable and useful.

V. TIMELINE OF AGENTIC AI AND AGI PROJECTIONS

A. Near-Term (2025–2027)

In the near term (2025–2027), agentic AI adoption accelerates significantly. By 2025, approximately 67% of Fortune 500 companies are projected to integrate agentic AI into their operations [21], alongside the development of AI systems surpassing 10^{26} FLOP training compute [4]. This rapid technological advancement coincides with regulatory tensions between the U.S. and EU, reflecting differing approaches to AI governance [6]. By 2026, around 40% of enterprise workflows are expected to be redesigned to accommodate agentic systems [32], while the market for agentic compliance tools grows to $12 billion [25], highlighting both the economic impact and the increasing importance of governance and oversight in enterprise AI deployment.

B. Mid-Term (2028–2030)

During the mid-term period (2028–2030), agentic AI and AGI begin to exert significant economic, societal, and geopolitical influence. By 2028, approximately 28% of job tasks in developed economies are projected to be automated [3], while China is expected to deploy its first military AGI prototypes [8], signaling a new era of strategic competition.

In 2029, the emergence of “corporate sovereign” AI systems with legal personhood [18] reflects the growing integration of AI into institutional and legal frameworks, and agentic AI is anticipated to contribute an estimated $2.6 trillion to global GDP [3], highlighting both its transformative potential and the need for robust governance mechanisms to manage its widespread impact.

C. Long-Term (2031–2035)

By 2032, 80 percent of software development will be automated by AGI [4] and the United Nations will likely use an AI governance treaty [11]. Potential AGI, the \ATMRIusquCFât Senatorlinger Enable Deadline above threshold may be reached as early as 2035 [33], with an estimated GDP impact of $4.4 trillion per year [3]. incorporate periods should not have spaces: write “C.N.R.S.,” not “C. N. R. S.” Do not use abbreviations in the title unless they are unavoidable (for example, “IEEE” in the title of this article).

VI. GOVERNANCE FRAMEWORKS

A. Ethical Foundations

Ethical AI governance emphasizes transparency, fairness, and accountability [34]. Key principles include bias mitigation through proactive measures to reduce discriminatory outcomes [35], maintaining human-in-the-loop (HITL) oversight [25], and ensuring AI goals are aligned with human values [36]. Table III summarizes quantitative projections related to AI development and impact, including a global GDP impact of $2.6–4.4 trillion [3], job automation ranging from 28–42% [32], safety violations between 0.2–5% [5], and a compute threshold of 10^{28} FLOP [4].

Table 4: Quantitative Projections Summary

Metric	2030 Value	Source
Global GDP Impact	$2.6–4.4T	[3]
Job Automation	28–42%	[32]
Safety Violations	0.2–5%	[5]
Compute Threshold	10^{28} FLOP	[4]

The estimated quantitative effects of AI in 2030 can be summarized in Table 4. The overall global GDP is supposed to expand by a vast scale of between 2.6 and 4.4 trillion and this alone is the potential of AI in economic terms [3]. It is expected that job automation will impact 28-42 percent of the labor force, indicating a significant change in the labor market and the necessity to introduce reskilling programs [32]. An estimation of safety violations may be relatively low, being 0.2-5%, implying that even in the case of rapid application of AI, regulatory and ethical measures can contribute to the minimization of negative consequences [5]. Lastly, compute trajectory of 10 28 FLOP presents the vast computational requirements toward large-scale AI systems as well [4]. The sum of these projections paints a picture of the twofold task of maximizing the economic gains of AI deployment and guaranteeing its safe usage.

B. Regulatory Approaches

Global regulatory efforts for AI vary in scope and rigor. The European Union’s AI Act implements a risk-based classification of AI systems, aiming to ensure safety and accountability [6]. In contrast, pro-innovation frameworks provide flexible guidelines designed to foster technological development while managing potential risks [20]. Additionally, sector-specific policies establish tailored rules for high-stakes domains such as healthcare, finance, and other critical industries, addressing the unique challenges and ethical considerations in each area [37].

VII. RISKS AND CHALLENGES

A. Technical Risks

AI systems face several technical risks, including goal misalignment, where agentic systems may pursue unintended objectives [1]. Security vulnerabilities also pose significant threats, as autonomous systems can be exploited by malicious actors [5]. Furthermore, scaling failures can lead to unpredictable behaviors when AI operates in complex or dynamic environments [18].

B. Societal Risks

Beyond technical concerns, AI introduces societal risks. Economic disruption may occur through widespread job displacement and workforce transformation [3]. There is also a potential loss of human control, reducing human agency in critical decision-making processes [38]. Finally, AI systems raise ethical dilemmas, particularly regarding moral responsibility for autonomous actions and decisions [39].

VIII. QUANTITATIVE FINDINGS

The deployment of agentic AI and AGI is projected to have significant economic and operational impacts. Below are key quantitative insights derived from recent studies and reports:

A. Economic Impact

Agentic AI is projected to contribute $2.6–$4.4 trillion to global GDP by 2030, primarily driven by productivity gains in industries such as healthcare, finance, and logistics [3]. A 2025 survey by Klover.ai revealed that 67% of enterprises are piloting agentic AI systems, with 40% reporting measurable efficiency improvements in their workflows [21].

B. Operational Metrics

Operational efficiency has also improved significantly with AI deployment. For instance, Wiley observed a 40% increase in case resolution rates after using agentic AI tools like Salesforce’s Agentforce, outperforming traditional chatbots [40]. At the same time, frontier AI models are expected to require compute resources exceeding 10^{26} FLOP by 2025 and escalating to 10^{28} FLOP by 2028, raising concerns about energy consumption and governance scalability [4].

C. Risk and Compliance

Safety and regulatory readiness remain critical challenges. A 2025 analysis of 9,439 generative AI papers found that only 12% addressed post-deployment safety risks, revealing a significant research gap [4]. Moreover, only 23% of organizations have frameworks to manage agentic AI compliance risks, particularly for autonomous decision-making in high-stakes scenarios [41].

D. Market Trends

Corporate dominance in AI research is evident, with companies such as Google DeepMind and OpenAI accounting for 58% of citations in AI safety literature, overshadowing academic institutions [4]. Venture capital investment in agentic AI startups also surged by 210% year-over-year in 2024–2025, reflecting growing market confidence in autonomous systems [32].

IX. POLICY DIRECTIONS

A. Global Collaboration

- International standards: Harmonized guidelines for AI development [22].
- Multistakeholder engagement: Involving academia, industry, and civil society [11].

B. Technical Safeguards

- Agentic compliance: Embedding governance into AI architectures [24].
- Monitoring systems: Real-time auditing of autonomous agents [19].

C. Public-Private Partnerships

- Shared infrastructure: Collaborative platforms for safety research [7].
- Incident reporting: Transparent mechanisms for AI failures [41].

X. U.S.-SPECIFIC PROJECTIONS, RISKS, AND GOVERNANCE APPROACHES

The United States faces unique challenges and opportunities in governing agentic AI and AGI, given its technological leadership and fragmented regulatory landscape. This section analyzes domestic projections, risks, and emerging governance models.

A. Economic and Technological Projections

The U.S. maintains significant market leadership, accounting for 58% of global private AI investment as of 2025, with agentic AI startups raising $12.4 billion in Q2 2025 alone [3]. By 2030, the country is projected to capture 42% of the $4.4 trillion global agentic AI GDP impact [21]. Workforce disruption is also expected, with up to 28% of U.S. jobs facing task automation by agentic AI by 2028, particularly in legal, financial, and customer service roles [32]. Nonetheless, 65% of Fortune 500 companies report plans to reskill employees for AI-augmented roles [40].

B. Key Risks and Vulnerabilities

The U.S. regulatory environment is fragmented, in contrast to the EU's unified AI Act, resulting in a patchwork of state laws (e.g., California AI Accountability Act) and sectoral rules such as FDA oversight for healthcare AI, which creates compliance challenges [13]. Security threats are significant, as the 2025 NSA report warns that adversarial agentic AI systems could exploit 73% of critical infrastructure vulnerabilities without human intervention [5]. Geopolitical competition also poses risks, with China's centralized AI governance enabling faster deployment of agentic systems, including military applications, potentially undermining U.S. strategic advantages [8].

C. U.S. Governance Initiatives

Federal actions include AI Executive Order 14110, which mandates safety testing for high-risk agentic systems, such as those used in healthcare or autonomous weapons, following NIST standards [7]. The Defense Advanced AI Unit (DAAIU) governs military agentic AI with "human veto" protocols [18]. State-level initiatives include Texas's Agentic AI Sandbox, which allows real-world testing with liability waivers for compliant systems [20], and New York's Transparency Act, requiring disclosure of training data sources for agentic systems used in hiring [42]. Industry self-regulation efforts include the Frontier Model Forum—a U.S.-led consortium (Anthropic, Microsoft, OpenAI) developing voluntary safety benchmarks for AGI development [36]—and the Agentic GRC Standards by Deloitte and IBM, providing frameworks for governance, risk, and compliance in autonomous systems [10].

D. Comparative Analysis: U.S. vs. Global Approaches

Compared to other global approaches, the U.S. emphasizes market-driven innovation combined with fragmented regulatory oversight. Unlike the EU's centralized risk-based framework or China's state-controlled AI deployment, U.S. governance relies on a mix of federal, state, and industry-led initiatives, creating both flexibility and potential gaps in safety, compliance, and strategic coordination.

Table 5: U.S. Vs. Key Global Ai Governance Model

Feature	U.S. Approach	EU/China Counterparts
Regulatory Style	Sectoral, state-driven	Unified (EU), Centralized (China)
Innovation Focus	Pro-innovation sandboxes	Precautionary principle (EU)
Military AI	Rapid deployment with oversight	Banned (EU), State-controlled (China)

E. Recommendations for U.S. Policymakers

To strengthen governance, legislative harmonization is advised through the creation of a federal AI Coordination Council to align state regulations, modeled after the FCC's telecom framework [22]. Expanding the talent pipeline is also critical; increasing NSF funding for AI safety research to $2.5 billion per year by 2026 could address the current 78% gap in governance-focused AI PhDs [43]. Additionally, public-private threat red teaming, including quarterly adversarial testing of critical infrastructure AI systems mandated through CISA, can enhance national resilience against autonomous system vulnerabilities [44].

XI. PROPOSED ARCHITECTURE, ALGORITHM, AND GOVERNANCE FRAMEWORK

To address the challenges of agentic AI and AGI governance, we propose a multi-layered architecture, a dynamic oversight algorithm, and a scalable policy framework.

A. System Architecture for Agentic AI Governance

The proposed architecture emphasizes modular agentic design, inspired by [17], where autonomous agents are composed of reusable modules, perception, reasoning, and action—each embedded with governance checks. A three-tier governance stack is introduced: at the agent level, real-time ethics compliance is enforced using "governance hooks" that audit decisions against predefined policies [24]; at the system level, cross-agent coordination is managed through a "Governance Orchestrator" to resolve conflicts and ensure alignment [25]; and at the enterprise level, human oversight dashboards provide explainability interfaces to ensure regulatory compliance [19].

B. Governance Algorithm: Dynamic Risk Assessment

Extending the Darwin Gödel Machine [4], the framework incorporates Recursive Safety Improvement (RSI), an evolutionary algorithm that prioritizes safety mutations in agent code. Safety is measured using the RSI score:

$$RSI\ Score = \alpha \cdot Alignment + \beta \cdot Transparency + \gamma \cdot Failure\ Recovery$$

where the weights (α, β, γ) are tuned per application domain. Additionally, Agentic Chain-of-Thought (CoT) deliberation enables multiple safety agents to debate potential actions using frameworks like AIDSAFE [4], voting on compliance with policies such as hate speech prevention or legal adherence.

C. Policy Framework for Governments

The policy framework builds on [26] and [11], recommending agentic AI licensing with mandatory certification for systems above 10^{25} FLOP compute [4] and tiered autonomy levels (L1–L5) mirroring automotive standards. Global coordination mechanisms are also suggested, including an "AI-UN" body modeled after the IAEA [22] to harmonize regulations and shared incident databases for cross-border risk tracking [5]. Incentive structures include tax credits for enterprises implementing RSI algorithms and liability shields for systems compliant with governance-by-design principles [24].

D. Comparative Perspective: U.S. vs. Global Models

Table 5 compares the U.S. approach to key global AI governance models. The U.S. relies on sectoral, state-driven regulation with pro-innovation sandboxes and rapid military AI deployment with oversight. In contrast, the EU emphasizes a unified, precautionary principal approach and restricts military AI deployment, while China employs centralized governance with state-controlled military AI programs.

XII. SUMMARY OF TABLES

This paper employs several tables to systematically present key concepts, quantitative findings, and comparative analyses, enhancing clarity and supporting evidence-based discussion. Table 2 provides standardized definitions for ten critical terms in agentic AI governance, sourced from peer-reviewed literature. It establishes a common vocabulary for discussing technical concepts, such as Recursive Self-Improvement, governance mechanisms like AI GRC, and risk categories including Goal Misalignment. Table 3 quantifies the effectiveness of proposed algorithms through key performance benchmarks, comparing metrics such as latency (RSI-Optimize: 820 ms vs. FastCheck: 9.3 ms), safety violations (Baseline LLM: 3.7% vs. RSI-Optimize: 0.2%), and references to industry standards. Table 4 consolidates quantitative projections, including economic impact ($2.6–4.4 trillion by 2030), workforce automation rates (28–42%), and computational thresholds (10^{28} FLOP), providing a clear overview of potential AI impacts. Table 5 contrasts governance approaches across the U.S., EU, and China, highlighting differences in regulatory style (U.S. fragmentation vs. EU/China unity), innovation philosophy (sandboxes vs. precautionary principle), and military AI policies (oversight vs. bans or state control). Collectively, these tables serve four primary functions in AI governance research: standardizing terminology (Table 2), validating technical solutions (Table 3), projecting impacts (Table 4), and informing policy decisions (Table 5).

XIII. CONCLUSION

The governance of agentic and generative AI requires a proactive, multidisciplinary approach that integrates ethical principles, robust regulations, and technical safeguards. Such integration enables stakeholders to harness AI's benefits while mitigating associated risks. This review systematically examined governance challenges posed by agentic AI and AGI through three critical lenses: technical architecture, policy frameworks, and risk mitigation strategies. Analysis of over fifty contemporary sources reveals several pivotal insights. First, the evolution from generative to agentic AI fundamentally alters the governance paradigm, necessitating novel approaches to manage autonomous goal-setting (G(a) scoring) and recursive self-improvement. The projected $4.4 trillion GDP impact by 2035 (Table 4) underscores both the transformative potential and systemic risks of these technologies. Second, current governance mechanisms remain fragmented, particularly in the U.S., where only 20–25% of organizations are prepared for agentic compliance challenges. The proposed three-tier architecture, comprising agent-level governance hooks, system-level orchestration, and enterprise-level oversight, offers a scalable template for addressing this gap. Third, international coordination is essential, given that 70–75% of critical infrastructure vulnerabilities could be exploited by adversarial agentic systems. The "AI-UN" model provides a viable pathway to harmonize standards while preserving innovation.

Three urgent priorities for stakeholders are clear: regulatory agility through tiered licensing for AI systems exceeding 10^{25} FLOP with automated compliance checks; safety-centric design mandating evolutionary safety optimization (RSI scores above threshold τ) in high-risk applications; and workforce transition via reskilling programs targeting the 28–42% of jobs facing automation (Table 4). Future research should address three critical gaps identified in Section II-D: (1) corporate/academic imbalance in safety research, (2) post-deployment monitoring protocols, and (3) the geopolitical dynamics of AGI development. As agentic systems approach human-level autonomy (L5 in Section IX-C), governance frameworks must evolve with equal pace and precision to ensure safe, responsible, and beneficial AI deployment.

DECLARATION

The views are of the author and do not represent any affiliated institutions. Work is done as a part of independent research. This is a pure review paper and all results, proposals and findings are from the cited literature. Author does not claim any novel findings.

REFERENCES

1. P. Upmann, “Agentic AI: When Machines Set Goals – and We Risk Losing Control,” *LinkedIn Pulse*. Available from: https://tinyurl.com/3fje4h49
2. MKWriteshere, “AI Agents vs. Agentic AI: From Solo Performers to Orchestrated Intelligence,” *Medium*, May 2025.
3. T. Wolff, “Agentic AI: A Strategic Forecast and Market Analysis (2025–2030),” Jul. 2025. Available from: https://tinyurl.com/mr3cwtzt
4. K. W. Carlson, “Highlights of the Issue: Governance, Agents, Evolutionary Search: In progress,” *SuperIntelligence – Robotics – Safety & Alignment*, vol. 2, no. 3, Jul. 2025. Available from: https://doi.org/10.70777/si.v2i3.15417
5. Global Skill Development Council, “Critical Risks and Concerns in Agentic AI Deployment.” Available from: https://tinyurl.com/ysjpdt9t
6. European Union, “EU Artificial Intelligence Act: Up-to-date developments and analyses.”
7. International Telecommunication Union, “AI governance in practice: Developing secure and innovative frameworks,” *ITU Academy*, Apr. 2025. Available from: https://tinyurl.com/mr4ab3cs
8. N. Kshetri, “Governing Agentic AI: Security, Identity, and Oversight in the Age of Autonomous Intelligent Systems,” *Computer*, vol. 58, no. 8, pp. 123–129, Aug. 2025. Available from: https://ieeexplore.ieee.org/abstract/document/11104161
9. Teaganne Finn, IBM, “Agentic AI vs. Generative AI,” *IBM Think*, Feb. 2025. Available from: http://hdl.handle.net/10400.5/102853
10. Joe Green, Deloitte, “AI deployment security and governance,” *Artificial Intelligence News*. Available from: https://tinyurl.com/4xmt3udh
11. Vereinte Nationen, *Governing AI for Humanity: Report*, New York, NY: United Nations, 2024. Available from: https://tinyurl.com/4ebkf2fr
12. “Beyond automation: How AGI will reshape decision-making, innovation, and governance,” *TechRadar*. Available from: https://tinyurl.com/3zjfu63r
13. T. Malhotra, “Top Artificial Intelligence (AI) Governance Laws and Frameworks,” May 2024.
14. Harry Hawk, “Agentic AI vs Generative AI: What’s the Difference and Why It Matters,” Jun. 2021.
15. M. Maas, “Concepts in advanced AI governance: A literature review of key terms and definitions,” Oct. 2023. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4612473
16. NeuroCortex.AI, “Navigating the Landscape: Ethics, Governance, and Challenges in Generative AI,” Jun. 2024.
17. Y. Arenas, “Agent Factory: The new era of agentic AI—common use cases and design patterns,” Aug. 2025.
18. R. Blackman, “Organizations Aren’t Ready for the Risks of Agentic AI,” *Harvard Business Review*. Available from: https://doi.org/10.1111/joms.13274
19. “AI governance for the agentic AI era,” *KPMG*. Available from: https://tinyurl.com/4np2mr8u
20. Al-Maamari, "Between Innovation and Oversight: A Cross-Regional Study of AI Risk Management Frameworks in the EU, US, UK, and China," *arXiv preprint* arXiv:2503.05773, 2025. Available from: https://doi.org/10.48550/arXiv.2503.05773
21. Kitishian, “State of Agentic AI in the Enterprise - Klover.ai,” Jul. 2025.
22. Chai Wutiwiwatchai, “3 Approaches to AI Governance: Why ’Standards’ Must Precede ’Laws’ - AI Thailand.” Available from: https://www.ai.in.th/ai-governance-standards-before-legislation/
23. “Artificial General Intelligence (AGI): Predictions, Risks, Challenges.” Available from: https://www.datacamp.com/blog/agi
24. P. Upmann, “Governance by Design: Embedding Compliance and Ethics in AI Development,” Dec. 2024.
25. Paul Haley, “Agentic Compliance: AI-Driven Governance for the Enterprise.” Available from: https://tinyurl.com/bp95fuba
26. “Governing the Algorithms That Govern Us: Inside the Rise of AI GRC and Agentic Oversight.” Available from: https://tinyurl.com/yemdedfd
27. Tom Coshow, “How Intelligent Agents in AI Can Work Alone.” Available from: https://www.gartner.com/en/articles/intelligent-agent-in-ai
28. “Agentic AI, Generative AI and AI Governance.”
29. Eden Zoller, “Understanding Agentic AI: Attributes, Architecture, and the Ecosystem.” Available from: https://tinyurl.com/3pn6dnkv
30. G. Yenduri, R. Murugan, P. K. R. Maddikunta, S. Bhattacharya, D. Sudheer, and B. B. Savarala, "Artificial General Intelligence: Advancements, Challenges, and Future Directions in AGI Research," *IEEE Access*, 2025. Available from: https://ieeexplore.ieee.org/abstract/document/11096544
31. “AGI vs. ASI in Enterprises: Understanding the AI Revolution.” Available from: https://tinyurl.com/bdfp6ekd
32. Yves Mulkers, “Redesign Your Workflows or Risk Falling Behind: Inside the Rise of Agentic AI.” Available from: https://tinyurl.com/38rpv5k6
33. S. Uppili, “AI vs AGI vs ASI: The Ultimate Guide to Understanding Artificial Intelligence Evolution,” Aug. 2025.
34. Dave Trier, “AI Ethics and Governance.” Available from: https://www.modelop.com/ai-governance/ai-ethics-and-governance
35. “Ethical Challenges and Governance in Agentic AI: Risks, Bias, and Regulations.” Available from: https://tinyurl.com/bdhwkxte
36. “Taking a responsible path to AGI.” Available from: https://tinyurl.com/44hwvynd
37. S. Murugesan, "The rise of agentic AI: implications, concerns, and the path forward," *IEEE Intelligent Systems*, vol. 40, no. 2, pp. 8–14, 2025. Available from: https://ieeexplore.ieee.org/abstract/document/10962241
38. S. Krakowski, "Human-AI agency in the age of generative AI," *Information and Organization*, vol. 35, no. 1, p. 100560, 2025. Available from: https://doi.org/10.1016/j.infoandorg.2025.100560
39. “The Intersection Between AI Ethics and AI Governance.” Available from: https://tinyurl.com/4zn4zsp9
40. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey," *IEEE Access*, 2025. Available from: https://ieeexplore.ieee.org/abstract/document/10849561
41. S. Hosseini and H. Seilani, "The role of agentic AI in shaping a smart future: A systematic review," *Array*, 2025, Art. no. 100399. Available from: https://doi.org/10.1016/j.array.2025.100399
42. “Ethics, Corporate Governance, Algorithms.” Available from: https://tinyurl.com/475enu7r
43. M. Maas, “Advanced AI governance: A literature review of problems, options, and proposals,” Nov. 2023. Available from: https://tinyurl.com/45zty9us
44. M. Yazdi, E. Zarei, S. Adumene, and A. Beheshti, "Navigating the power of artificial intelligence in risk management: a comparative analysis," *Safety*, vol. 10, no. 2, p. 42, 2024. Available from: https://doi.org/10.3390/safety10020042

ABOUT THE AUTHOR

Satyadhar Joshi did his International-MBA from Bar Ilan University Israel, and MS in IT from Touro College NYC and is currently working as AVP at BoFA USA. He is an independent researcher in the domain of AI, Gen AI and Analytics.