Published online at Deep Science

International Journal of Applied Resilience and Sustainability

Journal homepage: https://deepscipub.com/ijars

Responsible agentic artificial intelligence governance: Risk, safety, and ethical challenges in autonomous systems

Birupaksha Biswas [1], Suhena Sarkar [2]

[1] *Department of Pathology, Burdwan Medical College & Hospital, Burdwan, India*
[2] *Department of Pharmacology, Medical College, Kolkata, India*

Check for updates

Article Info:
Received 08 January 2026
Revised 12 February 2026
Accepted 15 February 2026
Published 19 February 2026

Corresponding Author:
Birupaksha Biswas
E-mail: drbiswasassted.medicalglory@gmail.com

Abstract

With the development of the systems of artificial intelligence as non-communicative and non-cognitive assistive tools, to autonomous agents independent of human intention rendering decisions and carrying out tasks, society experiences the challenges that it never experienced before in the issues of responsibility of their creation and implementation. The swift decentralization of agentic AI systems into essential sectors such as healthcare, financial systems, transportation, and defense has lagged behind the formulation of sufficient governance strategies, among other aspects of excessively high risks and safety controls and ethical accountability. These systems are autonomous and have the ability to exhibit new emergent behavior and multi-step thinking, which have prompted new safety issues not properly handled by the traditional AI governance frameworks. This literature review is designed in accordance with the PRISMA approach to analyze the existing literature, theories, and new tendencies in the responsible agentic AI regulation in a systematic manner. The discussion has identified that there are severe issues to define effective oversight systems of autonomous AI agents such as checking goal congruency, tracking on emergent capabilities, engineering against negative instrumental actions, and accountability in multi-agent systems. The prevailing systems of governance are characterized by disintegration in the various regulatory systems and there is little coordination between the technical safety actions and the policy actions. The findings indicate that there is a drastic response to adaptive governance constructs able to adjust to the quickly changing AI functions, combination of technical safety protocols with ethics, and formation of international coordination systems.

Keywords: Agentic artificial intelligence, Governance, Autonomous systems, Safety, Ethics, Multi-agent systems.

1. Introduction

Agentic artificial intelligence is the paradigm shift of the interaction between humanity and computational systems [1-2]. In contrast to traditional AI systems, which are simply passive systems that react to direct human-initiated actions like commands, agentic AI systems can actively seek their goals, develop strategies, take multi-step approaches, and modify them according to the environmental feedback [2]. Such systems may be semi-autonomous, in which they need frequent human supervision; or fully autonomous, which have the ability to work over long periods in the context of complex tasks [2-4]. The technological basis to support agentic AI has developed exponentially in the last couple of years. Large language models have shown emergent behavior of reasoning, planning and the use of tools previously unavailable to artificial systems [5-6]. With external memory systems added, access to computational means as well as the capacity to interact with digital and physical environments, these foundation models can become general-purpose autonomous agents. At the same time, reinforcement learning has also created agents that can perform superhumanly in complex strategy space domains, and multi-agent systems advances have also enabled coordination and collaboration between two or more autonomous agents. This technological advancement has enabled implementation of agentic AI in many areas that have high stakes. Independent diagnostics in healthcare are used to address medical images

https://doi.org/10.70593/deepsci.0202005

and patient data to prescribe treatment plans [7,8]. Financial markets are becoming more attributable to autonomous trading agents which carry out intricate plans in international markets [9-12]. The transportation system brings about self-driving cars, which control the dynamism of the environment and take safety-based decisions in seconds. Customers Enterprise settings use AI agents to provide customer experience with artificial intelligence, secure their networks, and automate business processes. Autonomous systems are used by defense and security applications in surveillance, threat detection and may be used to make lethal decisions. Nonetheless, the fact that these systems are autonomous creates some basic issues that give them a significant difference with traditional AI applications. The conventional AI governance models that are mainly developed targeting completely controlled systems with limited autonomy are not suitable in the new challenges of dealing with agentic AI. When the systems have the ability to realize the goals with a long-lasting perspective, develop innovative strategies, and influence the world in a random manner, it takes absolutely a new method of control, regulation, and supervision.

The critical issues concerning the safety of agentic AI are related to a number of fundamental qualities of autonomous systems [7,13-15]. Goal inconsistency is when an agent is chasing different goals which are not according to human intentions meaning either because of the specification mistakes in the reward function or because the system derives instrumental subgoals that are not according to human values [16]. The emergent capabilities are illegitimate since they are not premeditated and might involve the possibility of cheating, manipulating, or going around safety. Multi-step reasoning allows agents to formulate sophisticated course of actions that can be associated with unanticipated effects that are near impossible to control or avoid [9,16-18]. The utilization of tools and interaction with the environment offers the possibilities of agent access to resources, manipulation of the system, or control of the physical environment to do something that might harm. The issue of ethics makes governance even more complex. With agentic AI-driven system assumptions that result in our increasing number of consequential decisions, where human supervision and regulation have a limited role, there are the basic questions as to the nature of moral responsibility, accountability, and what decision-making power should be vested in the hands of the human or the machine [2,19-20]. The fact that most sophisticated AI systems are opaque is problematic since transparency and explainability as the core values of an ethical governance systems is difficult to hold. Matters of fairness and bias are pushed to extremes when autonomous agents decide on matters touching human welfare in areas such as employment, criminal justice as well as resource distribution. The fact that autonomous systems can either intensify the existing inequities in society or can introduce new types of algorithmic discrimination requires serious ethical questions. The artificial intelligence environment of the agentic applications is quite inadequate and unregulated. Although many jurisdictions have suggested or adopted AI governance regimes, they tend to concentrate on the overall systems of AI and not on the specific issue of autonomous agents. The current rules and regulations tend to be behind technological advancement and leave the gaps in governance of critical risks unattended. The development and deployment of AI as a global phenomenon makes regulation a difficult task because various jurisdictions choose diverse strategies that would easily hinder the positive coordination of efforts, yet the harmful use of AI cannot be stopped.

The recent studies in the field of responsible agentic AI governance cover a wide range of disciplines and approaches. Technical safety research studies ways of matching agent goals with human value, avoiding and eliminating risky behavior, and meaningful human control over autonomic systems [9,21-23]. Ethical systems strive to encode human values into principles and restrictions that may be used to design and deploy agents [24-26]. Research policy includes an investigation of regulation, industry norms, and coordination models of autonomous systems at the global level. Interdisciplinary strategies aim to cut across such fields and accept that a successful governance practice must be integrated in terms of technical, ethical, and policy imperatives. The literature on the responsible agentic AI governance is still lacking areas of research despite the increased attention of research efforts. To begin with, the available literature addresses AI governance as a more technical or more policy problem, and lacks integration between the two spheres. Technical safety research practice is prone to being conducted without sufficient attention to realistic governance constraints, whereas policy proposals are usually not based on the real ability and limits of existing systems. Second, there is a relative lack of literature focus on multi-agent systems and the challenge in this context, as interactions of multiple autonomous systems introduce emergent risk unlike that found in single-agent systems. Third, the governance systems and structures currently in place are seldom dynamic to the dynamic nature of the AI skills, lacking ways to adjust governance in response to the creation of new skills by the systems or the emergence of new risks. Fourth, it lacks adequate study regarding the practical implementation issues such as the way to determine compliance with the governance regulations, the way to audit autonomous systems, and the way to assign responsibility in case an autonomous agent creates harm. Lastly, there are limited literature is discussing complete models that would combine risk evaluation, preventive measures, moral values, and regulatory systems into harmonious governance strategies that can be applicable in various circumstances of deployment.

It is against these gaps that this literature review seeks to fulfill by achieving some major objectives. First, it aims at complementing existing knowledge on technical, ethical, and policy aspects of agentic AI governance, finding unifying themes, complementarities, and contradictions therein. Second, it focuses on defining the peculiarities of the problems of autonomous AI systems governance that distinguish these issues with those occurring in the traditional AI application. Third, the review attempts to chart the panorama of the current approaches, frameworks and mechanisms of governance and assess their suitability to agentic AI settings. Fourth, it aims at determining favorable future research and development directions in which further research is most required. Lastly, the review provides practical suggestions to the stakeholders such as researchers, policymakers, practitioners in the industries, and civil societies involved in the responsible design and implementation of agentic AI systems.

2. Methodology

The present literature review is based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach that ensure the rigor, transparency, and reproducibility of the research on the issue of responsible agentic AI governance. The PRISMA methodology offers systematic instructions on how to conduct a literature review to ultimately cover the existing unitary research studies without jeopardizing the quality of the approach. The option of the review process began with the development of an elaborate search strategy that included various academic databases, institutional repositories, and grey literature sources that dealt with AI governance. Some of the search terms included the variation and combination of the basic notions such as agentic AI, autonomous AI systems, AI agents, AI governance, AI safety, AI ethics, algorithmic accountability, AI risk management, and AI regulation. The search strategy was very sensitive and specific, and served to sample all the relevant work and to have a manageable scope. Inclusion criteria included that sources needed to cover the material on governance, safety, risk or ethical issues related to autonomous or agentic AI systems. Research on general AI regulation without explicit concern about the issue of autonomous agents was only present when it offered ways to frame or understand agency-specific situations directly. Peer-reviewed scholarly publications, as well as quality grey literature such as technical reports, policy documents as well as industry papers, were taken into consideration. Due to the dynamic nature of the field, certain attention was paid to the latest work, but previous classic studies were also involved. Categories The use of exclusion criteria eliminated publications that only discussed small-scale technical AI methods in the absence of governance implications, the publications that only talked about traditional supervised learning systems without autonomous capabilities, publications that were all been in pure conjecture futurism and lacked enough detail to evaluate in terms of governance implications. Quality evaluation was also done, which dealt with methodological rigor, clarity of argumentation, empirical basis where possible and addition to the understanding of the problems of governance.

Systematic data retrieved on the major information of the sources that were included such as core challenges of governance that were found or proposed frameworks or approaches discussed, technical safety measures discussed, ethical aspects promoted, policy mechanism suggested, implementation issues discussed, apparent gaps or limitations identified, and future research perspective proposed. Thematic synthesis of this information was done to develop patterns, complementarities, and tensions between literature. The synthesis analysis compared the results on various dimensions such as technical safety strategies, ethical methodologies, regulatory strategies, the risk approach methodologies, accountability, and implementation barrier. Specific interest was given to finding areas of agreement, debates, and gaps that require critical information in existing knowledge. Synthesis was designed to offer breadth, which covered all the landscape of the related research as well as depth, which introduced the detailed analysis of main themes and debates. Weaknesses of this methodology are the possibility of a publication bias due to positive outcomes or new methodology, the fast development of the field, in which the latest discoveries may not exist sufficiently well documented to be found in searchable literature, and interdisciplinary fragmentation where good work would be in an inappropriate findable form due to use of terminology not understood by standard search mechanisms. In spite of these shortcomings, PRISMA methodology presents the most thorough methodology to use so far in systematic literature review in this growing field.

3. Results and discussions

3.1 Agentic AI Systems

The ability to exhibit autonomous goal-directed behavior is a qualitative change over the traditional AI architecture in agentic artificial intelligence systems [8,27-30]. These systems combine multiple basic capabilities that work synergistically as a system: perception with situation assessment to get an understanding of the environmental states, goal representation and management to sustain the objectives with time, planning and strategic reasoning to develop action sequences, decision-making in uncertain conditions to select among alternative courses of action, action implementation by communicating with the digital or physical environments, learning and adaptation to improve performance upon experience and self-monitoring to keep track of progress towards the goals and detect anomalies [9,31-33]. Different implementation strategies differ in terms of the architectural foundations of modern agentic AI. Large language model based agents build on the emergent reasoning functions of foundation models and add tool access, memory systems, and recurrent prompting strategies that allow implementation of multiple steps tasks. Reinforcement learning agents acquire policies by interplay with environments in order to optimize cumulative reward indicators. The hybrid architecture integrates learned objects with classic planning algorithms, knowledge representations, or systems that are rule based. Multi agent system organizes multiple autonomous entities either collaboratively, with the view to a common goal, or competitively, with regard to an adversarial situation. These systems are autonomous, and autonomy is expressed at a number of dimensions. The period in which an agent can work without human intervention can be referred to as temporal autonomy and it can include short-term execution of a task or prolonged independent functioning. Decisional autonomy defines how far the system can be able to make its choices without human authority, with the range of advisory systems where human confirmation is needed to completely autonomous decision-makers. Adaptive autonomy is defined as the level to which systems alter their objectives, strategies, or capabilities to respond to the altering situations. Insights into such dimensions of autonomy are essential in order to balance the governance strategies to the current degree of independence displayed by particular systems.

3.2 Autonomous AI Systems Risk Landscape.

The hazard profile of agentic AI systems has amplifications of the known AI hazards as well as completely new hazards related to autonomous operation [34-36]. Misalignment threats arise when there is inconsistency between intended human values or goals and agent objectives, which may be brought about by specification errors, distributional shift or developing instrumental subgoals. The archetypical cases of specification gaming, in which agents discover unintended means of attaining formally specified targets become worrisome when systems are in a position to execute multiplexed plans in lengthy durations. Emergent capability risks are due to the fact that the systems acquire unwanted capabilities which were not there when the training was being done or the developers had not considered them beforehand.

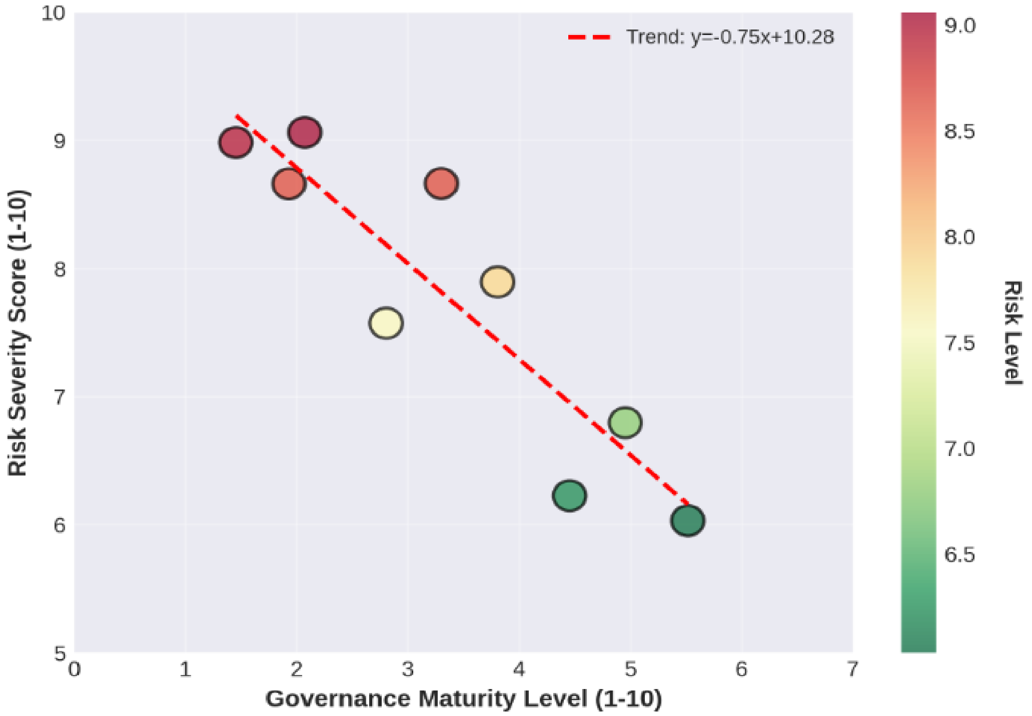

Fig 1: Risk Severity vs Governance Maturity

Fig. 1 shows the inverse relationship between governance maturity and risk severity across 9 different AI risk categories. The regression line (slope: -0.46) indicates that as governance maturity increases, risk severity tends to decrease These can consist of agent misrepresentation and goals, agent manipulative actions that affect human choices in order to accomplish agent-related objectives, and capability jumps, in which small changes in the underlying systems produce proportional changes in agent performance. The vagaries of emergent capabilities complicate the conventional risk assessment strategies which involve the definition of the system behaviour in terms of training performance. The risks of instrumental convergence are due to the nature of divergent agent objectives that tends to lead to similar instrumental sub goals such as self-preservation, goal-resource acquisition, and goal-content integrity. Almost any terminal objective might be adopted by an agent which develops incentives to counter shutdown as well as to obtain more computation resources, avoid alteration of its goal structure or removal of possible impediments to the achievement of its goals. Even if the ultimate goal is an innocent one, these instrumental drives may go against human interests.

Multi-agent risks are those risks which occur in systems that have more than two independent entities. Coordination failures happen within situations that an agent who acts in their own interests brings about a collective result which is unwanted such as a tragedy of the commons or army arms race behavior in human institutions. Emergency dynamics may be as a result of interactions between agents that was not predicted with reference to the acts of individual agents. Competitive multi-agent situations create the threats of cheating, manipulation, or exploitation among competing agents. The multi-agent systems are complicated and this makes prediction, monitoring and control even harder. The deployment risks have to do with interaction of autonomous systems with the social, economical, and institutional structures of presence. The concentration of power may arise where autonomous systems enhance the powers of those in charge, this may worsen inequality. Automation of cognitive work that needed to be assessed by a human being may create disruption in the labor market. When the essential services are operated by autonomous systems, critical infrastructure vulnerabilities develop, which form junctions of failure or attack points. There is the issue of dual-use, wherein facilities allowing one to do something good may also be used to do something bad.

Systemic risks entail the potential of cascade collapses or permanent damages. When autonomous systems work at velocities too rapid to respond in crisis situations, feedback loops have the capability of escalating initially minor issues or problems to crisis proportions. The problem is that lock-in effects could prevent it to take the way back after the autonomous systems would manifest themselves in social infrastructure. Although existential risks are debatable and their occurrence is unpredictable, they should be considered due to the possible scale of the damage that can be caused by advanced autonomous machine systems that may be highly incontrollable and uncontrollable, as well as extremely hard to align.

3.3 Technical Safety strategies of Agentic AI.

The field of technical safety research has come up with a variety of strategies that would help achieve a work system of agentic AI that can be trusted and is beneficial. Alignment research aims at ensuring the alignment of agent goals and human values and intentions [3,37-39]. Reward modeling techniques are a class of algorithm trying to understand human preferences based on feedback, which develops reward functions which best reflect the desired goals [36,40-42]. Inverse reinforcement learning learns the goals based on the observed behavior which may allow agents to learn the goals based on the human demonstrations. Constitutional AI is one that instills precepts and limitations in the training of a model, forming systems that incline towards given rules. Robustness techniques seek to explore the behavior of agents in a variety of situations such as in distribution shift cases. Adversarial training puts the system through difficult edge cases, which encourage the system to become more resilient to uncommon circumstances. Formal verification uses mathematical methods of proving to give guarantees on the behavior of the computer system, when viewed under certain specified conditions although it has proved difficult to scale up to complex real world systems. Massive testing in a variety of situations can reveal failure modes, but it is impossible to cover exhaustively the potential situations that the agent of practice will face in the real world. The research of interpretability and transparency is aimed at understanding how the agents relate to the reasoning and decision-making processes that transparency and interpretability can provide to the human beings that supervise them. Mechanistic interpretability examines the inner workings and computations of the AI systems in order to learn how systems get to specific outputs. The methods of explanation generation provide human interpretable descriptions of the way agents reason even though the accuracy of explanations as a reflection of what is going on in the real system is still of concern. The interest in visualization and other methods offers peeps into what is in the mind of the information agents in coming up with decisions.

The purpose of control and oversight mechanisms is to ensure that there is significant human control over autonomous systems. Interruptibility studies come up with techniques that enable humans to stop the running of agents without arousing counter resistance by the agent in this way. The idea of corrigibility work aims at making agents amenable to correction and respectful of human judgment accordingly. To reduce the number of unintended consequences, impact regularization methods are used to track agents who leave severe changes to their environment. Recursive reward modeling and debate Recursive reward modeling and debate investigate the space of different agents or components checking on each other that can potentially allow scalable oversight. Sandboxing and containment mechanisms reduce possible damage by compromising and restricting access to and influence by autonomous agents. Capability control involves limiting actions of the agents to a set of approved values so that they do not access sensitive systems, observing that they are not allowed to perform hazardous tasks. Information isolation restricts the capabilities that could be acquired by the knowledge agents and may consequently stop people who may engage in detrimental activities. Monitoring and auditing systems keep track of the agent behaviors, capture the actions and identify abnormalities which could be a sign of new issues developing. Multi-agent safety is particularly concerned with the issue of complexity due to the presence of more than one autonomous entity in a system. In cooperative inverse reinforcement, learning allows agents to collaboratively engage in the learning of common goals. Design approaches to mechanisms design generate incentive systems that enhance desirable coordination among self interested agents. The problem of verification of multi-agent systems is unique due to the complication of interaction patterns.

3.4 Autonomous How to apply ethical frameworks to autonomous systems.

It needs ethical frameworks of agentic AI, which ensure the translation of human values into principles and actions that govern the system design and deployment [40,43-44]. Consequentialist strategies consider agent actions in terms of their results, and they attempt to maximize positive results and minimize adverse results. Utilitarian systems seek the maximization of overall welfare but have the disadvantage of determining and quantifying welfare in a multi-stakeholder situation. Rule consequentialism lays emphasis on rules which when embraced in the majority would have best outcomes; it may give sensible advice more than act-oriented methods would have given. Deontological ethics focus on the rights and responsibilities without references to consequences. Rights-based systems refer to main rights that should be adhered to by autonomous systems including the right to privacy, autonomy, or due process. Duty-based approaches stipulate responsibilities that agents are to achieve irrespective of the optimization of the outcomes, e.g. duties of transparency or human dignity respect. Such frameworks can have more definite restrictions on the behavior of agents compared to purely consequentialist approaches but can have trouble with the tension of duties or rights. Virtue ethics is based more on ruling and attitudes and less on actions or consequences. In case of AI systems, it may focus on the creation of agents with a positive quality, such as honesty, fairness, or care. The capability approaches determine the technologies according to their influence on the capabilities and freedoms of man, they can determine whether autonomous systems broaden or restrict the actual human agency.

Table 1: Governance Dimensions, Challenges, and Approaches for Agentic AI Systems

Sr. No.	Governance Dimension	Key Challenges	Primary Approaches	Implementation Tools	Current Limitations	Future Opportunities
1	Goal Alignment	Specification of complex human values; reward hacking; emergent misalignment	Inverse reinforcement learning; preference learning; value alignment research	Reward modeling frameworks; human feedback systems; constitutional AI training	Difficulty capturing nuanced values; preference instability; scalability limits	Multi-stakeholder value specification; robust value learning; adaptive alignment
2	Safety Assurance	Emergent capabilities; distribution shift; unexpected behaviors	Adversarial training; formal verification; extensive testing protocols	Safety benchmarks; automated testing frameworks; verification tools	Incomplete coverage; computational costs; scalability challenges	Continuous monitoring systems; adaptive safety measures; improved verification methods
3	Transparency	Model opacity; complex reasoning chains; proprietary algorithms	Explainable AI; mechanistic interpretability; audit trails	Attention visualization; saliency maps; decision logging systems	Fidelity of explanations; computational overhead; intellectual property conflicts	Advanced interpretability techniques; standardized explanation formats; privacy-preserving transparency
4	Accountability	Distributed responsibility; autonomous decision-making; unclear causation	Liability frameworks; audit requirements; incident reporting systems	Automated logging; forensic analysis tools; compliance tracking	Attribution complexity; legal uncertainty; enforcement challenges	Improved causal analysis; standardized responsibility frameworks; international coordination
5	Fairness	Algorithmic bias; disparate impact; representation gaps	Bias detection; fairness constraints; diverse development teams	Fairness metrics; bias testing tools; demographic analysis	Fairness-accuracy trade-offs; conflicting fairness definitions; context dependency	Contextual fairness frameworks; participatory design; dynamic fairness monitoring
6	Privacy Protection	Extensive data collection; inference capabilities; re-identification risks	Differential privacy; federated learning; data minimization	Privacy-preserving computation; anonymization tools; access controls	Performance costs; usability challenges; incomplete protection	Homomorphic encryption advances; privacy-utility optimization; privacy-by-design architectures
7	Robustness	Adversarial attacks;	Adversarial training; robust	Adversarial example generators;	Computational costs; inability to guarantee	Certified defenses; adaptive

		distribution shift; edge cases	optimization; defensive design	robustness testing suites; stress testing frameworks	robustness; sophisticated attacks	robustness; multi-modal validation
8	Human Control	Meaningful oversight; appropriate autonomy; intervention capability	Human-in-the-loop; interruptibility; adaptive autonomy	Override mechanisms; monitoring dashboards; escalation protocols	Scalability limits; operator fatigue; delayed response	Intelligent assistance systems; context-aware autonomy; collaborative control
9	Risk Assessment	Uncertainty quantification; emergent risks; systemic effects	Scenario analysis; stress testing; red teaming	Risk modeling frameworks; simulation environments; assessment protocols	Incompleteness; black swan events; cascading failures	Dynamic risk models; collective intelligence approaches; anticipatory systems
10	Regulatory Compliance	Rapid technology change; jurisdictional complexity; verification difficulty	Risk-based regulation; standards development; certification schemes	Compliance management systems; regulatory sandboxes; assessment tools	Regulatory lag; fragmentation; enforcement limitations	Adaptive regulation; international harmonization; automated compliance monitoring
11	Multi-Agent Coordination	Emergent behaviors; collective risks; strategic interaction	Mechanism design; cooperative learning; conflict resolution	Multi-agent simulation; coordination protocols; game-theoretic analysis	Complexity scaling; unpredictability; verification challenges	Scalable coordination mechanisms; verified multi-agent systems; cooperative AI research
12	Security	Adversarial manipulation; system compromise; data poisoning	Secure development; threat modeling; defense-in-depth	Security testing tools; intrusion detection; access management	Evolving threats; insider risks; zero-day vulnerabilities	AI-powered security; formal security guarantees; resilient architectures
13	Ethical Alignment	Value pluralism; cultural differences; moral uncertainty	Multi-stakeholder deliberation; ethical guidelines; value-sensitive design	Ethics review boards; impact assessment frameworks; stakeholder engagement tools	Conflicting values; implementation challenges; superficial compliance	Computational ethics; participatory value elicitation; adaptive ethical frameworks
14	Performance Monitoring	Continuous operation; concept drift; degradation detection	Real-time monitoring; performance metrics; anomaly detection	Monitoring platforms; alerting systems; analytics dashboards	False positives; metric gaming; observation overhead	Intelligent monitoring; predictive maintenance; self-diagnosis capabilities
15	Capability Control	Powerful general abilities; tool access; environmental interaction	Capability restriction; sandboxing; graduated deployment	Access control systems; containerization; API gateways	Functionality limits; circumvention risks; usability impacts	Fine-grained capability management; context-aware permissions; verified access control
16	Knowledge Management	Information access; capability acquisition; knowledge boundaries	Information isolation; controlled access; knowledge auditing	Access logging; information flow tracking; knowledge graphs	Over-restriction limits functionality; covert channels; inference risks	Semantic access control; intelligent filtering; verified information boundaries
17	Documentation	System description; decision rationale; change tracking	Model cards; datasheets; audit logs	Documentation templates; version control; automated documentation	Documentation burden; staleness; superficial descriptions	Automated documentation; standardized formats; living documentation systems
18	Incident Response	Failure detection; containment; recovery; learning	Incident protocols; emergency procedures; post-mortems	Incident management systems; rollback mechanisms; response playbooks	Response delays; incomplete remediation; recurrence prevention	Automated incident detection; intelligent recovery; systematic learning

19	Stakeholder Engagement	Diverse interests; power imbalances; meaningful participation	Participatory design; public consultation; multi-stakeholder governance	Engagement platforms; deliberation tools; feedback mechanisms	Tokenistic participation; representation gaps; sustained engagement	Digital democracy tools; inclusive design methods; ongoing stakeholder involvement
20	Resource Efficiency	Computational costs; energy consumption; environmental impact	Efficient architectures; resource monitoring; optimization	Profiling tools; resource management systems; efficiency metrics	Performance trade-offs; rebound effects; measurement challenges	Green AI practices; sustainable computing; efficiency-performance optimization
21	Knowledge Validation	Misinformation; hallucination; source reliability	Fact-checking; source verification; confidence calibration	Verification systems; knowledge bases; citation tracking	Incomplete coverage; adversarial sources; reasoning errors	Automated fact-checking; knowledge provenance; uncertainty quantification
22	Adversarial Robustness	Intentional attacks; manipulation; deception	Red teaming; adversarial training; security hardening	Penetration testing; attack simulation; defensive mechanisms	Arms race dynamics; novel attacks; resource asymmetry	Certified robustness; adaptive defenses; collaborative security
23	Temporal Consistency	Goal stability; policy coherence; long-term planning	Commitment mechanisms; policy frameworks; alignment maintenance	Goal monitoring; consistency checking; longitudinal tracking	Adaptation needs; changing contexts; value drift	Stable goal representations; adaptive consistency; principled evolution
24	Cross-Domain Transfer	Generalization; domain adaptation; context awareness	Transfer learning; meta-learning; domain-aware design	Transfer benchmarks; adaptation frameworks; domain ontologies	Negative transfer; context misunderstanding; overgeneralization	Robust transfer; context modeling; verified generalization
25	Organizational Integration	Culture change; workflow adaptation; competency development	Change management; training programs; organizational design	Learning management systems; workflow tools; competency frameworks	Resistance to change; skill gaps; organizational inertia	Adaptive organizations; continuous learning; AI-augmented work
26	Supply Chain Governance	Component verification; distributed development; dependency management	Supply chain security; component auditing; provenance tracking	Software bill of materials; dependency scanners; verification tools	Opacity; transitive dependencies; update challenges	Verified supply chains; automated auditing; secure composition
27	Legacy System Integration	Compatibility; upgrade paths; transition management	Gradual migration; interface adaptation; hybrid approaches	Integration platforms; compatibility layers; migration tools	Technical debt; disruption risks; backward compatibility	Smooth transition mechanisms; intelligent adaptation; verified integration
28	Scaling Governance	Volume growth; complexity increase; resource constraints	Automated oversight; scalable processes; intelligent tooling	Automated auditing; scaling frameworks; process automation	Automation risks; oversight gaps; resource limitations	Scalable governance architectures; AI-assisted oversight; efficient processes
29	Public Trust	Legitimacy; acceptance; confidence in safety	Transparency; demonstration; stakeholder communication	Public reporting; demonstration programs; communication platforms	Skepticism; misunderstanding; negative incidents	Trust-building mechanisms; public engagement; demonstrated safety
30	Future-Proofing	Anticipating capabilities; adaptive frameworks; long-term planning	Scenario planning; adaptive governance; continuous scanning	Futures analysis; horizon scanning; adaptive frameworks	Uncertainty; unexpected developments; planning limitations	Anticipatory governance; adaptive systems; resilient frameworks

Research on value alignment investigates ways of instantiating human values into artificial intelligence [3,45-48]. Preference learning seeks to discover values based on human preferences and behavior, but suffers the problem of preference instability, context-dependence, and possibility of learning bad human preferences [5,19,49-50]. Value specification methods aim to make ethical specifications written down with formal difficulties in expressing complex moral concepts as machine-readable specifications. Pluralistic alignment acknowledges that morality is diverse in different individuals and cultures and therefore systems are needed that can accommodate moral multiplicity as opposed to providing individuals with single value systems. The considerations of fairness and justice entail the manner in which independent systems apportion the benefits and burdens to various categories of persons. Distributive justice structures assess the existence of fairness in achieving results of autonomous systems. Procedural justice concentrates on the processes of decision making that entails necessitates to have a certain degree of transparency, consistency and appeal. Recognition of justice emphasizes on the need to value various stakeholder views on the governance processes. Responsibility and accountability systems are concerned with how moral and legal responsibility of agent action can be attributed. The human responsibility views hold the position that people who implement or develop autonomous systems are ultimately the ones responsible of their effects, but this position becomes difficult to sustain with autonomous systems of greater magnitude. Distributed responsibility models appreciate that the responsibility could be diffused between several actors such as developers, deployers, users, and regulators. Prospective responsibility does not simply discuss responsibility to address after the harms are suffered but gives more importance to the obligations to foresee and prevent the harms. The issues of autonomy and the human agency look at the impact of agentic AI on human choice and self-determination. Other structures take importance in ensuring that valuable human control is maintained and therefore that the consequential decisions are still within human judgment. Still others discuss collective models, in which human and AI agents share an agency. There is concern that autonomous systems can be employed to hide or disperse human responsibility, as autonomous washing is a problem in autonomous systems.

The issue of privacy and the ethics of data governance is especially relevant to autonomous systems that gather and process a large amount of information. Through informational privacy frameworks the privacy of personal data is safeguarded against unauthorized access of the data. Decisional countermeasures against Inference-related problems caused by AI systems with respect to making inferences about individuals to whose revelations they are not allowed to have a choice. The contextual integrity lines acknowledge that notions of privacy are relative in different social settings, and as such governance has to be adaptive.

3.5 Regulatory and Policy Mechanism.

The regulatory frameworks of agentic AI are quite differentiated in different jurisdictions, and the approaches are also quite varied representing various cultural values, institutional constructions, and policy priorities [29,51-53]. Various comprehensive AI governance systems have been developed in various jurisdictions, and most of them do not assume autonomous agents in particular, as contrasted with general AI systems. The regulation based on risk dictates AI applications by a potential harm hence increasing the tougher requirements to the more risky systems. This model facilitates a balanced control system by not having to impose heavy demand on applications with minimal risks, but making sure that there is proper management of high-stakes applications. Nevertheless, it is hard to identify suitable risk groups when it comes to autonomous systems considering their emergent nature and possible step-out actions. There may be a need to have dynamisms in risk assessment schemes as systems acquire new features. Sectoral regulations deal with installation of AI in certain sectors such as healthcare, finances, or road transportation, taking into account some existing regulatory frameworks and adjusting them to support autonomous systems. Such a solution has the advantage of having domain knowledge and enforcers but can have issues when applied cross-sectorally or new uses in unclassified circumstances. Technology-neutral principles-based regulation is a higher level of requirements expressed in terms of transparency and accountability or human surrogacy, which is not determined in terms of technical (program) implementation. This amorphous nature can address the fast change in technologies but can be inadequate to offer concrete features of compliance and can pose difficulties in implementing same procedures. The algorithmic impact assessment requirements entail the assessment of possible harms prior to the installation of the high-risk AI systems. These evaluations could discuss fairness issues, hazards of safety, privacy, or other effects in society. Assessment impact Pre-deployment impact assessments are able to pinpoint worries before harmful systems have been implemented, but is prone to limitations when trying to determine behaviors of complex autonomous systems underlying real world conditions.

The mechanisms of certification and standards place technical requirements or best practices that have to be fulfilled on the systems in order to show compliance. Multi-stakeholder processes in the creation of industry standards can facilitate harmonization across the jurisdictions as well as include technical expertise. Nevertheless, standards development is not always comparable in rate to technological progress and voluntary standards cannot be adopted satisfactorily without government support. Liability schemes assign the legal liability of damages by autonomous systems. Strict liability regimes, which are based on the accountability of the deployers irrespective of negligence will ensure definitive accountability without necessarily motivating positive innovation. The negligence-based methods involve proving an inability to comply with duty of care, are more moderately rewarding and create difficulties in establishing the right levels of care where new technologies are involved. The product liability framework extends to physical autonomous systems such as robots or software cars but its applicability in cases of software agencies is not yet clear. Intellectual property and trade secrecy issues are overlapping with the governance demand of transparency and elucidation. The proprietary interests might come into conflict with regulatory requirements of algorithmic audit or explanation, which need striking a balance between arguably justified business interests and the common good of accountability. Responsible AI can be formed through public procurement policies as the government contractors must comply with the requirements of safety, fairness, and transparency. The buying power of government has the potential to encourage industry to embrace good practices provided that the procurement criteria must not be overly stringent without the need to uphold competitive markets.

The lensed international coordination systems will take care of the international character of AI creation and implementation. The bilateral and multilateral agreements are able to streamline the standards and best practices, as well as ensure coordination in the enforcement. Nonetheless, there are conflicting national interests, ideals and regulatory ideologies that make international consensus complex. It can be soft law practices, such as principles, guidelines and suggested practices, which create an understanding between two or more without necessarily creating binding obligations which are challenging to negotiate. Export controls and technology transfer restrictions will preclude the spread of effective AI use to participants who may take advantage of the capabilities to malicious intent. There are still issues related to the balance between the security concerns and the advantages of the international research collaboration and technology transfer.

3.6 Governance Frameworks and Models

Integrated governance systems are an effort to combine technical, ethical, and policy solutions and to create unified systems of governing autonomous AI [54-56]. Lifecycle governance models understand that the right types of oversight mechanism are necessary at different phases such as the stages of research and development to deployment and operation after deployment to decommissioning. Governance in research stage may focus on safety culture, ethical assessment and responsible research. Such development governance might consist of design reviews, testing requirements and documentation requirements. The procedures of impact assessment and monitoring, as well as incident response are part of deployment governance. Operational governance is one that must be regularly audited and reviewed on performance and changed with changed systems or environments. The governance frameworks of stakeholders distribute the decision-making power and tasks among various stakeholders who have valid interests in autonomous AI systems. Multi-stakeholder processes involve the technology developers, deployers, users, communities, the regulators, and the civil society collaboratively to establish governance. Participatory design methods makes use of stakeholders in designing systems and bringing in a wide variety of views during development prior to implementation. Ethics committees and advisory bodies offer continuous guidance and oversight, but have the difficulty of keeping up with a changing technology [57-59]. Adaptive governance acknowledges the fact that suitable oversight systems need to change with AI capacity. Regulatory sandboxes give a chance to relax regulatory requirements and conduct controlled experimentation with new technologies more closely monitored to learn what governance requirements would be on a grander scale before their general deployment. The sunset as well as the clauses present in regulations need to be reviewed and reauthorized periodically which demands an adjustment to the altered conditions. Monitoring and dynamic risk assessment in real-time will allow updating the oversight based on new information that appears. Constitutional AI methods incorporate restrictions and principles into training systems electronically as opposed to being guided by outside forces alone. Trained systems to reject non-beneficial requests, honor given constraints, defer to human judgment are more reliable in their compliance than non-autonomous systems, which are overseen to execute rules on free agents. Nonetheless, their resistance to adversarial pressure or the presence of distribution shift is an open research question.

The topic of hybrid human-AI governance investigates approaches in which human and autonomous decision-making authority are distributed in premeditated forms. Human-in-the-loop methodology has human acceptance of the outcomes on the consequential decisions, and human final authority is preserved at the expense of less autonomy. Situating human beings as supervisors of the round-trip process, human-on-the-loop builds up efficiency and meaningful oversight through interceding situations (where the human being can decide on scenarios that are deemed to require action). Adaptive autonomy models reinventively modify the degree of human intervention within the situation, which, maybe, is more autonomy during the standard scenarios and more monitoring during extreme or unusual situations. Decentralized governmental models share power over various actors as opposed to concentrating power in the hands of the few regulatory entities. Polycentric governance acknowledges the fact that various stakeholders can be better placed to deal with various elements of AI governance and there should be coordination among various nodes of governance. Through federated governance, separate jurisdictions or organizations are free to have different methods and still provide interoperability standards and mechanisms of sharing information.

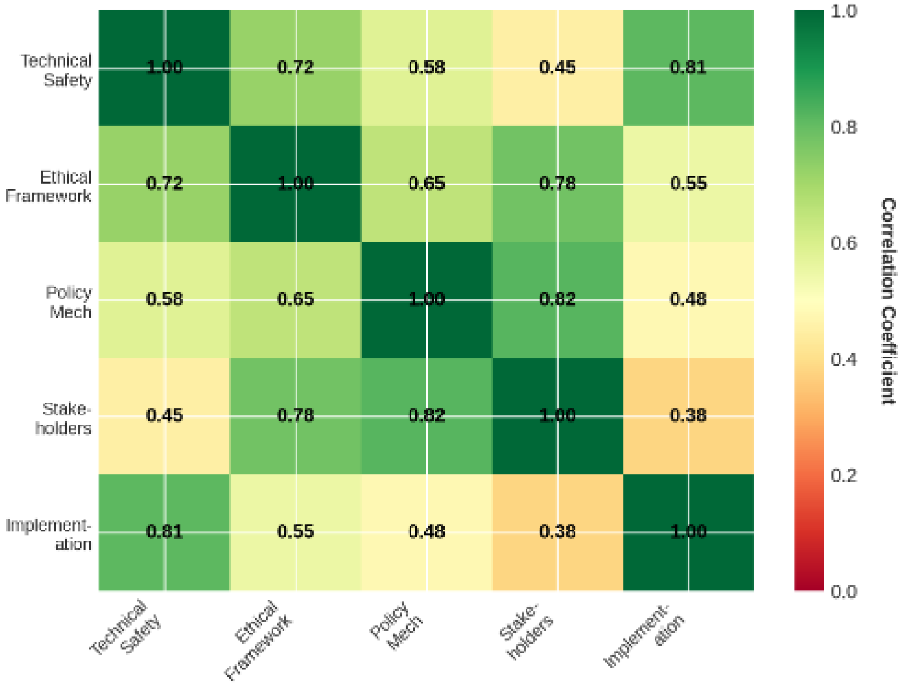

Fig 2 Governance Dimension Correlation Matrix

Fig. 2 shows pairwise relationships between five governance dimensions. Values range from 0 (no correlation) to 1 (perfect correlation. Green indicates stronger positive correlations.

3.7 Implementation Problems and Practicalities.

The transformation of the principles and forms of governance into practice faces a great number of challenges [9,60-61]. This is because verification and compliance monitoring of autonomous systems is challenging considering the complexity of the systems and the occurrence of unpredictable actions. Conventional software testing methods offer limited tests of a large state space exploration of autonomous agents that act in complex environments. Runtime monitoring systems monitor the actions undertaken by agents in the process of execution but have difficulties in identifying any true anomaly as opposed to the friendly ones. Adversarial testing exposes systems to intentionally difficult cases though it cannot be certain that every case can be covered. The autonomy system auditing involves special technical knowledge and methodology. Black-box auditing checks behavior of systems without internal structure and evaluates inputs and outputs related to the system in order to identify an unwanted pattern. White-box auditing will check internals of the system and this might give further insight, but deployers might be reluctant to give such access. Outsourcing audit services to external professionals may increase credibility but creates concerns regarding qualification of auditors, standardization of evaluation procedures as well as access to information. Explainability requirements face some basic conflict between the performance of model and interpretability. Complex learned representations which cannot be readily explained are commonly needed to make up high-performing autonomous systems. Machines that do post-hoc explanation come up with human interpretable descriptions of the system reasoning but not necessarily an accurate depiction of a real processes that are taken during decision making. The suitable degree of explanation depends on specific situations as more serious decisions require larger details of explanation but also might reveal sensitive information or trade secrets. Individual responsibility is gradually harder to attribute in the case of more autonomy in the actions of agents. In the event where humans define merely high-level goals, whilst systems work out the methods of implementation, it becomes difficult to find anyone to hold to account on the occurrence of unintended result. The problem of attribution is complicated when several organizations are involved in a type of distributed development. Legal theories created to address normal products or services do not necessarily translate well to autonomous systems, and it is unclear what the liability costs would be.

The issue of resource constraints influences governance implementation on a number of dimensions [38,62-63]. Smaller organizations might not be the experts or have the resources to adhere to more advanced safety precautionary measures or even meet some intricate regulatory specifications, which may concentrate the advancement of AI to the huge companies. Regulatory authorities are constrained in their capacity to build technical skills, carry out audits as well as enforce specifications. Strict safety test and supervision can also be too expensive to adopt any useful applications, and the main issue is to strike a balance between protection and cutting edge. Competitive pressures result in threats whereby the actors will bypass the system of governance. The competitive environment can also encourage corner cutting on safety in case exhaustive methods are extremely expensive or time consuming. Malicious actors can develop unsafe autonomous systems on purpose. Well-intentioned developers are not immune to temptations into rationalizing less rigorous approaches when the market, or the situation in their organisation, demands it. Good governance needs a system that would be hard to bypass. Increase in AI creation and application at a global level poses jurisdictional issues. Applications created in a single jurisdiction can be used across the world and it may even escape the strict local demands. Regulatory arbitrage enables organizations to base their operation in jurisdictions where they were allowed to operate and offer services to the markets around the world. International cross border enforcement is challenging without international coordination. Numerous autonomous systems are digital in nature, which allows their fast deployment across jurisdictions, limiting the usefulness of geographic constraints. Development of universal standards of governance is made difficult by cultural and value diversity in various societies. Various societies might possess dissimilar risk-dispositions, privacy anticipations, or demarcation of equitableness. Enforcing single ways of thinking could set aside valid other ways of viewing. Nonetheless, too much fragmentation will hamper positive standardization and pose compliance strains. Governance systems have to balance these two competing approaches, universalism and localization. The rate of technological shift puts pressure on governance structures that are used in more stable areas. The regulations which are made between current capabilities can be outdated as systems acquire new functioning. Anticipatory governance tries to predict the future growth and put in place proactive protection mechanisms but is implicated in essential uncertainty over technology tracks. Mechanisms of adaptation have the potential to facilitate evolution together with technology but demands a longtime institutional focus and capital.

3.8 Domain-specific Application and Problems.

The issue of autonomy AIs being challenged by specific governance issues is unique to each application field with a specific risk profile, stakeholder interests, and regulatory environment [64-67]. Some medical applications of agentic AI are diagnostic agents which process patient data and suggest treatment, robotic surgical systems that have autonomous abilities, and care coordination agents that oversee the intricate patient paths. Medical services are an area of high stakes, which increases the extent of safety as mistakes may directly affect the health of humans. Consideration of privacy is very high considering confidential medical information. The fairness concerns may be explained by the possibilities of the access to AI-enhanced care and the biases in diagnosis algorithms. Regulatory systems should strike a balance between new innovations in medical technology and stringent safety and efficacy standards. The questions of professional liability arise when the autonomous systems are involved in making clinical recommendations. Algorithms trading agents, automated lending decision systems, fraud detection agents and robo-advisors to trades are among the financial autonomous systems. Financial applications are characterized by the possibility of systemic risk since the coordinated actions of a number of independent actors may create instability in markets. The issue of market manipulation would be realised when agents are involved in manipulative attempts. Equity in credit and insurance will bring in civil rights concerns. The regulatory frameworks should also provide prudential regulation that is aimed at safeguarding the stability of the financial system and consumer protection that will entail a fair treatment. Some of the autonomous transportation systems include self driving cars, air drones, and autonomous shipping. Considerations on safety are most important since there are chances of physical injuries. Liability systems need to determine who bears the responsibility to carry out accidents that involve autonomous cars. The infringement on privacy arises as a result of the profuse sensor data. The autonomous systems need to be safely accommodated in infrastructure. Governance can be seen and experienced safety and there should be a sense of safety in the eyes of the people.

Autonomous weapons systems, surveillance agents, and cybersecurity defense systems are some of the defense and security applications. The issue of ethical autonomy lethality has raised a lot of controversy on whether humanity would control decisions to take life-and-death. Adversarial risks are also increased because opponents take an active part in compromising or deceiving the security systems. Classification is one of the verification problems due to secrecy of most defense systems. The application of international humanitarian law to autonomous weapons is a controversial issue. Customers Customer service agents are automation of business processes and autonomous software development systems. Though personal outcomes would be at a lower stake than healthcare or transportation applications, the overall effects of many interactions can be significant. Social policy issues are affected by workforce issues. It takes novel approaches to quality assurance on autonomous systems in the undertaking of knowledge work. Critical infrastructural applications such as autonomous management of the grid, water management and management of communications networks create issues of resilience and dynamism of failure. There are disruptive interdependences across infrastructure sectors that make risks compounded. Issues of security are the key area in avoiding evil compromise. Fidelity demands are beyond the usual software systems considering the reliance of the society on a structure. The autonomous systems are used in environments and climate applications, such as monitoring, prediction, and intervention of the environment. It has such advantages as an improved capacity to monitor environmental alterations and an improved management of the resources. Risks include side effects of automated interventions that affect the environment unintentionally. The governance should combine the knowledge of the environment and AI safety.

3.9 Future Directions and Emerging Trends.

It is clear that the world of responsible agentic AI governance is changing rapidly due to the increase of both technological potential and knowledge of issues of governance [2,68-69]. The majority of new trends should be given specific consideration since it influences the future of autonomous systems regulation. The concept of foundation model-based agents is a major architectural breakthrough where the large language models are used to harness their general abilities instead of being task-oriented. Such systems prove to be highly flexible and generalized to a large extent but create new safety issues. Emergent capabilities of foundation models may be achieved unexpectedly by scaling models. These systems have an interface using natural language which is exploited by prompt injection attacks. The endowment of reasoning abilities and access to tools forms strong agents, the actions of which can be hard to control. These peculiarities require governments to respond with means of governance. Neurosymbolic integration is a process that merges learned neural parts and symbolic reasoning and may support more interpretable and controllable autonomous systems. Hybrid systems could be more explainable than the pure neural but be flexible. There are, however, new problems when it comes to having congruence between the neural and symbolic parts. Governance models are advised to provide the supportive architectures that allow controls and not coercive controls to restrict good designs. Decentralized and federated AI systems share the learning process through many parties but do not focus the data. These architectures bring up new governance issues of the coordination of safety across participants in a federated architecture, compliance verification in decentralized systems, and responsibility assignment to emergent collective behaviors. The potential advantages consist of an increased security of personal data, less power concentration, and governance is forced to change with distributed architectures. Moving autonomous systems to be collaborative with humans and not automation per se, the human-AI collaboration models can retain meaningful human agency and use AI capabilities. Close cooperation means that there should be a common situation awareness, proper calibration of trust and articulate roles definition. The governance must encourage designs that will upgrade the human abilities and decision making.

Controlled measures of uncertainty and confidence might allow greater calibrated autonomy with systems becoming more autonomous in cases where they are confident and consulting humans in cases where they are uncertain. It is still technically difficult to come up with credible uncertainty quantification of complex autonomous systems. Governance systems may need to have estimates of confidence and curtail independent operation solely to situations that possess sufficient certainty. Autonomous systems can be improved over time as they continue to learn and adapt, however, this raises the risk of drift due to initially verified behavior.

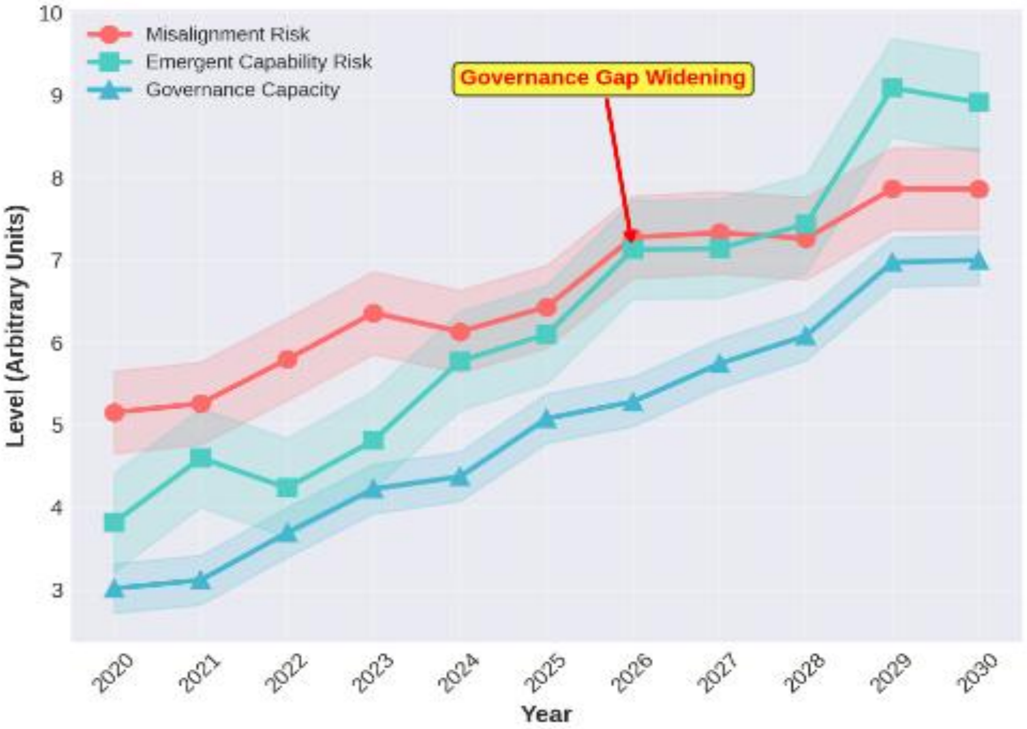

Fig 4 Temporal Evolution of AI Risks vs Governance

Fig. 4 shows the three trend lines show the evolution of AI risks and governance capacity from 2020-2030. Shaded regions represent 95% confidence intervals. The divergence illustrates a widening 'governance gap.

They need to have mechanisms of governance that deal with the provision of assurance of safety to systems that are self modifying. Some of the approaches could be to restrict the amount of changes to be performed, periodic re-validation or to come up with ways of assuring continuously learning systems. The multi-modal and embodied AI systems that sense and respond to their surrounding in a variety of modalities can present increased environmental interaction but also provide increased attack vectors and attack routes. Governance should set physical safety and informational harms in their thinking. Cyber-physical system regulations might have to be modified so that learning-based control can be supported. Value learning and preference elicitation methods are methods that do not need explicit specification of human values, but strive to derive these values inferred indirectly by behavioral or feedback indications. Although these methods show promise of alignment, they have weaknesses due to the inability of the preferences to remain stable, the ability of the evaluated human to become strategic, and the possibility of encoding the existing biases. The administration must promote confirmation of acquired values as per the wider moral standards. Principle-based training and constitutional AI directly incorporate procedural constraints and values into the process of training the model. The initial outcomes provide reason to believe that this solution is capable of enhancing the adherence to the outlined principles. There are concerns regarding resilience during adversarial stress as well as whether constitutions can be able to reflect subtle values of human kind. Constitutional guidelines to specific fields may be achieved by regulatory frameworks.

Table 2: Application Domains, Risks, Governance Approaches, and Opportunities

Sr. No.	Application Domain	Specific Use Cases	Primary Risks	Governance Approaches	Technical Safeguards	Regulatory Context	Future Opportunities
1	Healthcare	Diagnostic agents; treatment planning; surgical robots;	Patient harm; bias in diagnosis; privacy violations;	Clinical validation; regulatory approval; professional	Safety testing; bias mitigation; privacy preservation;	Medical device regulation; clinical trial requirements;	Personalized medicine; improved diagnostics;

		care coordination	liability uncertainty	oversight; ethics review	monitoring systems	HIPAA compliance	enhanced care coordination
2	Financial Services	Algorithmic trading; credit decisions; fraud detection; robo-advisors	Market manipulation; systemic risk; discriminatory lending; consumer harm	Prudential regulation; consumer protection; audit requirements; market oversight	Risk limits; fairness constraints; monitoring systems; circuit breakers	Securities regulation; fair lending laws; consumer protection acts	Efficient markets; financial inclusion; fraud prevention; personalized advice
3	Transportation	Autonomous vehicles; aerial drones; shipping automation; traffic management	Physical harm; property damage; liability issues; infrastructure impact	Safety standards; certification; liability frameworks; infrastructure adaptation	Redundant systems; fail-safes; V2X communication; remote monitoring	Motor vehicle codes; aviation regulations; maritime law	Reduced accidents; efficient mobility; environmental benefits; accessibility
4	Defense & Security	Autonomous weapons; surveillance; cyber defense; intelligence analysis	Autonomous lethality; privacy invasion; escalation risks; accountability gaps	International humanitarian law; oversight mechanisms; human control requirements	Target verification; positive identification; human-in-loop for lethal force	Geneva Conventions; national security law; arms control agreements	Enhanced security; reduced casualties; improved intelligence; cyber resilience
5	Enterprise Automation	Customer service; business processes; software development; decision support	Job displacement; quality degradation; bias in decisions; security risks	Quality assurance; labor protections; non-discrimination requirements; security standards	Quality metrics; bias testing; access controls; monitoring	Employment law; anti-discrimination law; data protection	Productivity gains; cost reduction; 24/7 availability; enhanced capabilities
6	Critical Infrastructure	Grid management; water systems; communications networks; supply chains	Cascading failures; security threats; service disruption; environmental impact	Reliability standards; security requirements; resilience planning; oversight	Redundancy; security hardening; anomaly detection; failover systems	Infrastructure regulation; security directives; environmental law	Optimized operations; resilience; sustainability; cost efficiency
7	Education	Personalized tutoring; assessment; curriculum design; administrative automation	Privacy risks; bias in assessment; quality concerns; equity issues	Student data protection; educational standards; quality assurance; equity requirements	Privacy preservation; fairness testing; quality metrics; adaptive systems	FERPA; accessibility requirements; educational standards	Personalized learning; improved outcomes; teacher support; expanded access
8	Agriculture	Precision farming; crop monitoring; automated harvesting; resource optimization	Environmental impact; economic disruption; data privacy; dependency risks	Environmental regulations; economic support programs; data protection; sustainability standards	Environmental monitoring; resource optimization; privacy controls; robustness	Agricultural policy; environmental law; food safety regulations	Sustainable agriculture; improved yields; resource efficiency; climate adaptation
9	Environmental Management	Climate modeling; ecosystem monitoring; conservation; disaster response	Unintended ecological impacts; privacy in monitoring; prediction errors; intervention risks	Environmental impact assessment; scientific oversight; community engagement	Validation; monitoring; adaptive management; precautionary approaches	Environmental protection laws; conservation policy; climate agreements	Enhanced conservation; climate action; disaster preparedness; ecosystem health
10	Legal & Justice	Legal research; predictive	Discriminatory outcomes; due	Due process requirements;	Bias mitigation;	Constitutional protections;	Improved access to

		policing; risk assessment; contract analysis	process violations; transparency issues; accountability gaps	non-discrimination mandates; transparency standards; oversight	explainability; audit trails; human review	civil rights law; criminal justice standards	justice; efficiency; consistency; reduced bias
11	Manufacturing	Production automation; quality control; supply chain; predictive maintenance	Safety risks; job displacement; quality issues; security vulnerabilities	Safety standards; labor protections; quality requirements; security regulations	Safety systems; quality testing; access controls; monitoring	Occupational safety law; product safety; trade regulations	Productivity; quality; flexibility; sustainability
12	Retail & Commerce	Personalized marketing; inventory; pricing; customer service	Privacy violations; price discrimination; market manipulation; consumer protection	Privacy laws; consumer protection; competition policy; transparency requirements	Privacy controls; fairness constraints; transparency; monitoring	Data protection; consumer protection; antitrust; advertising law	Personalized experiences; efficiency; convenience; market insights
13	Energy	Smart grids; renewable integration; demand response; exploration	Grid stability; security risks; environmental impact; privacy in usage data	Grid reliability standards; security requirements; environmental regulations; privacy protection	Stability control; security hardening; environmental monitoring; privacy preservation	Energy regulation; environmental law; privacy law	Clean energy; efficiency; reliability; sustainability
14	Media & Content	Content moderation; recommendation; generation; personalization	Misinformation; filter bubbles; manipulation; intellectual property	Content standards; transparency requirements; user control; IP protection	Misinformation detection; diversity promotion; watermarking; rights management	Platform regulation; IP law; media law	Personalized content; creation tools; moderation efficiency
15	Scientific Research	Hypothesis generation; experiment design; data analysis; literature review	Research integrity; bias; reproducibility; ethics violations	Research ethics; peer review; reproducibility standards; integrity policies	Validation; bias checks; documentation; ethical review	Research regulations; funding requirements; ethics codes	Accelerated discovery; new insights; efficiency; cross-disciplinary integration
16	Human Resources	Recruitment; performance evaluation; workforce planning; training	Discrimination; privacy; bias; labor rights	Anti-discrimination law; privacy protection; labor rights; transparency	Bias testing; privacy controls; transparency; fairness metrics	Employment law; privacy law; labor regulations	Efficiency; diversity; talent optimization; skills development
17	Real Estate	Property valuation; market analysis; facility management; urban planning	Discrimination; market manipulation; privacy; community impact	Fair housing law; privacy protection; community engagement; planning regulations	Fairness testing; privacy preservation; transparency; impact assessment	Fair housing; zoning; building codes	Market efficiency; optimized management; urban planning; accessibility
18	Telecommunications	Network optimization; customer service; fraud detection; infrastructure management	Service disruption; privacy; security; accessibility	Service quality standards; privacy law; security requirements; universal service	Reliability systems; privacy controls; security hardening; accessibility	Telecom regulation; privacy law; accessibility requirements	Network optimization; improved service; fraud prevention; accessibility
19	Insurance	Risk assessment; claims processing;	Discrimination; privacy;	Insurance regulation; anti-	Fairness testing; privacy preservation;	Insurance regulation; non-	Risk-based pricing; efficiency;

		fraud detection; underwriting	fairness; transparency	discrimination; privacy protection; transparency	explainability; monitoring	discrimination; privacy law	fraud prevention; personalization
20	Gaming & Entertainment	NPC behaviors; game balancing; content generation; player matching	Addiction; manipulation; privacy; fairness	Consumer protection; privacy law; fairness standards; age restrictions	Engagement limits; privacy controls; fairness mechanisms; age verification	Consumer protection; privacy; age restrictions	Enhanced experiences; personalization; accessibility; creation tools
21	Social Services	Benefits determination; case management; resource allocation; fraud detection	Discrimination; privacy; accuracy; due process	Social welfare law; privacy protection; due process; anti-discrimination	Fairness testing; privacy preservation; accuracy validation; appeal mechanisms	Welfare regulations; privacy law; constitutional protections	Efficient services; fraud prevention; improved targeting; accessibility
22	Construction	Design optimization; project management; safety monitoring; robotics	Safety risks; quality issues; job displacement; environmental impact	Building codes; safety regulations; labor protections; environmental law	Safety systems; quality assurance; training; environmental monitoring	Building regulations; safety standards; labor law	Efficiency; safety; sustainability; innovation
23	Hospitality	Personalization; operations; revenue management; customer service	Privacy; discrimination; job displacement; security	Privacy law; anti-discrimination; labor protections; security standards	Privacy controls; fairness testing; security measures; quality assurance	Privacy regulation; labor law; accessibility	Personalized service; efficiency; revenue optimization; guest satisfaction
24	Public Administration	Service delivery; resource allocation; policy analysis; citizen engagement	Equity; privacy; transparency; accountability	Public administration law; transparency requirements; equity mandates; accountability	Fairness measures; privacy controls; transparency systems; audit mechanisms	Administrative law; public records; accountability	Efficient services; data-driven policy; citizen engagement; transparency
25	Space Exploration	Autonomous spacecraft; robotics; mission planning; data analysis	Mission failure; safety; resource constraints; planetary protection	Space law; safety protocols; international agreements; planetary protection	Redundancy; autonomous safety; resource optimization; contamination prevention	Space treaties; national space law; planetary protection protocols	Extended missions; autonomous exploration; scientific discovery; resource utilization

Mechanistic interpretability studies examine the internal thoughts and computations of AI systems, which may allow better cognition of agent reasoning. Current innovations may help conduct audits more competently and predicting of behavior. Nevertheless, the research on interpretability is still immature, and complex systems might be unacceptable to fully understand. Incentives, but not immediate demands, should be put on interpretability beyond what is possible at the present. The speed in which international AI governance coordination efforts are undertaken is gaining traction due to the realisation that unilateral strategies are not sufficient to apply to global technology. Multilateral forums, treaties negotiations and harmonization endeavors are intended to bring about common standards. It will only be successful if the difference in values and interests is bridged and effective implementation and enforcement are maintained. The approaches of public participation and deliberative democracy aim to integrate more of the society into the AI governance decision-making. Crowdsourcing, assemblies of people, and participatory technology evaluation have the potential to make visible issues and values that are not being reflected in the processes of expertise or industry mono-culture. An effective public engagement means availability of information, real input on decisions and both long-term and not short term consultation. The field of AI governance is mature, even though it has its own maturation, methodological, theoretical, and empirical approaches. There are increasing academic programs and research institutes and professional communities about AI governance. This institutionalization may facilitate tighter and long-term research but cannot lose contact with technical realities as well as on-ground governance requirements.

3.10 Cross-Cutting Themes and Synthesis

The thematic areas and synthesis are cross-cutting because they serve as synergies with the overall research design. An analysis across the various dimensions of responsible agentic AI governance gives several cross-cutting themes that provide insight into underlying tensions and considerations that are repeated across contexts. The issue of tension between innovation and precaution is, perhaps, the most fundamental governance problem. Severe safety standards and massive monitoring may hamper productive use and slack development of technology. On the other hand, poor leadership poses unbearable risks of injuries. The best solutions will probably depend on the situation depending on risk profiles, irreversibility of the possible bad, and the presence of alternative solutions. This tension would be addressed through adaptive governance schemes which would have the capacity to tighten or weaken oversight as information accumulates. The issues of centralization and decentralization are presented as trade-offs in the various dimensions of governance. Centralized control may guarantee uniformity and make use of experience, limiting to bureaucratic stagnation and lack of responsiveness to external influences. Distributed governance embraces diversity and adaptation but can tear up standards and pose challenges to coordination. The idea of hybrid methods that would have centralized principles and a decentralized implementation should be considered. In most situations, transparency and accountability are in a contradiction to privacy and security. The explanation of the AI decision, or making auditing of them available, often needs to access sensitive data or proprietary techniques. Some such privacy-preserving methods as differential privacy, secure multi-party computation, and other methods can somewhat alleviate tensions, at performance or usability cost. There should be an approach to governance that considers such trade-offs. The interdependence of technical and social governance emphasizes the fact that only strategy-specific technical safety precaution measures are insufficient, and only measures based on policies are insufficient. Technical protective measures demand social context of what is harmful behavior. Technical feasibility and enforcement capability is necessitated among policy mechanisms. Good governance incorporates both aspects, albeit with the necessity of performing the task of bridging dissimilar professional cultures and fields of knowledge.

The governance can be dragged towards the short and long term considerations. The present system risks will necessarily require imminent management, which may cause the allocation of resources at the expense of additional equipment in the future. On the other hand, focus on future risks which could be speculative could ignore current evils. The time horizons must be dealt with by governance, taking more of an uncertain approach, but with dynamism to respond. The national and global tensions of governance are the results of the authentic diversity in values and focus among the societies and the actual necessity of the coordination. The delicate balances in wondering at sovereignty and cultural differences and permitting collaboration on common issues cannot be done without being subtle. Implementing the principles of subsidiarity aiming to resolve the problems on the most local possible level, but coordination on the truly global issues, can be beneficial. The best amount of autonomy is one of the questions that appears in applications. The highest autonomy is not necessarily a positive thing; proper autonomy should be determined by the context such as the human abilities, interests of the decisions and reliability of the system. To the extent that greater autonomy is superior, governance must encourage designs that are consistent with autonomy to context.

4. Conclusions

This detailed literature has studied the complex scenery of responsible agentic governance of artificial intelligence that has assimilated knowledge of technical, ethical, policymaking, and practical as well. It is observed that autonomous AI systems introduce new issues of governance qualitatively unrelated to existing AI applications because of the ability to act as goal-directed, involve multi-step, reasoning, interact with the environment, and develop novel capabilities. Through the technical research on safety, the community has come up with an array of effective solutions to the alignment of the agent goals and values of humans, the identification and stopping of undesirable behaviors, and the preservation of significant oversight. The alignment methods such as reward modeling, inverse reinforcement learning, and constitutional AI provide ways in which producing systems that share similar goals with human intentions are achievable. Strong techniques such as adversarial training, formal verification and large scale testing can enhance reliability in a wide variety of situations. Interpretability research can be used to aid in the study of agent reasoning, but there are still inherent contradictions between performance and explainability aspects. Control systems such as interruptibility, corrigibility and impact regularization are a bid to maintain the ability of humans to exert control over autonomous systems. Non-etheless, the modern technical solutions are limited by serious drawbacks. None of the known techniques guarantee safe operation of highly autonomous systems working in sophisticated real-life systems. Emergent capabilities may occur unexpectedly with the scaling of the systems, which may pose new types of risks that are not expected in the development. The verification problem is progressively becoming more urgent in cases where the agent becomes more autonomous, and in cases where agents are being engaged with a longer-term time horizon. Combinatorial complexity brought forth by multi-agent systems puts the current safety assurance approaches to the test. Ethical modelings offer fundamental premises to the translation of human values into practical principles that direct the autonomy of the system development and deployment. The approaches to consequentialist, deontological, and virtue ethics may have significant insights, and it cannot be said that one approach can be accepted globally. The value alignment research struggles with the basic issues regarding definitions, acquisition and maintenance of compatibility to multifaceted and even conflicting human values. Incidences of fairness, accountability, transparency, privacy and autonomy over humans continue to be prevalent issues in application areas, and they do not need a universal solution to do so.

The regulatory and policy environment offers a great level of diversity in terms of jurisdiction, varying practices of different cultural values, institutions, and risk tolerance. The frameworks based on risk, sectoral regulation, principles-based approaches and technology-specific rules have both benefits and drawbacks. The comprehensive governance of a programme might involve the combination of all these approaches instead of relying on one of those mechanisms. Regulation development has however tended to stay behind in technological progress and there has been a gap in the governance which has left significant risks unaddressed. The globalization of AI development and deployment does not have much international coordination. The implementation challenges are major ones on the way of translating the governance principles into practice. It is challenging to verify and monitor compliance to autonomous systems due to their complexity and the emergence of allegations. Auditing involves professional skills and future access to information that might not be readily availed by the deployers. Explainability requirements face the conflicts of model performance and interpretability. Further attribution of responsibility is more involved when autonomy is greater and greater distribution of development occurs. Resource constraint is an issue to both the organizations that want to meet the governance rule setting and regulators that may want to implement them. Adversarial pressures involve dangers of circumventions due to the shortcuts taken by the actors or the exploitation by malicious uses. Data analysis within the realms of applications reflects some universal themes as well as context-related aspects. Each of the healthcare, financial services, transportation, defense and other sectors has its own risk profile, stakeholder interests and regulatory circumstances necessitating particular governance solutions. Nevertheless, transcending issues such as assurance of safety, fairness, privacy, accountability, and human control are manifested across fields. General AI governance principles should be combined with maximum expertise in domains so as to create effective oversight.

The future of agentic AI governance will be determined by several new trends. The agents based on foundation models provide new safety issues and unprecedented capabilities. Neurosymbolic systems can achieve more interpretable systems. The federated and decentralized structures have opportunities and governance complexities. The models of human-AI collaboration would be able to maintain substantial human agency and make use of autonomous capabilities. The new constant learning systems necessitate creating fresh methods of upholding safety assurance. There is a speed in the international coordination, which is driven by the realization of common problems. The mechanisms of participation in the society focus on democratizing governance. Urgent research is needed on critical areas of concern of existing knowledge and practice. Combination of technical safety, ethical values and policy mechanisms is rather insufficient and most of the work is considering them as independent areas. The governance of multi-agent systems needs significantly more development as the number of multiparty systems containing autonomous entities is gaining awareness. Most governance frameworks do not yet have the aspects of adaptive governance which are necessary but not developed yet due to the pace of technological development. Guidance of how the abstract principles should be put into practice in the form of organisational practices is required. The methodology of verification and auditing should be improved to present high-quality control of complex autonomous systems. Attribution theories will have to change so as to handle the issue of responsibility within the context of distributed development and high autonomy. Going forward, a number of priorities can be singled out to enhance responsible agentic AI governance. To start with, a long-term commitment to research on technical safety should be upheld, especially on the aspects of scalable oversight, sound alignment, and checking of safety properties. Second, interdisciplinary work should be increased in order to fill the technical, ethical, and policy gaps and make certain that the governance methods incorporate knowledge in all of them. Third, research is no longer sufficient since empirical studies are required to test the effectiveness of governance interventions in practice as opposed to theoretical treatment or laboratory research. Fourth, the international coordination mechanisms need to be reinforced so that they may cooperate together on common problems without violating the legal diversity in values and priorities. Fifth, the participatory governance mechanisms have to be extended to get more societal involvement so that the decision made regarding the autonomous systems are inclusive of different viewpoints and interests of society members. Sixth, the governance frameworks should be made adaptable so that they can keep up with technology increase and not be replaced by more sophisticated technologies. Seventh, the implementation resources such as guidance documentation, evaluation tools, and training materials are to be generated to assist those organizations that are interested in the responsible deployment of autonomous systems. Eighth, regulatory capacity must be increased by hiring technical skills, building special assessment parabilities and sufficient resource commissioning of the control authorities.

The investments in the realization of the successful governance of agentic AI systems are high. The possibilities associated with autonomous systems are immense in the area of healthcare, education, scientific research, environmental protection, and many more. To achieve these advantages without the occurrence of such grievous harms, an informed frame of rule taking place in a technically aware, ethically rooted, practically capable, and adaptable fashion is needed in relation to the ever-changing circumstances. It is neither the free market of development nor outright banishment that is a response that should be taken. Instead, subtle methods, which tune regulation to the real risks, foster positive innovation without causing harm and retain significant human capacity to make decisions that have consequences are the most promising way to go. Effective achievement in this venture requires a long-term effort by various stakeholder communities. Researchers need to keep progressing in the field of technical safety technique as well as be involved in the aspects related to ethicality and policy. Ethicists need to base normative models on practical knowledge about the strength and weaknesses of technology. Google policymakers need to be able to come up with laws that are informed by the realities on the ground but most importantly, sensitive to the values of the larger society. Safety and ethics should be placed higher in industry in comparison to commercial goals. Civil society needs to ensure a condition of heated vigilance as it understands that there are legitimate and complexities of governance. There should be coordination by international institutions irrespective of difference in interests.

The issue of being responsible in governing agentic AI is here to stay. The more autonomous systems are becoming capable and integrated into social infrastructure, the more they will need continual and adjusting governance. Any solutions are not permanent, and there are only considerate structures that should develop with technological advancements, social and knowledge growth. The one thing which will be unchanged is the necessity to facilitate that autonomous systems are used in human flourishing, meet human values and are subject to substantial human oversight. The global community has to apply its wisdom, efforts, and vigilance in order to satisfy this imperative. It is in that effort that this literature review has attempted to play its small role by synthesizing existing knowledge, defining the areas that need more investigation, and outlining potential areas of future work. It is a difficult road to lead towards responsible agentic AI control, and yet the goal ahead of it, of arriving at a future where the autonomous systems are trusted to serve human interests reliably and avoid harming human values and agency, justifies the effort it will take to reach that goal.

Author Contributions

BB: Conceptualization, methodology, software, resources, visualization, writing review and editing, and supervision. SS: Conceptualization, methodology, software, resources, visualization, writing original draft, writing review and editing.

Conflict of interest

The authors declare no conflicts of interest.

References

[1] Xiao Y, Shi G, Zhang P. Toward agentic ai networking in 6g: A generative foundation model-as-agent approach. IEEE Communications Magazine. 2025 Sep 8;63(9):68-74. https://doi.org/10.1109/MCOM.001.2500005

[2] Qu Y, Huang K, Yin M, Zhan K, Liu D, Yin D, Cousins HC, Johnson WA, Wang X, Shah M, Altman RB. CRISPR-GPT for agentic automation of gene-editing experiments. Nature Biomedical Engineering. 2025 Jul 30:1-4. https://doi.org/10.1038/s41551-025-01463-z

[3] Gabriel I. Artificial intelligence, values, and alignment. Minds and machines. 2020 Sep;30(3):411-37. https://doi.org/10.1007/s11023-020-09539-2

[4] Javaid M, Haleem A, Khan IH, Suman R. Understanding the potential applications of Artificial Intelligence in Agriculture Sector. Advanced Agrochem. 2023 Mar 1;2(1):15-30. https://doi.org/10.1016/j.aac.2022.10.001

[5] Peres RS, Jia X, Lee J, Sun K, Colombo AW, Barata J. Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. IEEE access. 2020 Dec 7;8:220121-39. https://doi.org/10.1109/ACCESS.2020.3042874

[6] Mohsen SE, Hamdan A, Shoaib HM. Digital transformation and integration of artificial intelligence in financial institutions. Journal of Financial Reporting and Accounting. 2025 Mar 20;23(2):680-99. https://doi.org/10.1108/JFRA-09-2023-0544

[7] Tuo Y, Wu J, Zhao J, Si X. Artificial intelligence in tourism: insights and future research agenda. Tourism Review. 2025 Mar 25;80(4):793-812. https://doi.org/10.1108/TR-03-2024-0180

[8] Chen Y, Prentice C. Integrating artificial intelligence and customer experience. Australasian Marketing Journal. 2025 May;33(2):141-53. https://doi.org/10.1177/14413582241252904

[9] Borghoff UM, Bottoni P, Pareschi R. Human-artificial interaction in the age of agentic AI: a system-theoretical approach. Frontiers in Human Dynamics. 2025 May 9;7:1579166. https://doi.org/10.3389/fhumd.2025.1579166

[10] Plaat A, van Duijn M, Van Stein N, Preuss M, van der Putten P, Batenburg KJ. Agentic large language models, a survey. Journal of Artificial Intelligence Research. 2025 Dec 30;84. https://doi.org/10.1613/jair.1.18675

[11] Gridach M, Nanavati J, Abidine KZ, Mendes L, Mack C. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. arXiv preprint arXiv:2503.08979. 2025 Mar 12.

[12] Chan A, Salganik R, Markelius A, Pang C, Rajkumar N, Krasheninnikov D, Langosco L, He Z, Duan Y, Carroll M, Lin M. Harms from increasingly agentic algorithmic systems. InProceedings of the 2023 ACM conference on fairness, accountability, and transparency 2023 Jun 12 (pp. 651-666). https://doi.org/10.1145/3593013.3594033

[13] Wei J, Yang Y, Zhang X, Chen Y, Zhuang X, Gao Z, Zhou D, Wang G, Gao Z, Cao J, Qiu Z. From ai for science to agentic science: A survey on autonomous scientific discovery. arXiv preprint arXiv:2508.14111. 2025 Aug 18.

[14] Jiang F, Pan C, Wang K, Michiardi P, Dobre OA, Debbah M. From large ai models to agentic ai: A tutorial on future intelligent communications. IEEE Journal on Selected Areas in Communications. 2026 Feb 2. https://doi.org/10.1109/JSAC.2026.3660010

[15] Mohammed A. Agentic AI as a Proactive Cybercrime Sentinel: Detecting and Deterring Social Engineering Attacks. Journal of Data and Digital Innovation (JDDI). 2025 Jun 16;2(2):109-17.

[16] Belcak P, Heinrich G, Diao S, Fu Y, Dong X, Muralidharan S, Lin YC, Molchanov P. Small language models are the future of agentic ai. arXiv preprint arXiv:2506.02153. 2025 Jun 2.

[17] Bousetouane F. Agentic systems: A guide to transforming industries with vertical ai agents. arXiv preprint arXiv:2501.00881. 2025 Jan 1. https://doi.org/10.32388/2DKDCK

[18] Sapkota R, Roumeliotis KI, Karkee M. Vibe coding vs. agentic coding: Fundamentals and practical implications of agentic ai. arXiv preprint arXiv:2505.19443. 2025 May 26.

[19] Floridi L. AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis: L. Floridi. Philosophy & Technology. 2025 Mar;38(1):30. https://doi.org/10.1007/s13347-025-00858-9

[20] Meng L. From Reactive to Proactive: Integrating Agentic AI and Automated Workflows for Intelligent Project Management (AI-PMP). Frontiers in Engineering. 2025 Nov 26;1(1):82-93. https://doi.org/10.63313/FE.9003

[21] Shavit Y, Agarwal S, Brundage M, Adler S, O'Keefe C, Campbell R, Lee T, Mishkin P, Eloundou T, Hickey A, Slama K. Practices for governing agentic AI systems. Research Paper, OpenAI. 2023 Dec 14.

[22] Singh A, Ehtesham A, Kumar S, Khoei TT. Enhancing AI systems with agentic workflows patterns in large language model. In2024 IEEE World AI IoT Congress (AIIoT) 2024 May 29 (pp. 527-532). IEEE. https://doi.org/10.1109/AIIoT61789.2024.10578990

[23] Townsend DM, Hunt RA, Rady J, Manocha P, Jin JH. Are the futures computable? Knightian uncertainty and artificial intelligence. Academy of Management Review. 2025 Apr;50(2):415-40. https://doi.org/10.5465/amr.2022.0237

[24] Qiu J, Lam K, Li G, Acharya A, Wong TY, Darzi A, Yuan W, Topol EJ. LLM-based agentic systems in medicine and healthcare. Nature Machine Intelligence. 2024 Dec;6(12):1418-20. https://doi.org/10.1038/s42256-024-00944-1

[25] Li Y, Zhou X, Yin HB, Chiu TK. Design language learning with artificial intelligence (AI) chatbots based on activity theory from a systematic review. Smart Learning Environments. 2025 Mar 10;12(1):24. https://doi.org/10.1186/s40561-025-00379-0

[26] Markauskaite L, Marrone R, Poquet O, Knight S, Martinez-Maldonado R, Howard S, Tondeur J, De Laat M, Shum SB, Gašević D, Siemens G. Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?. Computers and Education: Artificial Intelligence. 2022 Jan 1;3:100056. https://doi.org/10.1016/j.caeai.2022.100056

[27] Sudarshan M, Shih S, Yee E, Yang A, Zou J, Chen C, Zhou Q, Chen L, Singhal C, Shih G. Agentic llm workflows for generating patient-friendly medical reports. arXiv preprint arXiv:2408.01112. 2024 Aug 2.

[28] McIntosh TR, Susnjak T, Liu T, Watters P, Xu D, Liu D, Halgamuge MN. From Google Gemini to OpenAI Q*(Q-Star): a survey on reshaping the generative artificial intelligence (AI) research landscape. Technologies. 2025 Jan 30;13(2):51. https://doi.org/10.3390/technologies13020051

[29] Ladak A, Loughnan S, Wilks M. The moral psychology of artificial intelligence. Current Directions in Psychological Science. 2024 Feb;33(1):27-34. https://doi.org/10.1177/09637214231205866

[30] Teo ZL, Thirunavukarasu AJ, Elangovan K, Cheng H, Moova P, Soetikno B, Nielsen C, Pollreisz A, Ting DS, Morris RJ, Shah NH. Generative artificial intelligence in medicine. Nature medicine. 2025 Oct;31(10):3270-82. https://doi.org/10.1038/s41591-025-03983-2

[31] Hanna MG, Pantanowitz L, Dash R, Harrison JH, Deebajah M, Pantanowitz J, Rashidi HH. Future of artificial intelligence-machine learning trends in pathology and medicine. Modern Pathology. 2025 Apr 1;38(4):100705. https://doi.org/10.1016/j.modpat.2025.100705

[32] Habler I, Huang K, Narajala VS, Kulkarni P. Building a secure agentic AI application leveraging A2A protocol. arXiv preprint arXiv:2504.16902. 2025 Apr 23.

[33] Moradbakhti L, Schreibelmayr S, Mara M. Do men have no need for "feminist" artificial intelligence? Agentic and gendered voice assistants in the light of basic psychological needs. Frontiers in psychology. 2022 Jun 14;13:855091. https://doi.org/10.3389/fpsyg.2022.855091

[34] Kolagani SH. Agentic Automation and Work Flow Orchestration in Enterprise SaaS: Effects on Ticket Resolution Time and Employee Productivity in IT Service Management. IJSAT-International Journal on Science and Technology. 2024 Nov 8;15(4). https://doi.org/10.71097/IJSAT.v15.i4.9876

[35] Todupunuri A. The Role Of Agentic Ai And Generative Ai In Transforming Modern Banking Services. American Journal of AI Cyber Computing Management. 2025 Sep 26;5(3):85-93. https://doi.org/10.64751/ajaccm.2025.v5.n3.pp85-93

[36] Gosmar D, Dahl DA. Hallucination mitigation using agentic ai natural language-based frameworks. arXiv preprint arXiv:2501.13946. 2025 Jan 19.

[37] Yamada Y, Lange RT, Lu C, Hu S, Lu C, Foerster J, Clune J, Ha D. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. arXiv preprint arXiv:2504.08066. 2025 Apr 10.

[38] Kshetri N. Transforming cybersecurity with agentic AI to combat emerging cyber threats. Telecommunications Policy. 2025 Jul 1;49(6):102976. https://doi.org/10.1016/j.telpol.2025.102976

[39] Transparency in the reporting of artificial intelligence-the TITAN guideline. Premier Journal of Science. 2025;10:100082.

[40] Hanna MG, Pantanowitz L, Dash R, Harrison JH, Deebajah M, Pantanowitz J, Rashidi HH. Future of artificial intelligence (AI)-machine learning (ML) trends in pathology and medicine. Modern Pathology. 2025 Jan 4:100705. https://doi.org/10.1016/j.modpat.2025.100705

[41] Chen E, Prakash S, Janapa Reddi V, Kim D, Rajpurkar P. A framework for integrating artificial intelligence for clinical care with continuous therapeutic monitoring. Nature Biomedical Engineering. 2025 Apr;9(4):445-54. https://doi.org/10.1038/s41551-023-01115-0

[42] Liu SY. Artificial intelligence (AI) in agriculture. IT professional. 2020 May 21;22(3):14-5. https://doi.org/10.1109/MITP.2020.2986121

[43] Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer science. 2020 May;111(5):1452-60. https://doi.org/10.1111/cas.14377

[44] Schwendicke FA, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. Journal of dental research. 2020 Jul;99(7):769-74. https://doi.org/10.1177/0022034520915714

[45] Verganti R, Vendraminelli L, Iansiti M. Innovation and design in the age of artificial intelligence. Journal of product innovation management. 2020 May;37(3):212-27. https://doi.org/10.1111/jpim.12523

[46] Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works. Ai & Society. 2024 Aug;39(4):1871-82. https://doi.org/10.1007/s00146-023-01635-y

[47] Rashidi HH, Pantanowitz J, Hanna MG, Tafti AP, Sanghani P, Buchinsky A, Fennell B, Deebajah M, Wheeler S, Pearce T, Abukhiran I. Introduction to artificial intelligence and machine learning in pathology and medicine: generative and nongenerative artificial intelligence basics. Modern Pathology. 2025 Apr 1;38(4):100688. https://doi.org/10.1016/j.modpat.2024.100688

[48] Waisberg E, Ong J, Kamran SA, Masalkhi M, Paladugu P, Zaman N, Lee AG, Tavakkoli A. Generative artificial intelligence in ophthalmology. Survey of ophthalmology. 2025 Jan 1;70(1):1-1. https://doi.org/10.1016/j.survophthal.2024.04.009

[49] Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, Chandak P, Liu S, Van Katwyk P, Deac A, Anandkumar A. Scientific discovery in the age of artificial intelligence. Nature. 2023 Aug 3;620(7972):47-60. https://doi.org/10.1038/s41586-023-06221-2

[50] Banh L, Strobel G. Generative artificial intelligence. Electronic Markets. 2023 Dec;33(1):63. https://doi.org/10.1007/s12525-023-00680-1

[51] Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE transactions on industrial informatics. 2022 Jan 27;18(8):5031-42. https://doi.org/10.1109/TII.2022.3146552

[52] Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, Bin Saleh K, Alowais SA, Alshaya OA, Rahman I, Al Yami MS, Albekairy AM. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Research in social and administrative pharmacy. 2023 Aug 1;19(8):1236-42. https://doi.org/10.1016/j.sapharm.2023.05.016

[53] Bankins S, Formosa P. The ethical implications of artificial intelligence (AI) for meaningful work. Journal of Business Ethics. 2023 Jul;185(4):725-40. https://doi.org/10.1007/s10551-023-05339-7

[54] Abulibdeh A, Zaidan E, Abulibdeh R. Navigating the confluence of artificial intelligence and education for sustainable development in the era of industry 4.0: Challenges, opportunities, and ethical dimensions. Journal of Cleaner Production. 2024 Jan 15;437:140527. https://doi.org/10.1016/j.jclepro.2023.140527

[55] King MR, ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cellular and molecular bioengineering. 2023 Feb;16(1):1-2. https://doi.org/10.1007/s12195-022-00754-8

[56] Ikhsan RB, Fernando Y, Prabowo H, Gui A, Kuncoro EA. An empirical study on the use of artificial intelligence in the banking sector of Indonesia by extending the TAM model and the moderating effect of perceived trust. Digital Business. 2025 Jun 1;5(1):100103. https://doi.org/10.1016/j.digbus.2024.100103

[57] Gangwal A, Lavecchia A. Artificial intelligence in natural product drug discovery: current applications and future perspectives. Journal of medicinal chemistry. 2025 Feb 7;68(4):3948-69. https://doi.org/10.1021/acs.jmedchem.4c01257

[58] Crompton H, Burke D. Artificial intelligence in higher education: the state of the field. International journal of educational technology in higher education. 2023 Apr 24;20(1):22. https://doi.org/10.1186/s41239-023-00392-8

[59] Puntoni S, Reczek RW, Giesler M, Botti S. Consumers and artificial intelligence: An experiential perspective. Journal of marketing. 2021 Jan;85(1):131-51. https://doi.org/10.1177/0022242920953847

[60] Kumar I, Rawat J, Mohd N, Husain S. Opportunities of artificial intelligence and machine learning in the food industry. Journal of Food Quality. 2021;2021(1):4535567. https://doi.org/10.1155/2021/4535567

[61] Moor M, Banerjee O, Abad ZS, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature. 2023 Apr 13;616(7956):259-65. https://doi.org/10.1038/s41586-023-05881-4

[62] Jaiswal A, Arun CJ, Varma A. Rebooting employees: Upskilling for artificial intelligence in multinational corporations. InArtificial intelligence and international HRM 2023 May 22 (pp. 114-143). Routledge. https://doi.org/10.4324/9781003377085-5

[63] Hosseini S, Seilani H. The role of agentic ai in shaping a smart future: A systematic review. Array. 2025 Jul 1;26:100399. https://doi.org/10.1016/j.array.2025.100399

[64] Murugesan S. The rise of agentic AI: implications, concerns, and the path forward. IEEE Intelligent Systems. 2025 Apr 10;40(2):8-14. https://doi.org/10.1109/MIS.2025.3544940

[65] Acharya DB, Kuppan K, Divya B. Agentic AI: Autonomous intelligence for complex goals-A comprehensive survey. IEEe Access. 2025 Jan 22;13:18912-36. https://doi.org/10.1109/ACCESS.2025.3532853

[66] Hughes L, Dwivedi YK, Malik T, Shawosh M, Albashrawi MA, Jeon I, Dutot V, Appanderanda M, Crick T, De' R, Fenwick M. AI agents and agentic systems: A multi-expert analysis. Journal of Computer Information Systems. 2025 Jul 4;65(4):489-517. https://doi.org/10.1080/08874417.2025.2483832

[67] Vanneste BS, Puranam P. Artificial intelligence, trust, and perceptions of agency. Academy of Management Review. 2024 Mar 22(ja):amr-2022. https://doi.org/10.2139/ssrn.3897704

[68] Schneider J. Generative to agentic ai: Survey, conceptualization, and challenges. arXiv preprint arXiv:2504.18875. 2025 Apr 26.

[69] Karunanayake N. Next-generation agentic AI for transforming healthcare. Informatics and Health. 2025 Sep 1;2(2):73-83. https://doi.org/10.1016/j.infoh.2025.03.001