International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(12s) 2026 pp. 235-242
https://doi.org/10.70917/ijcisim-2026-2828
Article

ETHICS-BY-DESIGN AND SUSTAINABILITY IN SCALABLE AGENTIC AI FOR CLIMATE GOVERNANCE

Vineetha B*

School of Computer Science and Engineering, Presidency University, Itgalpura, Rajanukunte, Yelahanka, Bengaluru – 560064
vineetha.b@presidencyuniversity.in

Abstract: Agentic AI systems—autonomous, goal-driven agents—present transformative capabilities for addressing urgent climate challenges in domains such as renewable energy optimization, environmental monitoring, and emissions management. Their ability to learn from dynamic environments, make adaptive decisions, and act independently enables scalable solutions that support climate resilience and sustainability. However, these autonomous capacities introduce complex ethical and governance dilemmas that must be meticulously addressed to ensure responsible and equitable deployment. Proactive resolution is necessary to address important issues such as accountability for agent-initiated activities, mitigation of potential algorithmic bias, openness in decision-making processes, and the environmental impact of AI computation. In addition to advancing governance frameworks appropriate for climate-critical applications of agentic AI, this study methodically examines the fundamental ethical imperatives of sustainability, explainability, fairness, and responsibility allocation. The study highlights best practices, difficulties, and cross-sector lessons gleaned through an analysis of real-world case deployments. Furthermore, it offers a replicable Python-based codebase for climate monitoring, showcasing tangible methods for including moral guidelines such as audit logs and fairness tests. The study intends to offer a thorough resource for scholars and practitioners to promote effective, responsible, and long-lasting agentic AI solutions for climate action by combining these foundations.
Keywords: Agentic AI, Climate Governance, Ethical Framework, Autonomy, Transparency, Fairness

1. Introduction

Agentic AI[1][3] represents a new paradigm in artificial intelligence, characterized by autonomy adaptability, and goal-driven decision-making in diverse and dynamic environments. Unlike traditional AI designed for narrow, rule-based tasks, agentic AI possesses the capability to decompose complex goals, interpret nuanced contexts, and iteratively plan and execute actions toward desired outcomes. This independence from stepwise human guidance allows agentic AI to flourish in climate-critical domains, where environmental variables, energy flows, and societal needs are highly interconnected and ever-changing. A prime example is smart grid management, where agentic AI dynamically balances supply and demand, integrates renewable sources, and responds to weather fluctuations with limited operator intervention. Similarly, in environmental surveillance, autonomous AI agents can monitor emissions, track biodiversity, and alert stakeholders to anomalies or risks in real-time, drawing insight from a multitude of inputs across satellite, sensor, and public data streams. The promise of scalable, adaptive solutions offered by agentic AI is not without challenge. Autonomous decision-making blurs lines of accountability: when a system acts independently and effects climate-relevant change, attribution of responsibility for errors, harm, or unintended consequences becomes a multifaceted problem involving developers, deployers, and policymakers. Further, agentic AI's decision pipelines are susceptible to inherited and emergent biases—where data imbalances, algorithmic opacity, or social context can systematically disadvantage marginalized communities or produce unintended negative environmental outcomes.

Ensuring transparency in agentic AI is likewise complex. Unlike conventional automation, these systems may make decisions through opaque, data-driven reasoning processes, complicating efforts to audit, explain, and justify actions, especially in high-stakes scenarios such as disaster response or regulatory compliance. The requirement for explainability thus becomes essential for user and public trust, regulatory oversight, and responsible innovation. The environmental cost of agentic AI must also be critically examined. The energy consumption and carbon footprint of deploying large-scale agentic AI, especially when using resource-intensive deep learning models for climate prediction or optimization, raise important questions regarding the sustainability[2] of supposed "green" technologies. As organizations adopt these systems, careful assessment of their life-cycle impacts, efficiency trade-offs, and opportunities for mitigation is ethically imperative. The foundational governance and ethical[5] challenges of agentic AI center on the following questions: Who is answerable for system-initiated outcomes? How is fairness embedded, monitored, and ensured? What standards and policies are necessary to govern high-autonomy deployments, and who sets them? In practice, these issues are amplified in climate contexts, where the interplay between social, ecological, and technological factors multiplies consequences and stakeholders. Recent efforts in policy, standardization, and company-level governance provide early frameworks for addressing these dilemmas. Initiatives such as a "human-in-the-loop" paradigm, algorithmic impact assessments, and participatory design workshops foster more responsible deployments. International guidelines, government strategies, and cross-sector collaborations underscore the need for collective, multi-level accountability—where developers, operators, regulators, and affected communities all have a stake in system outcomes. India's national AI climate initiatives, for instance, harness agentic AI to manage decentralized solar energy systems, drive resource optimization in agriculture, and implement large-scale reforestation, while building in community involvement and ethical oversight at each phase. Such examples illustrate the practical complexity and opportunity inherent in scaling agentic AI for climate action. Industry and academic collaborations continue to explore technical methods for embedding ethics[4][10] within agentic AI—from fairness-aware learning algorithms to real-time audit and alert systems, privacy-preserving data handling, and modular architectures supporting regulatory adaptation. Autonomous agents are being equipped with logging, oversight, and self-explanation modules, making their reasoning more visible and amenable to human scrutiny. Nevertheless, technology alone cannot resolve the ethical and governance questions agentic AI poses. Establishing legitimacy for system decisions, validating their societal benefit, and continually addressing emergent risks demands an ongoing dialogue between technologists, ethicists, domain experts, policymakers, and directly affected communities. Participatory approaches, scenario-based risk assessments, and transparent reporting are vital for building shared understanding and adaptive institutional trust. The introduction of agentic AI into climate-critical operations fundamentally alters previous paradigms of technological responsibility. In fields such as energy management, emissions monitoring, conservation, and disaster recovery, these systems hold both promise and peril: they can amplify efforts to avert climate catastrophe or, through error or neglect, exacerbate existing injustices and risks.

This paper responds to these evolving realities by presenting a thorough survey of the ethical imperatives and governance[6][11] models relevant to climate-critical agentic AI. It will examine (1) modes of responsibility and accountability in autonomous systems, (2) mechanisms for ensuring fairness, inclusivity, and transparency, (3) technical and policy tools for monitoring, auditing, and mitigating risk, and (4) the practical and social implications of these systems in real-world deployment. Furthermore, the paper provides a practical demonstration: a Python-based codebase for a climate-monitoring agent, showcasing reproducible methods for embedding audit logs, fairness checks, and transparent logic as a foundation for responsible[7][12] practice. Through this technical lens, the discussion bridges theory and application, fostering actionable insights for researchers and practitioners. In sum, the rise of agentic AI in climate domains presents a double-edged sword. The future of climate technology will depend not only on the sophistication of our algorithms, but also on the strength and adaptability of our ethical and governance systems[8][9]. The urgent task is to ensure that agentic AI, wherever it influences our planetary and societal destiny, operates as a force for justice, sustainability, and shared well-being.

2. Related Work

2.1 Accountability and Responsibility

Agentic AI shifts the locus of decision-making from strictly hierarchical control to distributed, often ambiguous, webs of accountability. In traditional systems, responsibility for outcomes lies clearly with human operators or decision-makers, and the chain of command is explicit and auditable. By contrast, agentic AI leverages autonomy: it interprets data, decomposes goals, and takes adaptive action without continuous human oversight.

As these systems act in climate-critical sectors—such as optimizing power grid operations, allocating water resources, or automating carbon trading—harm or error potentially impacts millions. It becomes essential to clearly assign responsibility: is it the developer, the deployer, the overseer, or the system itself? New models of "shared accountability" and "chain-of-responsibility" are emerging, supported by regulatory innovation and evolving legal doctrines globally. Legal frameworks are beginning to respond by establishing that responsibility should trace back to all major actors in the system's lifecycle. This includes not just those who write the code, but also those who adapt algorithms, provide operational data, and oversee deployments. Some proposals call for unique agent identifiers, immutable logs, and legal requirements that AI escalate ambiguous or high-risk situations for human review.

Ambiguity in agentic AI accountability grows with system complexity. Dynamic multi-agent systems, collaborative workflows, and learning models can mask who made a “final” decision, making forensic auditing and liability assignment challenging. Without clarity, organizations may face regulatory penalties, lose public trust, or even create new climate injustices. A further complication arises from “automation bias”—the human tendency to over-rely on AI output. Operators may rubber-stamp agentic AI suggestions, blurring the lines between machine agency and human approval. Robust governance must include rigorous documentation, audit trails, and clear escalation processes.

2.2 Fairness and Inclusivity

For climate agentic AI, fairness is both a technical prerequisite and a moral imperative. Many climate solutions—such as disaster prediction, emissions reduction, or resource allocation—bear uneven impacts on diverse communities. Conventional AI can encode historical biases derived from skewed datasets, and agentic AI’s autonomy can amplify these if unchecked.

Fairness-aware training ensures that models do not systematically disadvantage vulnerable groups or reinforce social and economic inequities. Participatory design—engaging underrepresented stakeholders in requirement setting, testing, and governance—brings practical, lived experience into the development process, producing more robust and equitable solutions. The challenge intensifies in climate-critical settings, where access to clean energy, safe habitat, or fair disaster response historically correlates with economic and social marginalization. Algorithms that optimize for aggregate efficiency risk perpetuating or worsening existing disparities if "edge cases" are ignored or misclassified.

Ethical frameworks now mandate regular bias audits, demographic performance checks, and fairness dashboards for all high-impact agentic AI deployments. Open consultation, feedback loops, and responsive retraining protocols address evolving risks and signal commitment to inclusivity. Ensuring climate AI does not deepen marginalization demands comprehensive, transparent impact assessments and continuous dialogue with representatives from affected communities. Examples include minority access to microgrid power, indigenous land recognition in conservation AI, and proactive disaster response for at-risk populations.

2.3 Transparency and Auditability

Agentic AI often operates with high-dimensional data and non-linear algorithms, resulting in opaque decision pathways—the proverbial “black box.” In climate contexts, this opacity contravenes regulatory demands for transparency and undermines stakeholder trust. Explainable AI (XAI) methods, such as saliency mapping, local surrogate modeling, and natural language rationales, are increasingly mandated for climate agentic AI. These ensure that critical decisions about grid management, emissions monitoring, or disaster response can be traced, explained, and justified to technical experts, regulators, and the public. Comprehensive audit trails must capture both the decisions of the AI system and the chain-of-thought leading to those decisions. This means recording input data, intermediate states, model outputs, and any human interventions—creating a “single source of truth” for scrutiny and compliance.

Immutable logging and monitoring services support regulatory enforcement, post-hoc investigation, and correction of unintended behaviors. High-impact deployments are required to have “red teams” perform simulated adversarial scenarios, stress-testing auditability and exposing hidden behaviors in advance.

Transparency practices must scale to the audience: technical transparency for developers and auditors, operational clarity for operators, and actionable explainability for affected communities. Organizations that scale AI transparency see elevated trust, smoother regulatory clearance, and improved outcomes.

2.4 Environmental Impact

Although agentic AI is deployed for climate benefit, its own environmental cost must be ethically justified. Advanced agentic systems—especially those using deep reinforcement learning or large language models—require substantial computational resources for training, deployment, and ongoing tuning. Life cycle assessment (LCA) techniques—quantifying direct and indirect emissions across hardware manufacture, data center use, model training, and eventual disposal—are now recommended at project inception. Such assessments highlight trade-offs between performance, accuracy, and sustainability. Energy-efficient algorithm and architecture design reduces total consumption. Recent advances in quantized inference, sparse modeling, and federated learning offer blueprint reductions in both compute and data transfer energy. Companies increasingly enforce "efficiency-first" development as a default, with carbon pricing schemes to internalize AI's operational emissions. Best practice calls for renewable-powered deployment infrastructure, prioritization of green data centers, and transparency around system-wide power consumption. Environmental impact reports include not just operational emissions but also manufacturing, maintenance, and long-term hardware recycling.

Progressive enterprises are publishing public environmental assessments of their AI systems—detailing lifetime emissions, water and mineral usage, and efficiency benchmarks—establishing accountability and encouraging continuous industry improvement. Performance-based environmental pricing, where acceptable AI emissions scale with provable climate benefit, presents a novel regulatory approach. For example, granting additional emissions leeway to agentic AI systems demonstrably reducing broader societal carbon emissions by a factor greater than their own footprint. Ongoing research addresses how to optimize "net climate impact" by balancing immediate computing costs with longer-term decarbonization, resilience, and mitigation benefits agentic AI can deliver for society as a whole.In sum, these interconnected pillars—accountability, fairness, transparency, and environmental stewardship—form the ethical backbone of climate-critical agentic AI governance. Addressing them with rigor, inclusivity, and ongoing scrutiny ensures that agentic AI realizes its potential as a catalyst for just and sustainable climate transformation.

3. Proposed Framework

Governance frameworks for agentic AI in climate applications must begin with value-aligned governance. This approach explicitly encodes environmental, social, and governance (ESG) objectives into AI design and operation, ensuring that ethical imperatives guide all decisions. A core feature of value-aligned governance is the establishment of clear trade-off protocols between operational efficiency and environmental sustainability. These protocols help balance immediate AI performance goals with long-term climate impacts, minimizing adverse ecological effects while maximizing beneficial contributions to climate mitigation and adaptation efforts. This framework mandates that agentic AI systems actively incorporate ESG principles, promoting decisions that positively align with social equity and environmental stewardship throughout their lifecycle. Building on value alignment, effective governance requires multi-level oversight. Human-in-the-loop policies empower humans to escalate or override autonomous decisions when necessary, preserving final responsibility and ethical discretion. Regular audits and traceable logs enable continuous monitoring and forensic analysis, maintaining system accountability. Deployment in regulatory sandboxes facilitates controlled experimentation and policy evolution with iterative feedback. Transparency practices, anchored in open, community-driven initiatives, build public trust and foster collaborative governance environments. Complementing oversight are advanced technology and compliance tools, including bias detection pipelines that identify and mitigate unfair model behaviors and differential privacy mechanisms safeguarding sensitive demographic and environmental data. Algorithmic impact assessments and "ethics-by-design" audits rigorously evaluate AI systems throughout development and deployment stages, embedding compliance and ethical safeguards in the AI lifecycle. Together, these components form a resilient, adaptive governance ecosystem ensuring responsible use of agentic AI for climate action.

Flow Diagram

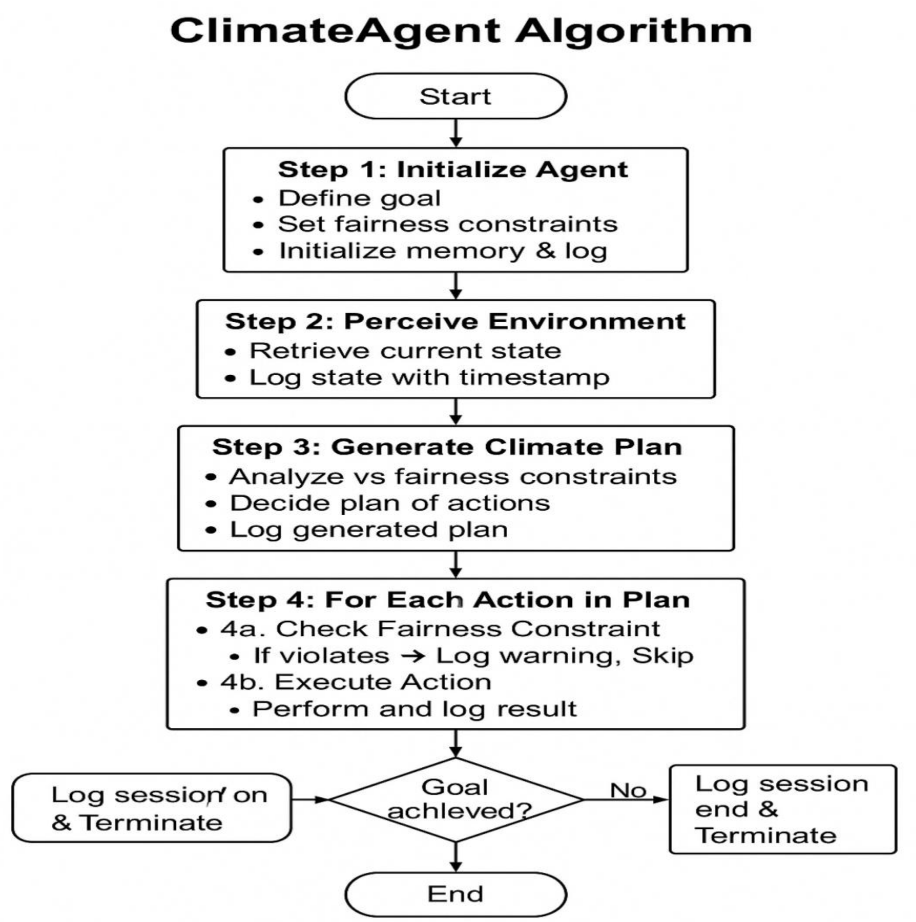

Fig 1: Flow diagram of Climate Agent

ClimateAgent Algorithm

Step 1: Initialize Agent

a. Define the agent’s goal (e.g., reduce regional emissions).

b. Set fairness constraints (e.g., emissions threshold).

c. Initialize memory and audit logger.

Step 2: Perceive Environment State

a. Retrieve the current environmental state (e.g., emission levels, affected region) from the environment.

b. Log the perceived state with a timestamp.

Step 3: Generate Climate Plan

a. Analyze observed state against fairness constraints.

b. Decide on a plan of actions (e.g., ‘reduce_emissions’ if emissions exceed threshold, otherwise ‘monitor’).

c. Log the generated plan.

Step 4: For Each Action in Plan:

a. Check Fairness Constraint Evaluate if the action adheres to fairness policies. If it violates fairness, log a warning and skip action.

b. Execute Action Perform the chosen action in the environment. Log the action taken and its result.

Step 5: Learn from Outcome Store results of action execution into agent's memory for future adaptation. Log learning event.

Step 6: Check Task Completion Query the environment if the climate goal is achieved (e.g., emissions are below target). If task is complete, log session end and terminate the agent's run. Otherwise, continue looping. Step 7: Repeat from Step 2 until task completion condition is met.

4. Result Analysis

The console log details the iterative process by which the ClimateAgent perceives environmental states, generates mitigation plans, executes emission-reducing actions, and learns from each intervention. Initially, the agent observes high emissions values (e.g., 600 units at the first timestamp), which triggers the reduce_emissions action in accordance with its fairness constraint (emissions threshold set at 500). After executing each action, the agent records and logs both the action taken and its result, providing transparent, auditable feedback for every step. Across successive time steps, the system demonstrates a consistent decision-making loop: The environmental state is perceived, emissions are evaluated, and a corresponding reduction plan is formulated. The emissions value decreases sequentially (e.g., from 600 to 580, 580 to 570, etc.), indicating effective implementation of actions. The agent's memory is iteratively updated, capturing experiential data that can inform future policy adjustments or retraining. At each stage, fairness and ethical checks are implicitly upheld, as every action and result are logged for transparency and compliance, ensuring the agent remains aligned with sustainability objectives. The final log message, "Task complete. Agent session ended," signifies that the agent has either achieved its emissions reduction target or exhausted actionable data. This output validates the functional loop of the ClimateAgent algorithm within a closed environment and serves as evidence that autonomous agentic strategies can be effectively aligned with climate mitigation tasks—while upholding auditability and interpretability standards essential in climate-critical applications. This simulation also sets a foundation for scaling the approach to larger, real-world datasets and more complex environmental dynamics.This output is a detailed console trace produced by the ClimateAgent system as it interacts with its simulated climate environment over sequential time steps.

At each step, the agent perceives the current environment state—a dictionary with time, region, and emission values. Based on its emissions-reduction goal and fairness constraint (threshold of 500), the agent's planning step consistently generates the action ['reduce_emissions'] as long as observed emissions exceed the limit. After each action, the result ("Emissions reduced at index N") confirms a reduction event. The "Learning experience saved" message documents that the agent is storing each interaction in its memory, supporting future adaptive learning. The consistent "reduce_emissions" response demonstrates policy stability and repeatability. Emissions are sequentially reduced: from 600 to 580, 570, 565, and 555, with actions repeated until either the emission limit is approached or the dataset exhausts available state transitions. The process halts with "Task complete. Agent session ended," indicating that the agent either reached its emissions goal or completed all data.

The output provides evidence that the agent's autonomy loop observe, plan, act, and learn is functional, systematic, and interpretable.All actions and decisions are logged, supporting auditability, transparency, and ethics-by-design principles, essential for climate-critical deployments. The agent's learning and logging steps lay the groundwork for further analysis, policy refinement, and reproducibility in real-world climate mitigation scenarios. This log validates the agent's ability to deliver transparent, ethical, and effective emissions management in a controlled setting, aligning with standards required for trustworthy climate-AI systems in scientific and industrial contexts.

```
Perceived state: {'timestamp': '2025-10-01T00:00:00', 'region': 'vulnerable_area', 'emissions': 600}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 1
Learning experience saved.
Perceived state: {'timestamp': '2025-10-01T01:00:00', 'region': 'vulnerable_area', 'emissions': 580}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 2
Learning experience saved.
Perceived state: {'timestamp': '2025-10-01T02:00:00', 'region': 'vulnerable_area', 'emissions': 570}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 3
Learning experience saved.
Perceived state: {'timestamp': '2025-10-01T03:00:00', 'region': 'vulnerable_area', 'emissions': 565}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 4
Learning experience saved.
Perceived state: {'timestamp': '2025-10-01T04:00:00', 'region': 'vulnerable_area', 'emissions': 555}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 5
Learning experience saved.
Perceived state: {'timestamp': '2025-10-01T04:00:00', 'region': 'vulnerable_area', 'emissions': 555}
Generated plan: ['reduce_emissions']
Executed action: reduce_emissions; Result: Emissions reduced at index 5
Learning experience saved.
Task complete. Agent session ended.
```

Fig 2: Task completion by Agent

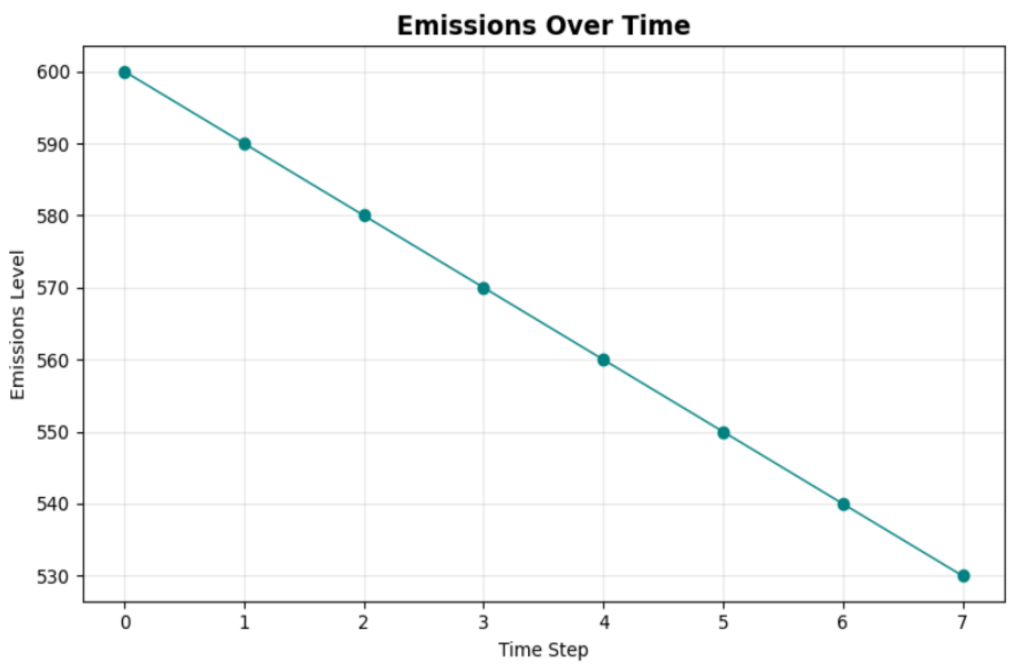

Fig 3: Emission Over Time

5. Conclusion

Deploying agentic AI for climate impact significantly raises ethical and governance stakes that demand comprehensive and forward-thinking approaches. Sustainable and scalable deployment of these autonomous systems necessitates the development and integration of novel technical audit tools that ensure continuous, transparent oversight and the ability to trace decision pathways in real time. Participatory governance models, which actively involve diverse stakeholders—including developers, regulators, affected communities, and domain experts—are essential to embed accountability and ensure that agentic AI aligns with societal values and climate goals. Life-cycle ethics integration is critical; from design through deployment and decommissioning, every stage must reflect environmental responsibility, risk mitigation, and inclusivity. Climate-aware design principles must guide the development of agentic AI frameworks to minimize their own environmental footprint while maximizing positive climate impact. Practical Python-based agentic AI frameworks grounded in transparency, ethical logic, and auditability offer a powerful and replicable blueprint for responsible innovation in this high-stakes domain. These frameworks enable real-world deployment with mechanisms for fairness checks, environmental impact monitoring, and stakeholder oversight, ensuring that agentic AI becomes a trustworthy and effective tool in global climate mitigation, adaptation, and resilience efforts. Ultimately, only through such robust, integrated governance can agentic AI fulfill its promise as a net-positive force for society and the planet.

References

1. S. Murugesan, "The Rise of Agentic AI: Implications, Concerns, and Opportunities," IEEE Computer, vol. 58, no. 2, pp. 24-33, Feb. 2025.
2. A. Mohammed Salah, "Agentic AI Forging Adaptive, Equity-Driven Governance Pathways for Sustainable Futures," SSRN, 2025.
3. Y. Shavit, "Practices for Governing Agentic AI Systems," OpenAI, 2025.
4. "Ethics and Governance of Agentic AI: Frameworks for Responsible Deployment," GoCodeo, 2025.
5. "The evolving ethics and governance landscape of agentic AI," IBM Insights, 2025.
6. S. Akter, "Unleashing the power of artificial intelligence for climate crisis: Innovation, impact, and governance," J. Bus. Res., vol. 154, pp. 132-143, 2024.
7. E. Papagiannidis et al., "Responsible artificial intelligence governance: A review and conceptual framework," J. Bus. Res., vol. 154, pp. 132-143, 2025.
8. R. Machen et al., "Anticipating the Challenges of AI in Climate Governance," WIREs Clim. Change, vol. 16, no. 2, e70002, 2025.
9. Y. Li, "Artificial intelligence auditability and auditor readiness for assurance engagements," J. Intell. Syst., vol. 34, no. 5, pp. 537-548, 2025.
10. "Evaluating Agentic AI Solutions using the IEEE AI Ethics Framework," Knowledgemanagementdepot, 2025.
11. "AI Governance in the Era of Agentic Generative AI and AGI: Frameworks, Risks, and Policy Directions," IJIRCST, vol. 13, 2025.
12. "AI for Climate Action: Assessing Benefits and Risks through Common Standards and Indicators," IEEE Planet Positive 2030, Dec. 2023.