

Privacy-Aware Agentic AI Frameworks for Secure Human-Agent Collaboration

Calcedonio Valmarana

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

valmarana30@uab.edu

Panji Samosir

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.

hellopanji@colostate.edu

Hemangini Buch

Department of Computer Science, University of Houston, Houston, TX, USA.

contacthemangini@uh.edu

Abstract

Agentic artificial intelligence systems are increasingly conceived as autonomous or semi-autonomous actors that plan, remember, negotiate, and execute tasks across organizational boundaries. This shift from passive prediction to situated action creates a distinctive privacy problem: agents must collaborate with humans and other agents while continuously handling sensitive context, preferences, credentials, and inferred intent. This paper develops a system-level framework for privacy-aware agentic AI in secure human-agent collaboration. It argues that privacy cannot be treated as a post hoc filter or a single technical control; rather, it must be designed into the architecture, governance, and operational lifecycle of agentic infrastructures. The paper examines layered agent architectures, secure aggregation, federated learning, differential privacy, contextual integrity, accountability auditing, and policy compliance as interlocking concerns. It analyzes structural trade-offs among utility, latency, autonomy, explainability, robustness, and regulatory alignment. It also considers deployment realities such as edge-cloud coordination, model provenance, intermittent connectivity, energy sustainability, and multi-stakeholder governance. Through cross-domain comparisons in healthcare, finance, public services, and industrial operations, the paper shows that secure human-agent collaboration depends on negotiated disclosure, least privilege, purpose limitation, and auditable delegation. The conclusion proposes that privacy-aware agentic frameworks should be evaluated not only by task success but also by their capacity to preserve contextual norms, distribute accountability, and remain resilient under adversarial and institutional pressure.

Keywords

privacy-aware agentic AI, human-agent collaboration, secure multi-agent systems, governance, federated learning, differential privacy, accountability, socio-technical infrastructure.

1. Introduction

The emergence of agentic artificial intelligence marks a transition from systems that answer queries to systems that pursue goals, invoke tools, coordinate with other agents, and act on

behalf of users or institutions. This transition is not merely an increase in model capability; it changes the socio-technical surface of computing. An agent that can schedule meetings, retrieve records, negotiate with external services, and adapt its plan over time becomes a persistent participant in human workflows. As a result, privacy concerns multiply because the agent must observe, retain, infer, and transmit information that may be intimate, proprietary, regulated, or contextually sensitive. Traditional privacy engineering often assumes bounded data flows and identifiable processing stages, but agentic systems generate diffuse, recursive, and long-lived data trajectories that are difficult to enumerate in advance [8], [7].

The problem is especially acute in human-agent collaboration, where trust is both a precondition and an outcome. Humans must delegate authority to agents without surrendering control over disclosure, identity, or future uses of data. Agents, in turn, must operate under uncertainty about human preferences, legal constraints, and adversarial manipulation. Ethical principles such as beneficence, non-maleficence, autonomy, justice, and explicability provide normative orientation, but they do not specify how an agent should decide whether to share a medical detail with a scheduling service, whether to retain a conversational memory, or whether to reveal uncertainty to a human partner [14], [15]. A privacy-aware framework therefore needs to translate principles into architectural commitments, governance rules, and operational practices that remain coherent across domains.

This paper argues that secure human-agent collaboration requires privacy to be treated as a first-class systems property. That means privacy constraints must shape agent memory, planning, tool use, inter-agent communication, model training, deployment topology, and auditability. The discussion is organized around structural trade-offs rather than isolated mechanisms. It considers how differential privacy, secure aggregation, federated learning, contextual integrity, and accountability auditing interact when agents are embedded in real institutions. It also examines deployment constraints such as edge-cloud partitioning, energy use, intermittent connectivity, and model provenance. The central claim is that privacy-aware agentic AI cannot be achieved by adding a privacy filter to an otherwise autonomous pipeline; it must be designed as a negotiated, auditable, and resilient collaboration infrastructure.

2. Conceptual Foundations of Privacy-Aware Agentic AI

Agentic AI draws on longstanding research in multi-agent systems, where autonomous entities interact through protocols, shared environments, and strategic reasoning. Classical work emphasizes coordination, communication, and delegation, but often assumes relatively stable ontologies and bounded task domains [12], [13]. Contemporary agentic models extend these assumptions by combining large-scale language understanding, tool invocation, memory, and open-ended planning. Recent work on scaling agentic models for full-scenario applications suggests that generality and cross-domain competence intensify the privacy surface because agents operate across contexts and retain traces that can be recombined later [11]. The more capable an agent becomes, the more it may need to know, and the more it knows, the more carefully its knowledge must be governed.

Privacy in this setting is not equivalent to secrecy. Contextual integrity frames privacy as the appropriate flow of information according to contextual norms, roles, purposes, and transmission principles [20]. An agent that shares a user’s location with an emergency service may act appropriately, while sharing the same information with an advertising broker may violate contextual expectations even if both actions are technically authorized. This distinction is crucial for agentic systems because agents often act across multiple contexts in a single workflow. A human-agent collaboration may begin in a healthcare context, move into

insurance verification, and then touch employment or financial services. Each transition changes the relevant norms, and the agent must recognize when a data flow crosses a contextual boundary.

Differential privacy offers a formal approach to limiting inferential disclosure by adding controlled noise to computations or outputs [1]. In agentic systems, differential privacy can protect aggregate learning, telemetry, and preference modeling, but it does not by itself resolve problems of consent, purpose limitation, or contextual appropriateness. An agent may produce differentially private statistics while still violating a user's expectation that a conversation will not be used for model improvement. Similarly, membership inference and related attacks show that even models trained on sensitive data can leak information about individuals through predictions or confidence patterns [2]. Privacy-aware agentic frameworks therefore need both quantitative guarantees and contextual governance. The technical and the normative must be co-designed rather than sequenced.

A further foundation is the distinction between user-facing agents, institutional agents, and infrastructure agents. User-facing agents represent individual interests, institutional agents represent organizational policies, and infrastructure agents manage routing, authentication, billing, or safety. These roles may conflict. An institutional agent may seek to optimize compliance and cost, while a user-facing agent seeks to minimize disclosure and maximize autonomy. Secure collaboration requires explicit negotiation mechanisms that make role, purpose, retention, and revocation visible. Without such mechanisms, privacy becomes an implicit bargaining chip buried in system prompts, terms of service, or opaque model behavior.

3. Architectural Layers and Secure Human-Agent Collaboration

A privacy-aware agentic architecture can be understood as a layered system in which perception, memory, planning, tool use, communication, and governance are separated but mutually constrained. At the perception layer, agents ingest user input, environmental signals, documents, and sensor data. At the memory layer, they store short-term context, long-term preferences, and task histories. At the planning layer, they decompose goals, select tools, and coordinate with other agents. At the action layer, they execute transactions, send messages, and modify external systems. At the governance layer, they enforce policies, log decisions, and support audit. Privacy failures can occur at any layer, but they are most dangerous when data crosses layer boundaries without corresponding changes in authority and accountability.

Secure human-agent collaboration also depends on authentication and authorization that go beyond traditional user login. An agent acting on behalf of a human needs a verifiable delegation chain that specifies what it may do, for how long, under which conditions, and with what evidence. This resembles capability-based security, but it must accommodate probabilistic reasoning and natural-language instructions. A user might say that an agent may share calendar availability with colleagues but not with external vendors. The agent must translate that instruction into enforceable policy and recognize when ambiguity requires clarification rather than assumption. Least privilege is necessary but insufficient if the agent can infer restricted information from legitimately accessible data.

Inter-agent communication introduces additional risks because agents may exchange summaries, embeddings, plans, or intermediate results. Secure aggregation techniques allow multiple parties to compute joint functions without revealing individual inputs to a central server [5]. Federated learning similarly enables model improvement across distributed data

holders without centralizing raw data [4], [6]. These approaches are valuable for agentic infrastructures, but they do not eliminate leakage through gradients, metadata, or repeated interactions. An adversarial agent might infer sensitive attributes by observing timing, query patterns, or model updates. Therefore, privacy-aware architectures must combine cryptographic protection, statistical disclosure control, and interaction-level policy enforcement.

Tool use is another critical boundary. Agents often call external application programming interfaces, databases, search engines, and productivity services. Each tool call may disclose user intent, identity, or context to a third party. A privacy-aware framework should treat tool invocation as a disclosure event that requires purpose binding, data minimization, and logging. It should also support confidential computing or trusted execution environments when sensitive operations must occur on untrusted infrastructure. However, such mechanisms add complexity and cost. The architectural challenge is to make privacy-preserving paths the default for sensitive tasks while allowing efficient paths for low-risk interactions. This requires risk classification that is dynamic, explainable, and subject to appeal.

4. Privacy Mechanisms and Their System-Level Trade-offs

Differential privacy, secure aggregation, federated learning, and encrypted computation are often presented as complementary privacy technologies, but their system-level effects differ substantially. Differential privacy provides a tunable privacy budget, yet the choice of budget is inherently political and domain-specific. A healthcare agent may require strong guarantees for diagnostic data, while an industrial agent may accept weaker guarantees for operational telemetry. The framework must make these choices explicit and auditable rather than burying them in technical defaults. It must also account for composition: repeated interactions across many tasks can cumulatively weaken privacy even if each individual release is protected [1].

Federated learning reduces central data concentration but introduces coordination, heterogeneity, and communication costs [6]. In agentic systems, federated learning could allow personal assistants to improve without uploading private conversations, but the resulting models may still encode sensitive patterns. Secure aggregation can protect individual updates during training, yet it assumes a set of participating clients and a threat model that may not hold in open agent ecosystems [5]. Moreover, federated approaches may disadvantage users with limited connectivity or compute resources, creating fairness concerns. A privacy-aware framework must therefore consider not only whether data are centralized but also who can participate, who benefits, and who bears the operational burden.

Human feedback mechanisms such as reinforcement learning from human preferences and instruction tuning are central to aligning agents with human intent [9], [10]. These methods often rely on collecting demonstrations, comparisons, or corrections, all of which may contain sensitive information. If feedback data are reused across contexts, agents may learn preferences that users intended for a specific relationship or task. Privacy-aware training pipelines should separate personalization from general model improvement, support local adaptation, and provide meaningful opt-out or opt-in mechanisms. They should also document provenance so that later audits can trace how a behavior was learned and whether it depended on protected data.

The trade-offs among privacy, utility, latency, and autonomy are not static. Stronger privacy protections may increase computational overhead, require additional human confirmation, or reduce the agent’s ability to act proactively. In time-critical settings such as emergency

response, excessive confirmation may be harmful. In low-risk settings such as recommending a restaurant, excessive privacy overhead may erode trust and adoption. A mature framework should therefore support graded privacy modes, contextual risk assessment, and reversible delegation. It should also recognize that privacy is relational: what is sensitive depends on the parties involved, the purpose, the setting, and the potential for future inference.

5. Governance, Policy, and Accountability

Governance is the connective tissue of privacy-aware agentic AI. Technical controls cannot anticipate every context, and policies cannot be enforced without architectural support. Accountability requires clear allocation of responsibility among users, developers, deployers, and institutions. Internal algorithmic auditing frameworks emphasize documentation, testing, and ongoing monitoring across the lifecycle [17]. For agentic systems, auditing must extend to planning traces, tool calls, memory updates, and inter-agent messages. It must answer not only what the agent did but also why it believed the action was permissible. This is a demanding standard because agent behavior may emerge from complex interactions rather than explicit rules.

Principles alone cannot guarantee ethical or privacy-respecting outcomes [16]. The global landscape of AI ethics guidelines demonstrates convergence on values such as transparency, fairness, and privacy, but also reveals gaps in enforcement and operationalization [15]. Regulatory instruments such as the European Union’s artificial intelligence framework introduce risk-based obligations, documentation duties, and prohibitions for certain practices [18]. Standards bodies and international coordination efforts can help translate legal requirements into technical specifications, but they must remain adaptable as agentic capabilities evolve [19]. A privacy-aware framework should therefore map legal categories to system properties, such as purpose limitation, data minimization, retention limits, and human oversight.

Accountability also requires mechanisms for redress. If an agent discloses information inappropriately, the affected person needs a path to understand what happened, who was responsible, and how to prevent recurrence. This is difficult in multi-agent systems where causation is distributed. Logging every internal state may itself create privacy risks, so auditability must be balanced against minimization. Techniques such as selective disclosure, zero-knowledge attestations, and privacy-preserving audit logs can help, but they require governance to define what evidence is necessary and who may access it. Without such safeguards, accountability infrastructures can become surveillance infrastructures.

Policy implications extend beyond compliance. Governments and organizations must decide whether agents may negotiate on behalf of citizens, how public services should authenticate delegated authority, and how competition policy applies when dominant platforms control agent ecosystems. Privacy-aware agentic AI is thus a matter of infrastructure governance, not only application design. It intersects with antitrust, consumer protection, labor, healthcare regulation, and national security. A narrow technical framing risks missing these structural dimensions.

6. Deployment, Infrastructure, and Sustainability

Deployment topology shapes privacy outcomes. Cloud-hosted agents can leverage large-scale models and centralized monitoring, but they concentrate data and create jurisdictional complexity. Edge-hosted agents can keep sensitive data local and operate under intermittent connectivity, but they face resource constraints and may be harder to update. Hybrid

architectures can partition tasks so that sensitive reasoning occurs locally while non-sensitive coordination occurs in the cloud. This partitioning must be dynamic because task sensitivity can change as context accumulates. A local calendar lookup may become sensitive when combined with location and health data, so the framework must continuously reassess disclosure risk.

Infrastructure sustainability is also relevant. Privacy-preserving computation often increases energy consumption through encryption, noise generation, redundant computation, or secure hardware. Federated learning can reduce data transfer but may increase client-side computation and coordination overhead [4]. If privacy protections are deployed inefficiently, they may not be sustainable at scale. A privacy-aware agentic framework should therefore consider carbon intensity, device battery life, network bandwidth, and lifecycle costs. Sustainability and privacy are not identical goals, but both require avoiding unnecessary data movement and computation. Designing for data minimization can support both.

Model provenance and supply chain security are additional deployment concerns. Agents may use foundation models, fine-tuned adapters, third-party tools, and external knowledge bases. Each component can introduce privacy risks or malicious behavior. A compromised tool could exfiltrate context, and a poisoned model could leak information through subtle output patterns. Provenance metadata, signed updates, and runtime attestation can help, but they must not become another vector for tracking users. The challenge is to establish trust in components without building a centralized identity graph that undermines privacy.

Operational resilience matters because privacy failures often occur during exceptions. When systems fail, engineers may disable safeguards to restore service. When users are frustrated, they may grant broad permissions. When agents encounter ambiguity, they may default to sharing more information to complete a task. A robust framework should therefore include fail-safe privacy modes, graceful degradation, and clear escalation paths. It should assume that humans will sometimes make mistakes and that adversaries will exploit those mistakes. Resilience is not only about uptime; it is about preserving rights under stress.

7. Robustness, Fairness, and Failure Modes

Privacy-aware agentic AI must be robust against adversarial inference, manipulation, and collusion. Membership inference attacks demonstrate that models can reveal whether specific data were used in training [2]. In agentic settings, similar attacks may target memory, tool logs, or preference profiles. An adversary might probe an agent with carefully crafted requests to reconstruct private context or infer relationships. Defenses must operate at multiple levels: model training, inference, interaction policy, and monitoring. No single mechanism is sufficient because attackers can combine information from multiple sources.

Fairness concerns arise because privacy protections are not distributed evenly. Users with less technical literacy may accept default settings that disclose more data. Users in under-resourced regions may have devices that cannot run local privacy-preserving computation. Institutional agents may prioritize efficiency for powerful clients while imposing stricter verification on marginalized groups. A privacy-aware framework should therefore include equity assessments, accessible consent interfaces, and independent oversight. It should also examine how privacy harms compound with other forms of exclusion, such as credit denial, employment screening, or immigration status.

Failure modes in human-agent collaboration include over-disclosure, under-disclosure, misplaced trust, and automation bias. Over-disclosure occurs when an agent shares more than

necessary due to ambiguous objectives or poor policy enforcement. Under-disclosure occurs when an agent withholds information necessary for safety or informed consent. Misplaced trust occurs when humans treat fluent agent output as authoritative. Automation bias occurs when humans defer to agent recommendations without scrutiny. These failures are not purely technical; they reflect interface design, organizational incentives, and training. Privacy-aware frameworks must therefore integrate human factors research and participatory design.

Security and privacy are related but distinct. A system can be secure against unauthorized access yet still violate privacy through legitimate but inappropriate data flows. Conversely, a privacy-preserving system may be vulnerable to denial of service or integrity attacks. Agentic systems require both confidentiality and integrity, because a manipulated agent may disclose private data or act against the user's interests. Robustness against adversarial inputs, prompt injection, and tool poisoning is thus part of privacy engineering. The framework should treat privacy as a property that depends on the entire socio-technical system, including human norms, institutional policies, and adversarial incentives.

8. Case Illustrations and Cross-Domain Comparisons

Healthcare provides a demanding case for privacy-aware agentic collaboration. A clinical agent might coordinate appointments, summarize patient history, and communicate with insurance providers. It must comply with legal frameworks, protect highly sensitive data, and support clinical judgment rather than replace it. Contextual integrity is essential because information shared within a care team may not be appropriate for insurers, employers, or researchers [20]. Differential privacy may be useful for population health analytics, but individual care requires precise information and strong access control [1]. The agent must distinguish between therapeutic purpose, billing purpose, and public health purpose, and it must enforce those distinctions in real time.

Financial services illustrate a different balance. Fraud detection agents must share information across institutions to identify coordinated attacks, but sharing too much can expose customer behavior and create systemic risks. Federated learning and secure aggregation can support collaborative fraud detection without centralizing raw transactions [4], [5]. However, financial agents also make consequential decisions about credit, pricing, and access. Fairness and explainability requirements are therefore tightly coupled with privacy. A privacy-aware framework must ensure that anti-fraud measures do not become pretexts for discriminatory profiling or unaccountable scoring.

Public services introduce questions of citizenship, equity, and due process. A government agent might help residents apply for benefits, verify eligibility, and coordinate across agencies. The efficiency gains are substantial, but the risks include function creep, exclusion by algorithmic error, and surveillance. Regulatory frameworks such as the European Union's risk-based approach offer guidance, but public-sector deployment requires stronger democratic oversight [18]. Standards can help with interoperability, yet they cannot replace political deliberation about what data government agents should collect and retain [19]. Privacy-aware design in this domain must include appeal mechanisms, human review, and transparency about delegated authority.

Industrial operations show how privacy can intersect with safety and intellectual property. An industrial agent might coordinate maintenance, optimize supply chains, and share telemetry with equipment vendors. The sensitive data may be commercial rather than personal, but employees' behavior and performance may also be inferred. Cross-domain comparison

reveals that privacy is not only about individual identity; it also concerns organizational autonomy, competitive advantage, and workplace power. A general framework must therefore support multiple privacy subjects, including individuals, groups, and institutions.

9. Forward-Looking Perspectives

Future agentic systems will likely operate in federated ecosystems where no single provider controls all capabilities. Interoperability standards will be necessary, but they must preserve privacy and competition. If a few platforms control identity, memory, and tool access, they may become privacy chokepoints. Conversely, fragmentation can create inconsistent protections and make accountability difficult. A pluralistic ecosystem requires portable privacy preferences, verifiable credentials, and open audit interfaces. It also requires governance mechanisms that can evolve as agents become more capable.

Another forward-looking direction is the development of privacy-aware agent constitutions. These are explicit, machine-readable specifications of what an agent may do, what it must not do, and how conflicts should be resolved. Such constitutions could combine legal rules, organizational policies, and user preferences. They could be tested through simulation, monitored in deployment, and revised through participatory processes. The required reference on scaling agentic models for full-scenario applications indicates that broad capability expansion is already underway, which makes constitutional constraints more urgent rather than less [11]. Capability without constraint increases the risk of contextual violations.

Human oversight will remain essential, but it must be designed realistically. Humans cannot review every agent action, and excessive oversight can create alert fatigue. Effective oversight should be risk-based, sampling high-impact decisions, supporting post hoc investigation, and giving users meaningful control over delegation. Explainability should focus on actionable reasons: what data were used, what policy applied, what alternatives were considered, and what can be changed. This kind of explanation is different from revealing internal model weights or chain-of-thought. It is a governance interface that connects technical behavior to accountability.

Finally, privacy-aware agentic AI should be evaluated as infrastructure. Metrics should include not only task completion and accuracy but also disclosure minimization, contextual fit, revocation success, audit completeness, fairness across groups, energy cost, and resilience under attack. Regulation and standards can set baselines, but continuous evaluation is necessary because agent ecosystems change rapidly [17], [18], [19]. The goal is not a perfectly private agent, which may be impossible in a connected world, but a system that makes privacy violations rarer, more visible, and more correctable. That is a realistic and socially necessary ambition.

10. Conclusion

This paper has argued that privacy-aware agentic AI frameworks must treat privacy as a structural property of secure human-agent collaboration. Agentic systems differ from conventional AI because they plan, remember, act, and coordinate across contexts. These capabilities create persistent and recursive data flows that cannot be governed by a single filter or consent screen. The paper examined layered architectures, delegation, inter-agent communication, tool use, differential privacy, secure aggregation, federated learning, contextual integrity, accountability, deployment, sustainability, robustness, fairness, and policy. Across these dimensions, the same lesson recurs: privacy depends on the fit between technical controls and contextual norms.

Secure human-agent collaboration requires mechanisms for negotiated disclosure, purpose limitation, least privilege, revocation, and audit. It also requires governance that allocates responsibility and provides redress. Technical guarantees such as differential privacy and secure aggregation are valuable, but they must be embedded in institutional processes and human oversight. Otherwise, they may provide formal protection while leaving substantive privacy harms unaddressed. The framework proposed here emphasizes co-design among architecture, policy, and operational practice.

The future of agentic AI will be shaped by choices about infrastructure, standards, competition, and public accountability. If privacy is treated as an afterthought, agentic systems may normalize pervasive inference and opaque delegation. If privacy is designed into the foundations, agentic systems can support collaboration while preserving human autonomy and contextual integrity. The challenge is both technical and political. It requires building systems that are capable, auditable, fair, and resilient, and it requires institutions willing to govern them accordingly.

References

1. Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming (ICALP 2006)*, Lecture Notes in Computer Science, Vol. 4052, pp. 1–12. Springer.
2. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pp. 3–18. IEEE.
3. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy*, pp. 399–414. IEEE.
4. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
5. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191. ACM.
6. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60.
7. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
8. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
9. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, pp. 4299–4307.
10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pp. 27730–27744.

11. Jiang, H., Xin, G., Huang, Y., Zhu, G., Zhang, J., Lin, X., ... & Cao, Z. (2026). AgentOmnia: Scaling Agentic Models for Full-Scenario Applications. arXiv preprint arXiv:2607.23124.
12. Wooldridge, M. (2009). An introduction to multiagent systems (2nd ed.). Wiley.
13. Jennings, N. R., & Wooldridge, M. J. (1998). Applications of intelligent agents. In *Agent technology: Foundations, applications, and markets*, pp. 3–28. Springer.
14. Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
15. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
16. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.
17. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. ACM.
18. Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4), 97–112.
19. Cihon, P. (2019). Standards for AI governance: International standards to enable global coordination in AI research and development. Oxford Internet Institute.
20. Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1), 119–157.