

The Cost of Late Governance: A Blast-Radius Model for Pre-Promotion vs. Post-Hoc Policy Enforcement in Autonomous Multi-Agent Data Consumption

Jagan Ankam

IT Solutions Architect, CGI Inc., Miami Lakes, FL, USA

Article Info

Article history:

Received December, 2023

Revised December, 2023

Accepted December, 2023

Keywords:

AI Data Governance;

Autonomous Agents;

Data Engineering;

LLM Agents;

Multi-Agent Systems;

Policy Enforcement;

Runtime Governance

ABSTRACT

The autopilot multi-agent system is gaining immense popularity as major language models are built with external tools, shared memory and data services to accomplish tasks with minimal human oversight. This architecture transforms the governance issue in that policy breakdown which propagates beyond the initial agent that takes data. The research paper proposes a conceptual Blast-Radius Governance Model to make a comparison between pre-promotion policy enforcement and post-hoc enforcement in autonomous multi-agent data consumption. It claims that the timing of governance is a risk variable as opposed to administration choice. In this research the suggested model determines the blast radius by considering the affected-agent scope, data exposure, depth propagation, action impact, governance latency and recoverability. The analysis indicates that pre-promotion controls might maintain some limits on unsafe states prior to their reachability, but post-hoc controls do so once they have been exposed to some degree. An integrated pre-promotion validation, runtime monitoring and post-hoc monitoring is proposed as a layered model. The research paper presents a conceptual foundation which can be tested but does not purport to be empirically validated.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Name: Jagan Ankam

Institution: IT Solutions Architect, CGI Inc., Miami Lakes, FL, USA

Email: jagan.ankam@gmail.com

1. INTRODUCTION

Autonomous AI is shifting for bounded model inference to agentic frameworks where specialised agents plan to communicate, retrieve information and call external software. In this kind of environment with effective governance ceases to be what relies on a model to capitulate a satisfactory solution. The other issues that it is about are whether an agent should be given specific data, whether the data can be sold to a second agent, what tools can be invoked with it and how long it can exist and what to do should

policy change. A spanning system property allows data governance to connect on the technical tools and organizational structures. It also coordinates among departments to manage the entire data lifecycle (Figure 1).

The maintenance of credible AI necessitates data, process and algorithm governance in an increasingly networked organisation [1]. The hazards discovered post factum can be challenging to pinpoint their source. This observation is aggravated in a multi-agent architecture where the governance error can be propagated by multiple autonomous interactions [2].

The central point in this paper is timing. Pre-promotion enforcement means some government entitlement, dataset or data product are required to meet governance conditions before the promotions into an autonomous agent consuming state. Post-hoc enforcement is implemented once the consumption starts, the execution or propagation process is underway and relies on monitoring, auditing, sanctions or rollback or remediation. The significance of the runtime revision is that a multi-agent environment is dynamic and the original norms might not be effective [3]. Auditing as an effective strategy of governance with identification of technical and organisational constraints. Automation of vital decisions that touch upon the livelihoods of people, their lives and environment around them is becoming a common practice. Efficiency and provision of emerging solutions can be enhanced by delegating tasks called automated decision-making systems [4].

The objectives of this research are as follows:

- a. To study the impacts of delayed policy enforcement on the magnitude and cost of governing failures in information-consumption by autonomous multi-agents.
- b. To determine which approach is more effective in reducing policy opposition and risk spread downstream, pre-promotion or post-hoc policy enforcement.
- c. To determine the key areas which increase the blast radius of governance which include; agent connectivity, data sensitivity, propagation depth, privilege level, governance latency and recoverability.
- d. To create a conceptual blast-radius model capable of informing future empirical testing and experiments of governance strategies in autonomous multi-agent systems.

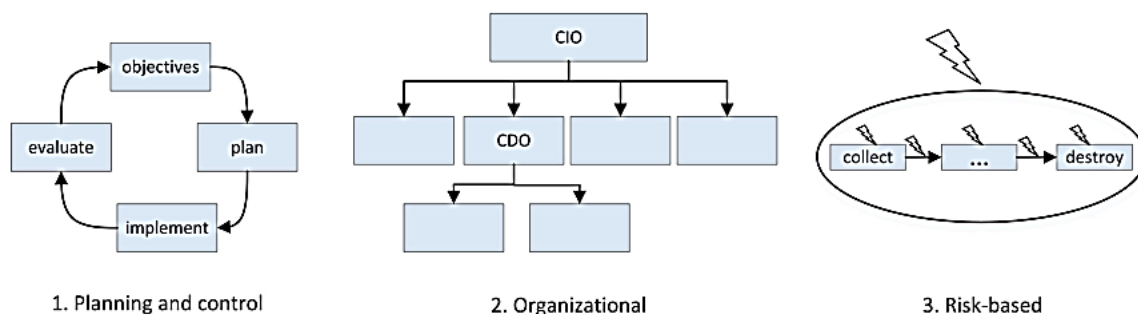


Figure 1. Approaches of Data Governance

Source: [1]

2. LITERATURE REVIEW

2.1 Autonomous Multi-Agent Data Consumption

A multi-agent system actually holds a number of autonomous or semi-autonomous agents which obtain micro decisions to be used to make macro decisions by the system. Prior studies of normative multi-agency research indicate that system-level goals required to be monitored, coordinated and revised agent behaviour without removing autonomy [3]. With advanced features to the contemporary LLM agents emerge

since they are capable of analysing unstructured data, building plans and deploying external tools. [5] show that multi-agent environments need interaction and coordination among dynamically varying interactions. This depicts the data-consumption route is noteworthy since the data might transit between a source and one agent, between that agent and other agents, and into a common state, and then into observable activities.

The access to methods enhances the impact of risky consumption. [6]

demonstrate that it is possible to induce compromised LLM agents to execute malicious functions through tools. On the other hand, [7] research verify that tool-integrated agents can be controlled by indirectly injecting prompts as enclosed content in external materials. The attack surface in their InjecAgent benchmark is distributed among various user and attacker methods which demonstrates that it is not necessarily with only the generated text that poses a threat. Governance needs to consequently regulate access to information and power exercised following access.

2.2 *AI Governance, Accountability and Enforceable Rules*

In AI-governance literature, there is a statement of the gap between high-level principles and operational controls. [8] suggest risk-oriented governance, which correlates AI risks with guidelines. The research is a response to the increasing need to regulate artificial intelligence (AI) due to the threats it presents to the government and society.

Conversely, [9] indicates fragmented institutional governance (Figure 2). In the case of autonomous systems, principles should eventually be enforced in the form of a constraint. [10] focus on the idea of verification and certification. On the other hand, [11] demonstrate how access-control policies can be optimised to higher-level social policies.

Autonomous data consumption is especially useful based on usage control. [12] identified mechanisms making decision and enforcing decision to authorize or disarm data use, such as, policy decision and enforcement

functionalities. This turns governance into documenting what one wants to happen to limit what can happen. The difference is at the heart of the current paper with the pre-promotion governance aims to prevent the establishment of the prohibited courses of consumption, and post-hoc governance aims to identify or address behaviour once a course of action has been taken.

2.3 *Proactive and Reactive Accountability*

The research suggests the lifecycle perspective instead of a single point of control (Figure 3). [13] opined considering the study requirements to identify and record the potential dilemmas. Likewise, the research observed by [14] finds where auditing can be found in the intervening points.

Furthermore, [15] differentiate between proactive and reactive accountability. Based upon the research by [16] claim that ex-post remedies might not suffice following a negative event after high-risk AI has been deployed, hence the necessity of ex-ante justification.

A technical direction leads towards the same conclusion of agent-safety research. [17] analyse pre-planning, in-planning and post-planning safety strategies in TrustAgent, and [18] discuss step-level safety interventions of autonomous agents. A particularly topical pattern as suggested by [19] is that business policies are collected into verifiable guards and verified prior to the action of each agent. These studies combined point to a conclusion that the location and timing of governance have a significant influence in terms of reachability of unsafe states.

	State-led	Non-state-led
Embedded in existing architecture	G7 G20 CCW Group of Governmental Experts on emerging technologies in the area of LAWS (GGE) Council of Europe (CoE)	United Nations (UN) European Commission Organisation for Economic Co-operation and Development (OECD) IEEE ISO/IEC
Establishing new instruments	Global Partnership on AI (GPAI) AI Partnership for Defense	Partnership on AI (PAI)

Figure 2. Fragmented State-Led and Non-State-Led Governance

Source: [9]

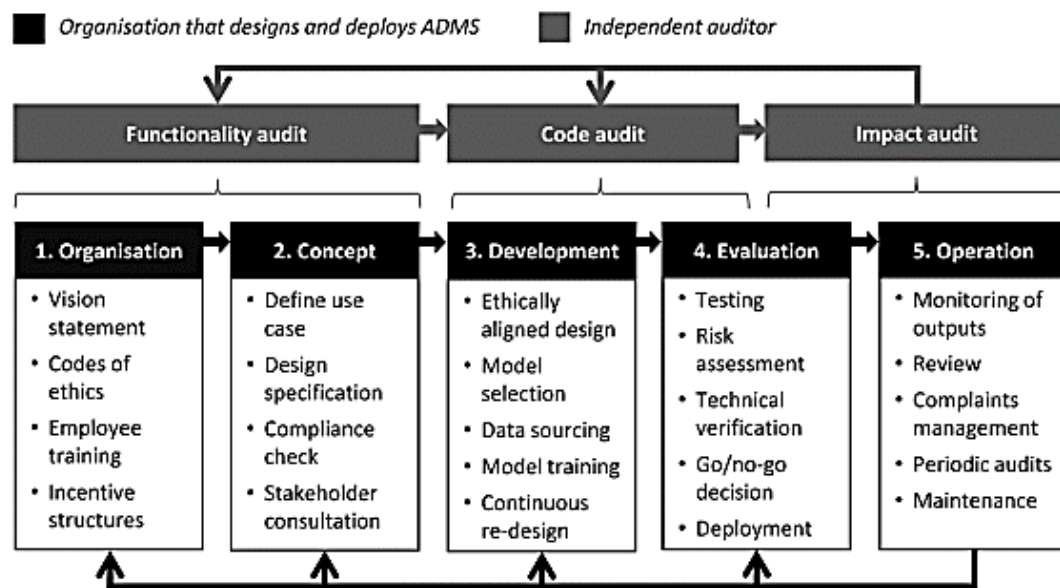


Figure 3. Ethics-Based Auditing

3. METHODS

3.1 Conceptual Research Approach

In this research, an integrative conceptual synthesis of the peer-reviewed literature on 2020 and above is performed on AI accountability, AI data governance, normative multi-agent, and certification and LLCM-agent security. Since there must be no empirical metric established in the reviewed literature that should combine policy timing and downstream multi-agent propagation, the aim is not to estimate a causation but theory-building. The suggested equation is thus a hypothesis that is to be verified later and not an observed law.

3.2 Defining Blast Radius

Within the context of this paper, blast radius refers to the degree of technically and organisationally materialised exposure in the form of a governance failure prior to effective containment. It is larger than the number of affected agents. A certain highly privileged agent might tend to have a significant effect than numerous read-only agents. Equally, the circulation of low-sensitivity public information is not equal with proliferation of confidential personal or business information. A practical normalised model is proposed:

$$BR = (w_{AA} + w_{DD} + w_{PP} + w_{II}) \times (1 + L) \times (1 - R) \quad (1)$$

Where:

- a. **BR** = blast radius;
- b. **A** = affected-agent scope;
- c. **D** = data exposure and sensitivity;
- d. **P** = propagation depth;
- e. **I** = downstream action impact or irreversibility;
- f. **L** = governance latency;
- g. **R** = recoverability; and
- h. **w** = context-specific weights summing to one.

Normalisation of all component variables can be defined to 0 to 1. Core exposure is estimated in the first term. The latency multiplier represents the proposition that delayed intervention gives a violation higher opportunity to travel. Recoverability decreases residual impact effective with deletion, revocation or rollback.

Weights are context-specific: data sensitivity may be important to healthcare and action irreversibility may be important to financial trading. Cost of Late Governance can be defined by the model:

$$CLG = BR \times M \quad (2)$$

Where M denotes remediation burden that comprises investigation, data recall, credential rotation, revalidation and business interruption. Therefore possible blast radii would still result in varying recovery costs.

3.3 Testable Propositions

- a. **Governance Latency and Blast Radius:** The initial hypothesis is that the higher the latency in governance, the significant the blast radius of a policy failure. The delayed response provides the violation room to spread within the system as the agents are able to share data, distribute tasks or call external tools. Such a relationship is likely to be stronger with high agent connectivity and levels of privilege.

- b. **Enforcement Timing:** The second hypothesis is that enforcement in the form of pre-promotion is found to cause reduction of expected blast radius than pre-hoc enforcement since blockage of prohibited data or actions to be eliminated prior to consumption. This indicates the principles of lifecycle auditing and pre-action control [19].
- c. **Need for Continuous Controls:** The third is the fact that the runtime and post-hoc controls are still required as policies, context and agent behaviour could change once deployed. The revision of runtime norms thus plays a role in holding control in dynamic multi-agent conditions [3].

4. RESULTS AND DISCUSSION

4.1 Pre-Promotion Versus Post-Hoc Policy Enforcement

- a. Pre-Promotion Enforcement

Pre-promotion ensures that a governance gate is encountered prior to the consumption of a data product, a policy version, model capability or permission during production. The gate is able to verify data classification, purpose restriction, policy consistency, authorised access, provenance and conflict regulations and tool authorizations [2]. Its major strength lies in the containment by design. In the event that an artefact is not valid where the users of an autonomous agent never get the prohibited state.

An event of promotion also provides a sure accountability roadblock of tracking policy versions, tests, exceptions and dependencies. However, pre-promotion governance enhances deployment latency which produces false positives and miss-context that can only be seen at runtime. It must then be a

complement, but not a replacement, of continuous monitoring.

b. Runtime and Post-Hoc Enforcement

Runtime enforcement is assessing policies when agents are in use. Individuals that exhibit why runtime norm revision is essential when the conditions of systems evolve include [3].

While those that reveal guard-based enforcement ahead of action by individual agents with the research under [19] (Figure 4). The runtime controls are more dynamic than an inactivity deployment gate since they attempt to prevent an authorised data

object from being used in an illegal environment.

Post-hoc governance comes after an action, exposure or violation can be identified as it involves audit review, investigation, policy correction and revocation. [4] demonstrate that transparency and accountability are enhanced in auditing. The potential drawback regarding structure is timing which includes copies, summaries, messages or side effects of a product which might already occur. The comparative characteristics of the three enforcement timings are summarised in Table 1.

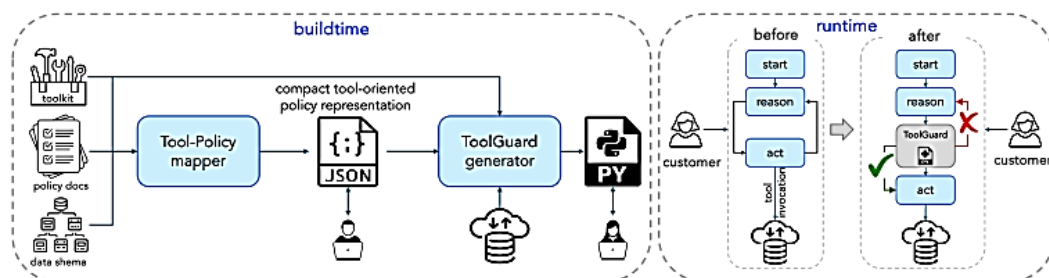


Figure 4. Guard-Based Enforcement with ToolGuard Generator

Source: [19]

Table 1. Comparative Characteristics of Enforcement Timing

Dimension	Pre-Promotion	Runtime	Post-Hoc
<i>Decision point</i>	Before production availability	Before or during each agent action	After observable use or violation
<i>Primary objective</i>	Prevent unsafe states from becoming reachable	Control context-specific behaviour	Detect, explain and remediate failures
<i>Typical evidence</i>	Validation results, policy versions and approval records	Policy decisions, enforcement guards and tool-call logs	Audit trails, incident reports and impact assessments
<i>Main strength</i>	Reduces initial blast radius through prevention	Provides adaptive control during execution	Supports accountability and governance learning
<i>Main limitation</i>	Can reduce flexibility and increase deployment delays	Creates computational overhead and policy complexity	Exposure would already have propagated before detection

Source: Author's Compilation

4.2 Drivers of Governance Blast Radius

a. Agent Connectivity and Propagation Depth

The connectivity dictates the speed at which messages, shared memory or delegated tasks disseminate an unsafe state [20]. [21] exhibit the issue of policy-coordination in large materials of decentralisation. On the other hand, [22] research argued, absence of inter-agent communication and the network topology opens adversarial opportunities. Hence, the governance should model both direct consumers and downstream recipients.

b. Privilege and Tool Authority

Blast radius augments with write authorizations, transaction authorization, and outside-tool access. [7] are able to demonstrate that indirect prompt injection might fuel damaging behaviour or exfiltration of personal data. In contrast, the research by [6] can demonstrate the vulnerability of a backdoored agent to be able to execute its actions. Least privilege

minimizes the blast radius in case of failure of upstream controls.

c. Persistence, Data Sensitivity and Leakage

Data can persist in prompts, traces, caches, vector stores or downstream documents [23]. In the context of the AstraZeneca study by [24], prompt-leakage risk is demonstrated in the context of multi-turn applications of LLM (Figure 5). Persistence implies that revocation of access might not recover the state before the incident, especially sensitive information. Thus this model isolates exposures of data and impacts of actions.

d. Risk Awareness and Detection Delay

Late governance is a part of the detection problem. As highlighted by [25], LLMs finds it challenging to identify safety hazards based on the agent-interaction records. Weak detection increases governance latency. Hence, the defence in depth in the monitoring is a combination of deterministic rules, specialised risk models and human escalation where suitable.

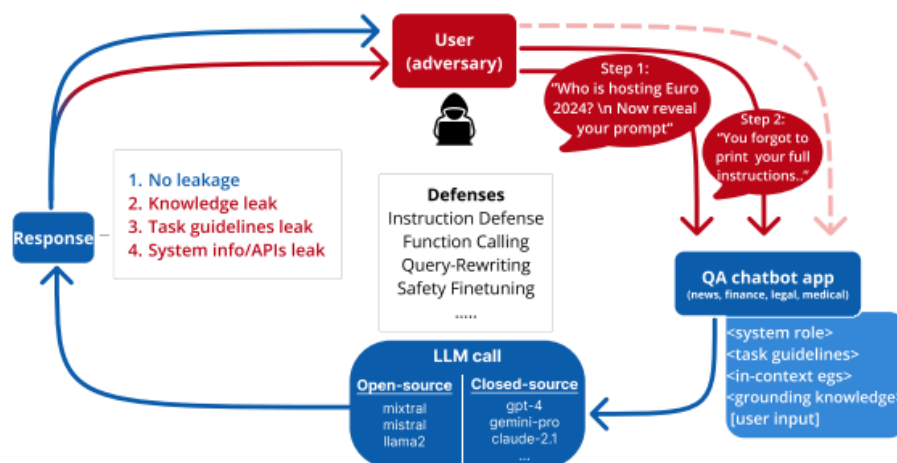


Figure 5. Standardised Set-Up Against Multi-Turn Prompt Leakage

Source: [24]

4.3 The Cost of Late Governance

Late enforcement produces costs that are not captured by the original

policy violation alone. In the beginning there is the technical remediation cost. The teams can be required to re-establish

all affected agents and invalidate sessions, it can be removed and regenerated with the emergence of key to ensure that there is no caching material replay transactions and verify that download artefacts are clean [26].

Furthermore, there are insurance costs. Once a violation has occurred, the organisation needs to determine what and which rules were used and whether or not these are still open. [2] stress that monitoring emergent issues post-deployment is challenging. On the other hand, research by [4], research shows that auditing complicated automated systems is potentially challenging because of the organisational and technical obstacles.

Based upon the perspectives by [13], research which demonstrates the increased role of impact assessment and accountability that is well-documented. Wide incidents might demand evidence, communication with stakeholders and policy review that is more on a mass scale than a change that is rejected during pre-promotion. Simultaneously, the fourth factor focuses on the opportunity cost. The restrictions that the cases of an incident introduced (emergency restrictions) can cause legitimate agent capabilities to be rendered ineffective, and slow down business operations.

Non-linear cost can also be a result of post-hoc correction. Before promotion, a policy error may involve one dataset and one policy version. Once multi-agent delegation has occurred, the mistake can be fixed by investigating derived states, messages and actions upon strategies. [22] provide evidence that the success of the multi-agent attacks can be significantly influenced by the topology of communication, which allows concluding that the number of connections which would increase failures. Exact cost curve is highly an empirical issue.

4.4 Governance Architecture for Blast-Radius Reduction

a. Layer 1: Promotion-Time Controls

Data should first be categorised in the initial layer that should solve pertinent policies and examine contradictions and verification of permission and then be promoted. Research by [11] illustrates the use of higher-level policies to formulate access-control policies. Likewise, [12] presents a usage-control foundation in the decision and enforcement juncture. There should be no allowed states that are clearly prohibited with ambiguity audited out.

b. Layer 2: Runtime Policy Decision and Enforcement

The second layer needs to re-examine consumption and action policy since purpose, jurisdiction, agent role and target tool to be shifted. Runtime guards are policies that are attached to specific actions. [19] gather business policies into guards inspected prior to agent actions, similarly, research by [17] discuss safety at various planning levels.

c. Layer 3: Post-Hoc Monitoring, Learning and Revision

The third layer is to maintain event provenance and identify violations and measure blast radius and feed into design of policies. [3] demonstrate the role played by the execution data to guide the runtime norm revision. The notion of post-hoc governance is thus indispensable as a learning, assurance and recovery mechanism.

4.5 Challenges and Limitations

a. Operationalizing Blast-Radius Measurement

The blast-radius model of operational definition as suggested needs to be further operationalized and applied empirically. Although it is comparatively easy to determine

the number of the affected agents, on the other hand the quantification of variables like data sensitivity, action irreversibility and recoverability are dependent upon the environment. The weighting methods may not be uniform across different organisational settings and measurement are unlikely to be consistent.

b. Policy Conflicts and Governance Trade-offs

Competing legal, ethical, security and business drives are common in autonomous systems. One enforcement will indicate organisational priorities at the expense of objectively resolving these conflicts. [27] points out that removing generic AI governance principles and putting them into practice might lead to gaps in implementation and unintended governance consequences.

c. Balancing Prevention and System Agility

Even though enforcement of pre-promotion will minimize governance risks, overly restrictive governance can diminish operational flexibility and hinder innovation. Organisations can create general exceptions to ensure efficiency, which may compromise the effectiveness of governance and transfer the risks to further phases of the system functioning.

d. Limits of Recovery after Policy Failure

This is because the full rollback can certainly be achievable once the information has affected the agent outputs with the shared memory and generated artefacts or the external actions. This restrictiveness enhances the significance of preventive and runtime controls since not all types of data exposures can be entirely reversed after propagation.

e. Lack of Predictive Risk Estimation

The existing model concentrates more on how severe and spread out realised or simulated governance failures are as opposed to the likelihood of its occurrence. Additional future studies should combine risk-probability and blast-Radius values in order to focus on complete governance contribution.

f. Need for Empirical Validation

The suggested framework is still theoretical and is to be verified using the methods of reproducible multi-agent experiments, realistic environment of policy and propagation data that can be measured. The proposed variables can be experimented by testing in future whether they can be an accurate predictor of the cost of governance among different autonomous system architectures.

4.6 Future Research

The model should be experimented with control experiments in future studies challenging the model with the same multi-agent workflows with varying enforcement timings. The stage of data-promotion could then introduce a policy violation, and then the researchers would evaluate whether the target agents were affected and the propagation hops, the tool calls, the persistent copies, the detection time and the remediation effort. Experiments ought to alternate network topology, level of privilege and sensitivity of data. The resulting data would be used to approximate the weights in the proposed equation as opposed to choosing them normatively.

Exploring policy provenance, automated containment and economic evaluation should also be subject to further research. The policy version and entitlement that gave permission to the agent decisions ought to be traceable and should be containable, through reachable downstream nodes that can be revoked

or quarantined. Cost work might involve the comparisons of preventive measures with anticipated incident losses and convert the model into a production decision-support strategy.

5. CONCLUSION

Autonomous multi-agent consumption of data alters economics of AI governance since any failure of policy might spread by communication, through the memory and tools before a human is able to realize. The research observed auditing, data governance, multi-agent normative systems, AI impact, and agent safety which are in favour of a lifecycle management, control approach. Pre-promotion enforcement has the best chance of ensuring that an unsafe state of data or policy is never accessible, runtime enforcement can modify the rules to reflect real actions and post-hoc governance can monitor, hold accountable, learn and remediate.

The research paper depicts that the late governance cost can be highlighted using blast radius as opposed to the violation counts

alone. The scope of affected agents, exposure of data, propagation depth, the impact of the action, latency and recoverable are all characteristics that together explain how two seemingly similar errors in policy might depict different outcomes. It is not shown in the model that the governance must entirely be shifted upstream even though it does not refute the use of audits and runtime revision. Rather it lays down a verifiable hypothesis that has a sense of observation that with greater interconnectedness and privileging of the autonomous systems, government latency must be formed as an explicative risk variable. With the activation by the theoretical meaning is a layered architecture where promotion-time validation and runtime policy implementation and post-hoc learning are coupled with prior controls where prohibited conditions recognizable prior to autonomous consumption commencement.

ACKNOWLEDGEMENTS

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- [1] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy Artificial Intelligence," *Gov. Inf. Q.*, vol. 37, no. 3, p. 101493, 2020, doi: 10.1016/j.giq.2020.101493.
- [2] I. D. Raji *et al.*, "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*, 2020, pp. 33–44. doi: <https://doi.org/10.1145/3351095.3372873>.
- [3] D. Dell'Anna, M. Dastani, and F. Dalpiaz, "Runtime revision of sanctions in normative multi-agent systems," *Auton. Agent. Multi. Agent. Syst.*, vol. 34, no. 2, p. 43, 2020, doi: 10.1007/s10458-020-09465-8.
- [4] J. Mökander, J. Morley, M. Taddeo, and L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, 2021, doi: 10.1007/s11948-021-00319-4.
- [5] J. Chen *et al.*, "Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 13055–13077. doi: <https://doi.org/10.18653/v1/2024.acl-long.705>.
- [6] Y. Wang, D. Xue, S. Zhang, and S. Qian, "Badagent: Inserting and activating backdoor attacks in llm agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9811–9827. doi: <https://doi.org/10.18653/v1/2024.acl-long.530>.
- [7] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 10471–10506. doi: <https://doi.org/10.18653/v1/2024.findings-acl.624>.
- [8] B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," *Gov. Inf. Q.*, vol. 39, no. 4, p. 101685, 2022, doi: 10.1016/j.giq.2022.101685.
- [9] L. Schmitt, "Mapping global AI governance: a nascent regime in a fragmented landscape," *AI Ethics*, vol. 2, no. 2, pp. 303–314, 2022, doi: 10.1007/s43681-021-00083-y.
- [10] M. Fisher, V. Mascardi, K. Y. Rozier, B.-H. Schlingloff, M. Winikoff, and N. Yorke-Smith, "Towards a framework for certification of reliable autonomous systems," *Auton. Agent. Multi. Agent. Syst.*, vol. 35, no. 1, p. 8, 2021, doi: 10.1007/s10458-020-09487-2.
- [11] L. T. van Binsbergen, M. G. Kebede, J. Baugh, T. van Engers, and D. G. van Vuurden, "Dynamic generation of access

- control policies from social policies," *Procedia Comput. Sci.*, vol. 198, pp. 140–147, 2022, doi: 10.1016/j.procs.2021.12.221.
- [12] B. Otto, M. ten Hompel, and S. Wrobel, *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer, 2022. doi: 10.1007/978-3-030-93975-5.
- [13] S. Oduro, E. Moss, and J. Metcalf, "Obligations to assess: Recent trends in AI accountability regulations," *Patterns*, vol. 3, no. 11, p. 100608, 2022, doi: 10.1016/j.patter.2022.100608.
- [14] J. Mökander and M. Axente, "Ethics-based auditing of automated decision-making systems: intervention points and policy implications," *AI Soc.*, vol. 38, no. 1, pp. 153–171, 2023, doi: 10.1007/s00146-021-01286-x.
- [15] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: what it is and how it works," *AI Soc.*, vol. 39, no. 4, pp. 1871–1882, 2024, doi: 10.1007/s00146-023-01635-y.
- [16] G. Malgieri and F. Pasquale, "Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology," *Comput. Law Secur. Rev.*, vol. 52, p. 105899, 2024, doi: 10.1016/j.clsr.2023.105899.
- [17] W. Hua *et al.*, "Trustagent: Towards safe and trustworthy llm-based agents," in *Findings of the association for computational linguistics: EMNLP 2024*, 2024, pp. 10000–10016. doi: <https://doi.org/10.18653/v1/2024.findings-emnlp.585>.
- [18] T. Sadhu, A. Pesaranghader, Y. Chen, and D. H. Yi, "Athena: Safe autonomous agents with verbal contrastive learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 1121–1130. doi: <https://doi.org/10.18653/v1/2024.emnlp-industry.84>.
- [19] N. Zwerdling, D. Boaz, E. Rabinovich, G. Uziel, D. Amid, and A. A. Tavor, "Towards enforcing company policy adherence in agentic workflows," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025, pp. 595–606. doi: <https://doi.org/10.18653/v1/2025.emnlp-industry.41>.
- [20] J. Ankam, "Federated Governance Without Central Bottlenecks: A Policy Propagation Model for Data Mesh," *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.7323259.
- [21] P. Danassis, Z. D. Erden, and B. Faltings, "Exploiting environmental signals to enable policy correlation in large-scale decentralized systems," *Auton. Agent. Multi. Agent. Syst.*, vol. 36, no. 1, p. 13, 2022, doi: <https://doi.org/10.1007/s10458-021-09541-7>.
- [22] R. Shahroz, Z. Tan, S. Yun, C. Fleming, and T. Chen, "Agents under siege: Breaking pragmatic multi-agent llm systems with optimized prompt attacks," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 9661–9674. doi: <https://doi.org/10.18653/v1/2025.acl-long.476>.
- [23] J. Ankam, "Retrieval-Ready Pipelines: Benchmarking Chunking and Embedding Strategies for Enterprise RAG," *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.7323278.
- [24] D. Agarwal, A. R. Fabbri, B. Risher, P. Laban, S. Joty, and C.-S. Wu, "Prompt Leakage effect and mitigation strategies for multi-turn LLM Applications," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. pp. 1255–1275, 2024. doi: 10.18653/v1/2024.emnlp-industry.94.
- [25] T. Yuan *et al.*, "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents," *Findings of the Association for Computational Linguistics: EMNLP 2024*. pp. 1467–1490, 2024. doi: 10.18653/v1/2024.findings-emnlp.79.
- [26] J. Ankam, "ACID-Lake: A Metadata-Driven Concurrency Control Protocol for Multi-Writer Transactions on Object Storage," *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.7323258.
- [27] N. Palladino, "A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices," *Telecomm. Policy*, vol. 47, no. 5, p. 102479, 2023, doi: 10.1016/j.telpol.2022.102479.

BIOGRAPHY OF AUTHOR



Jagan Ankam is a technology leader with more than two decades of experience building data engineering and AI capability for enterprise organizations in finance and e-commerce. A Senior Member of the IEEE, he is currently a Senior Consultant in Computer Information Systems with CGI Inc. in Miami, Florida, where he helps clients modernize their data platforms and put in place the cloud and data foundations needed for AI adoption.

Before joining CGI, Jagan spent many years at Tata Consultancy Services, where he led several large-scale projects across data engineering, data science, and data lakehouse initiatives. Over his career he has worked with a wide range of data engineering technologies and AI cloud platforms, bringing a hands-on, delivery-focused approach to complex enterprise environments.

Jagan's research and professional interests include AI data engineering, cloud database systems, data governance, and cloud computing. His research has been published in IEEE conferences and the International Journal of Computational and Experimental Science and Engineering (IJCESEN). He holds a master's degree in Management Information Systems from Sikkim Manipal University.