Towards Governance-Aware Local and Hybrid AI Agents for Web Applications

Lucas Schröder[1][0009−0009−5540−4495] and Martin Gaedke[1][0000−0002−6729−2912]*

Chemnitz University of Technology, Chemnitz, Germany
{lucas.schroeder, martin.gaedke}@informatik.tu-chemnitz.de

Abstract. Governance of AI systems is becoming a critical issue as both societal and regulatory circumstances demand requirements such as safe information handling, transparency, and explainability. This is complicated by the widespread use of third-party-operated LLMs, which limits organizations' control over the embedding and enforcement of governance and organizational policies directly within the agent. Consequently, users are required to exercise caution or are prevented from using these agents when interacting with web applications that contain potentially sensitive information or require transparent and explainable processing. We propose a framework and architecture for embedding governance and organizational policies for data handling and explainability within web agents based on local, tool-calling small language models. This core will be expanded with governance-aware hybrid routing to remote LLMs for non-critical tasks, balancing the high capabilities of large, third-party provided systems with the safe and transparent environment established through local processing.

Keywords: Governance · Web Agents · Human-AI Collaboration · Explainable AI · Small Language Models

1 Introduction

The field of artificial intelligence (AI) has seen enormous advancements in recent years, leading to a rapid adoption of the new technologies in almost every field of life. While initially using primarily reactive, chat-based interfaces, autonomous systems have emerged in the form of AI agents that can perform a variety of actions to solve tasks independently. For that, agents are often interacting with existing systems, including enterprise web applications such as knowledge, project, or customer relationship management tools. [4, 6, 11, 13]

As both the capabilities and adoption of these AI systems increase rapidly, governance over them is catching up slowly [6, 7, 13]. This makes the issue of ensuring the ethical, safe, and explainable use of AI increasingly relevant, as

* PhD supervisor of the first author.

evident in the growing number of regulations, guidelines, and frameworks such as the OECD AI Principles[1], EU AI Act[2], 2024 PLD[3], or ISO/IEC 42001[4].

Two major points of interest in these frameworks are the involvement of data and the explainability of the agentic systems. Because of the established model of cloud-based processing using third-party-operated AI models, the autonomous access to large amounts of data, often including sensitive information, inherently includes risks regarding privacy and data protection [13, 14]. At the same time, the agentic systems often act as a black box, making the agents' decision-making and actions non-transparent and unexplainable at times [6, 8, 13]. Both of these aspects are detrimental to the governance that users and organizations have over AI systems interacting with web applications, other agents, or other users.

Governance of AI systems generally concerns several aspects, including technical, ethical, and regulatory ones [1, 7, 8, 11]. In the context of this work, the focus will be on technical operationalization, namely on enforcing policies about data access and data flows, in addition to transparency and explainability of the involved actions. This leads to our central research question:

How can AI agents that interact with web applications, other AI agents, and even other users be systematically designed to comply with governance or organizational policies, especially regarding data handling and explainable processing of data?

In this work, we propose exploring the use of locally running AI agents based on Small Language Models (SLMs) with governance-aware tool calls for interacting with web applications, other agents, or other users, complemented by hybrid routing to external Large Language Models (LLMs) for non-critical tasks.

2 Problem Statement & Motivation

Embedding governance policies in AI agents interacting with web applications is particularly challenged by the prevalent cloud provider-based approach: By sending information for processing to what is essentially a third-party-operated black box, governance control over safe and explainable processing of potentially sensitive data is delegated to be handled by providers instead. [10, 13]

As a consequence, organizations do not generally have the ability to embed their own governance policies deep into the agent loop, which is especially relevant for organizations that require higher levels of governance than enabled by the provider [10]. Users who interact with such third-party-operated systems, therefore, first need to filter any data that is to be sent to the AI agent to avoid unintentionally exposing any sensitive information [4]. At the same time, they might be unable to use the agent at all for tasks that require auditability or explainability, because intermediate actions taken by the AI agents, such as tool

[1] https://www.oecd.org/en/topics/ai-principles.html
[2] https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[3] https://eur-lex.europa.eu/eli/dir/2024/2853/oj
[4] https://www.iso.org/standard/42001

calls, are opaque to the user, unless the provider gives sufficient insight into the processing steps, for instance, by providing very detailed logs or traces.

These circumstances lead to users being unable to use these AI agents within their workflows or when interacting with existing web applications that either contain potentially sensitive information or require additional oversight or explainable processing. As a result, the usefulness of these agentic systems is reduced for affected users, as the additional effort to ensure compliance with their organizations' governance policies can outweigh the benefits of using the AI agent, if they can even be made compatible at all.

Both ethical and regulatory reasons demand that governance is applied to AI systems, including the safe handling of potentially sensitive information or the explainability of processing. As such, AI agents interacting with web applications or other agents on behalf of their users also need to be aware of these governance objectives and implement the applicable regulations. Consequently, there is a growing need to embed governance-awareness into these AI agents, allowing affected users to make full use of the capabilities of these agents. [1, 5, 6, 11]

3 Related Work

Agentic AI is increasingly employed to interact with web applications, including enterprise web applications, for example, through agentic web browsers. While initially relying on a mixture of DOM- or vision-based approaches for direct, human-like UI interactions in web applications, there are first approaches to translate MCP-like tool use to web applications, such as WebMCP [9, 12, 13]. While these approaches enable AI agents to interact with web applications on behalf of their users, comprehensive governance-aware architectures for these kinds of agents are still missing.

A growing body of work is focused on establishing governance for AI-based systems in the form of regulations, such as the EU AI Act, 2024 PLD, or frameworks and guidelines like ISO/IEC 42001 or the NIST AI RMF [8]. They consider governance aspects throughout the entire life-cycle of AI systems, from model training to the use of the deployed system, but often focus on principles, describing qualities that should be achieved or requirements that AI systems, or organizations using these systems, need to fulfill. However, these frameworks often do not yet consider how to operationalize these principles within tool-calling AI agents interacting with web applications, other agents, or human users. [7]

Guardrail-based systems provide a first step in this direction by providing enforcement capabilities for automatic filtering of user inputs or agent outputs. While useful for implementing some governance principles, they are not yet fully developed to deal with more complex scenarios, such as access to sensitive information across multiple web applications and other systems. [3]

A different aspect of the topic of governance of AI agents and the handling of potentially sensitive information is the topic of hybrid systems, such as Hybrid LLM [2], which combine local and remote processing. While often focusing on cost-saving or performance, systems such as PRISM [14] also consider sen-

sitive information handling, routing calls depending on sensitivity to either a local/edge model or to a remote model. How to embed governance or organizational policies into such hybrid systems, specifically for AI agents interacting with web applications via tool calls, however, has not yet been explored in-depth.

This highlights a clear gap: Existing work on web-interacting agents, AI governance, and hybrid local/remote setups has not yet produced a comprehensive framework for embedding governance policies within the tool-calling and routing of AI agents interacting with web applications. In order to better analyze existing work and to gain a more comprehensive overview of the current state of the art, we plan to develop indicators to evaluate and compare prior work.

4 Objectives & Contributions

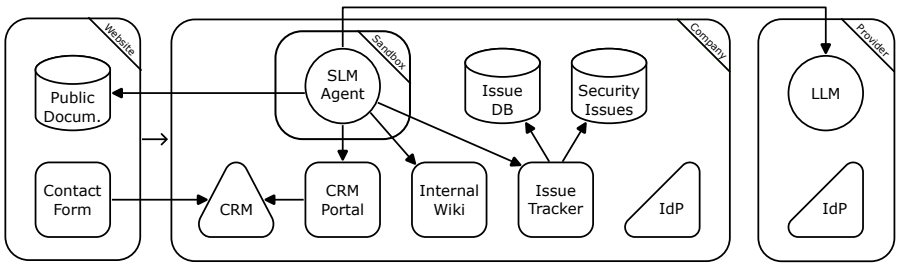

Fig. 1. An example of an agent integrated into the systems of a company. Access to multiple systems is necessary for its functionality, but poses governance issues, e.g., for data handling. This is combated through the use of a local, sandboxed SLM, governance-aware system interactions, and hybrid routing to the third-party LLM.

To better understand the objectives of this work, consider the following short scenario, which is also visualized in Figure 1: A software company has deployed an agent to help customer support staff in handling incoming inquiries in a customer relationship management (CRM) application, acting as first-level support. To fulfill its tasks, the agent has access to the company's public product documentation, the internal wiki, and the development team's issue tracker. Naturally, there are several constraints for the agent's processing: Customer information requires local processing, vulnerabilities from the issue tracker should not be exposed, replies need to be accurate and harmless, and the agent's actions need to be explainable and auditable.

This scenario highlights how governance requirements affect agents and why their enforcement is critical. To solve this, we propose an architecture based on sandboxed local processing, governance-aware interactions, and hybrid routing. Our main contributions will thus be represented by the following research objectives:

Sandboxed, local processing: Relying on third-party AI providers puts control over governance into the providers' hands, leading to potential compliance risks [10, 13, 14]. To prevent this and ensure reasonable control over and visibility into the agents' actions, we propose local processing using SLMs in a sandboxed environment. To maintain the required accuracy for tool calling, necessary optimizations need to be made to prompt and context management.

Enforcement of governance objectives in prompts and tool calls: Operationalizing the governance objectives determined by relevant regulations and organizational policies requires enforcement through technical means [5–7]. Current approaches to operationalize governance requirements need to be expanded using both prompt- and guardrail-based approaches, in addition to logging and other measures enabling explainability and auditability.

Routing decisions for hybrid systems: While local processing will be necessary sometimes, not every request will require it. As such, making use of the high capabilities of third-party operated models makes sense for some requests, which require adequate routing decisions [13, 14] to identify when this is the case. Existing approaches need to be expanded and integrated with the other techniques to form a comprehensive framework for enforcing governance and organizational policies directly within the AI agent.

5 Current State & Conclusion

So far, this work is still in an early stage of conceptualization: We have identified the research gap of governance-awareness in AI agents interacting with web applications, reviewed core literature, and created an initial outline of the problem formulation and research objectives for closing the gap, as well as an early concept for a solution architecture. The next steps will be to further analyze the state of the art in the involved areas over the next few months, before refining the conceptual architecture further until the beginning of next year. An implementation and evaluation of an initial, primarily local prototype of an AI agent for governance-aware interactions with web applications will follow and, after evaluation, be expanded with hybrid routing capabilities to remote LLMs.

Governance of AI systems is an increasingly relevant topic, including technical, ethical, and regulatory aspects. Embedding enforcement of governance policies directly into the agent is critical to allow organizations to make full use of the capabilities provided by these agents. To tackle the issue of governance-awareness in AI agents interacting with web applications, we propose the use of local SLMs and lightweight tool calls for governance-aware interactions with web applications, while routing non-critical tasks to third-party operated LLMs.

Acknowledgments. This work is supported by the European Union's Erasmus+ Programme under grant agreement No 101177485, project Across (European University for Cross-Border Knowledge Sharing), and by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chaudhry, U., Jones, J., Casovan, A., Burke, L., Bryant, N., Resmerita, L., et al.: AI governance in practice report 2024. Tech. rep. (Jun 2024), https://iapp.org/resources/article/ai-governance-in-practice-report
2. Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., et al.: Hybrid LLM: Cost-efficient and quality-aware query routing (Apr 2024). https://doi.org/10.48550/arXiv.2404.14618
3. Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., et al.: Position: Building guardrails for large language models requires systematic design. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 11375–11394. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/dong24c.html
4. Du, Y., Li, Z., Li, N., Ding, B.: Beyond data privacy: New privacy risks for large language models. Bulletin of the Technical Committee on Data Engineering **49**(4), 47–75 (2025)
5. Gaurav, S., Heikkonen, J., Chaudhary, J.: Governance-as-a-service: A multi-agent framework for AI system compliance and policy enforcement (Aug 2025). https://doi.org/10.48550/arXiv.2508.18765
6. Kraprayoon, J., Williams, Z., Fayyaz, R.: AI agent governance: A field guide (May 2025). https://doi.org/10.48550/arXiv.2505.21808
7. Lucaj, L., van der Smagt, P., Benbouzid, D.: Ai regulation is (not) all you need. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 1267–1279. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594079
8. National Institute of Standards and Technology: Artificial intelligence risk management framework: AI RMF 1.0. Tech. Rep. NIST AI 100-1 (Jan 2023). https://doi.org/10.6028/NIST.AI.100-1
9. Ning, L., Liang, Z., Jiang, Z., Qu, H., Ding, Y., Fan, W., et al.: A survey of WebAgents: Towards next-generation AI agents for web automation with large foundation models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. p. 6140–6150. KDD '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3711896.3736555
10. Petrin, M.: The impact of AI and new technologies on corporate governance and regulation. Singapore Journal of Legal Studies p. 90 (2024)
11. Tallam, K.: From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence (Mar 2025). https://doi.org/10.48550/arXiv.2503.13754
12. Walderman, B., Sagar, K., Farolino, D.: WebMCP. Tech. rep. (Mar 2026), https://webmachinelearning.github.io/webmcp
13. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., et al.: The rise and potential of large language model based agents: a survey. Science China Information Sciences **68**(2), 121101 (Jan 2025). https://doi.org/10.1007/s11432-024-4222-0
14. Zhan, J., Shen, H., Lin, Z., He, T.: PRISM: Privacy-aware routing for adaptive cloud-edge LLM inference via semantic sketch collaboration (Nov 2025). https://doi.org/10.48550/arXiv.2511.22788