Oversight Structures for Agentic AI in Public-Sector Organizations

Chris Schmitz[1], **Jonathan Rystrøm**[2], **Jan Batzner**[3,4]
[1]Centre for Digital Governance, Hertie School, Germany,
[2]Oxford Internet Institute, University of Oxford, UK,
[3]Weizenbaum Institute Berlin, Germany,
[4]Technical University Munich, Germany
Correspondence: ch.schmitz@hertie-school.org

Abstract

This paper finds that the introduction of agentic AI systems intensifies existing challenges to traditional public sector oversight mechanisms — which rely on siloed compliance units and episodic approvals rather than continuous, integrated supervision. We identify five governance dimensions essential for responsible agent deployment: cross-departmental implementation, comprehensive evaluation, enhanced security protocols, operational visibility, and systematic auditing. We evaluate the capacity of existing oversight structures to meet these challenges, via a mixed-methods approach consisting of a literature review and interviews with civil servants in AI-related roles. We find that agent oversight poses intensified versions of three existing governance challenges: continuous oversight, deeper integration of governance and operational capabilities, and interdepartmental coordination. We propose approaches that both adapt institutional structures and design agent oversight compatible with public sector constraints.

1 Introduction

Artificial Intelligence (AI) technologies, particularly Large Language Model-based agents, hold the potential to fundamentally transform Public Sector Organizations (PSOs) via efficiency enhancements and process automation (Ilves et al., 2025; Straub et al., 2024; Shavit et al., 2023). However, PSOs often lack the institutional capacity and agile governance[1] structures required for responsible adoption (Lawrence et al., 2023). These challenges are driven by characteristic features of bureaucratic organizations, such as procedural rigidity, hierarchical accountability, and dual demands for transparency and effectiveness, which complicate both AI deployment and oversight (Madan and Ashok, 2023; Neumann et al., 2024).

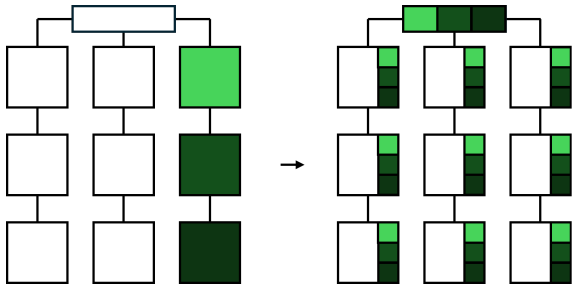

Figure 1: To enable agent oversight, PSOs may move from standalone digital governance functions (in green) towards centrally coordinated, distributed governance.

Although recent work has proposed best practices for AI agent governance (Shavit et al., 2023; Chan et al., 2024), it therefore remains unclear whether PSOs are structurally and procedurally equipped to implement these recommendations. The governance of agent systems likely requires both technical implementation and institutional adaptation — affecting processes, oversight mechanisms, and digital infrastructure (Madan and Ashok, 2023; Neumann et al., 2024). We ask the following research question:

How do existing public sector AI governance structures meet the requirements for responsible agent deployment?

Based on interviews with civil servants, we find that agent governance requirements exasperate existing digital governance challenges in PSOs. We first outline the potential of AI agents in the public sector (§2) and emerging agent governance tasks (§3) . We derive a hypothesis on challenges these tasks pose to public-sector oversight structures from public administration literature (§4), and describe our interview methodology (§5). We summarize our findings on existing oversight structures

[1]Throughout this article, we use the term "governance" to refer specifically to tasks and structures for oversight of LLM agents and other digital projects, rather than to the political-scientific definition, which may encompass all PSO activities.

Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 298–308
July 31, 2025 ©2025 Association for Computational Linguistics

(§6) and, finally, chart the ways agent oversight intensifies existing governance challenges in PSOs (§7). Our core contribution is an evaluation of the feasibility of agent oversight in current real-world public-sector contexts.

2 Potential of Agents in PSOs

Recent work demonstrates the application of LLM-based agents to tasks in narrow domains. Examples include implementations of agents designed to emulate domain experts such as academic researchers (Schmidgall et al., 2025), office workers (Gur et al., 2023), developers (Jimenez et al., 2023) or healthcare professionals (Zhao et al., 2025). Moreover, major LLM providers started introducing their Computer Use Agents for end-users in fall 2024 (Anthropic, 2024; OpenAI, 2025; Pichai et al., 2024).

While the AI community acknowledges the potential of such increasingly agentic systems (Liu et al., 2023; Lu et al., 2025), which integrate autonomous learning, tool use, and complex reasoning, Kapoor et al. (2024) emphasize their substantial current limitations, including susceptibility to benchmark overfitting, prohibitive operational costs, and unnecessarily complex computational architectures — which each scale with domain generality. Narrow, domain-specific agentic systems may therefore represent a more practical and immediately valuable development pathway.

The public sector is such a domain, as public-sector tasks possess ideal characteristics for automation: they are comprehensively documented, exhibit consistent structural patterns, and involve high-frequency repetition (Bullock et al., 2020). Focusing agent development on these well-structured tasks provides a pragmatic pathway toward realizing the benefits of AI agents while mitigating their current limitations.

Previous waves of digitization standardized data formats, process explanations, and decision frameworks (Bovens and Zouridis, 2002), contributing to *bureaucratic rationality*: the formal legibility and transparency of bureaucracies, from which they gain their legitimacy. LLM-based agents potentially continue this process, e.g., by supporting documentation requirements in specific contexts (Newman et al., 2022; Mökander and Schroeder, 2024).

Conversely, due to their well-known mechanistic inexplicability (Sharkey et al., 2025), AI systems including agents may also hinder, rather than enhance, the further rationalization of state entities. Given the fundamental importance of correctly-attributed responsibility to bureaucratic legitimacy (Cetina Presuel and Martinez Sierra, 2024), successful integration of agentic systems therefore depends on the careful design of accountability mechanisms.

The realization of this potential further depends on the resolution of myriad practical challenges commonly faced in PSOs: resource and capacity strain, varying levels of digitalization of existing processes, and growing internal and external reporting requirements, to name a few (Lawrence et al., 2023).

3 Agent Governance Challenges

Deploying LLM-based agents in the public sector presents novel governance challenges over and above those introduced by non-agentic software. In this section, we identify five interdependent governance areas that collectively enable responsible agent deployment.

Distributed Implementation: Compared to traditional IT projects, LLM agent-based systems add layers of complexity due to their many interacting components (Wang et al., 2024), which often span organizational units. A crucial challenge is to appropriately assign responsibility for different parts of the systems and ensure well-functioning cross-unit communication channels. For instance, implementing a customer query agent requires coordination between IT, customer service, and business intelligence units (for data access; Rome et al., 2024).

Visibility: Visibility into agent activity is an essential prerequisite for both organizational awareness (Straub et al., 2023) and the operational capabilities to identify and intervene in misbehaving systems (Chan et al., 2024). Visibility is generated at both the system level, via agent indexes similar to model cards (Mitchell et al., 2019; Derczynski et al., 2023), and the operational level, via real-time monitoring (Chan et al., 2024). Such monitoring enables human-in-the-loop safeguards for continuous oversight and proper attribution of responsibility (Ananny and Crawford, 2018).

Evaluation: Pre-deployment evaluation ensures that agent systems work as intended (Wang et al.,

2024). These evaluations must include system-level, task-specific, and comprehensive metrics (e.g., Jimenez et al., 2023; Kapoor et al., 2024), rather than general, component-level evaluations (e.g., Ye et al., 2024). Evaluation should enable socio-technical comparisons with existing human-based systems, and therefore should be designed in close collaboration with operational workers (Selbst et al., 2019).

Security: AI Agents present novel security challenges due to their complexity (Deng et al., 2025). Threats include jailbreaking (Tian et al., 2024), data exfiltration (Zeng et al., 2024), and exposure to DDoS attacks (Zhang et al., 2024). While many of these challenges fit well under existing cybersecurity frameworks (Krumay et al., 2018) and cybersecurity hygiene practices (Gupta and Furnell, 2022), the autonomy of Agents might require novel governance strategies and practices (Deng et al., 2025). Addressing these challenges requires cross-departmental threat mapping, mitigation strategies, and red-teaming efforts (Inie et al., 2025).

Auditing: Agent auditing capacity should build on the previous capacities to ensure holistic compliance across processes, components, and applications (Mökander et al., 2024). Agent auditing provides external compliance validation which ensures that efforts within Visibility, Evaluation, and Security are up to proper standards (Sandvig et al., 2014). While no formalized auditing frameworks for agents are well-established (Chan et al., 2023), PSOs should nonetheless strive to implement both entities and standards for independent auditing procedures.

4 Public Administration Context

This section draws on public administration theory to hypothesize that governing LLM-based agents challenges existing governance structures in PSOs.

PSOs are overwhelmingly Weberian bureaucracies (Weber, 1991) with strict delineation of responsibilities, specialization, hierarchical structure, and well-documented, impersonal processes (Clegg and Lounsbury, 2009). Tasks are differentiated into secluded units (Blau, 1970) with limited, formalized intra-unit communication, in the case of governance often in the form of approval processes. Governance structures for AI, where existent, often inherit these features, characterized by episodic approvals and externalization (Lawrence et al., 2023). Bureaucracies may arrive at these structures via "isomorphic" processes, in which they mimic similar organizations without considering whether these structures are well-suited for their individual case (DiMaggio and Powell, 1983).

The well-documented inefficiencies of these structures (Niskanen, 1971) have led to the development of several more dynamic governance frameworks, notably New Public Management (Hood, 1991), Digital-Era Governance (Dunleavy et al., 2006), "governing by network" (Goldsmith and Eggers, 2004) and derived meta-governance approaches (Sørensen and Torfing, 2017). Each of these have found limited success in practice, in part due to the persistence of external factors forcing rationalization, such as administrative law and freedom of information regulation (Pozen, 2018).

We therefore hypothesize that PSOs' existing governance structures are only partially compatible with the requirements posed by the integration of LLM-based agents. While some novel agent governance tasks may be well-situated within existing bureaucratic oversight structures, holistically governing LLM-based agents will require adaptations that challenge the typical segmentation, timing, and location of governance activities within PSOs.

5 Interview Methodology

In addition to our literature review, we conducted six semi-structured interviews with German officials across six agencies: three federal, two state, and one municipal agency. All participants were recruited through a certificate course on AI in the public sector coordinated by the first author. The interviews followed the interview guide in Appendix A. We qualitatively analyzed the interviews using open coding (Charmaz, 2006). Finally, we present the COREQ checklist in Appendix B (Tong et al., 2007) following best practices in the field (Adeoye-Olatunde and Olenik, 2021).

6 Findings on Existing Oversight Structures

Here we summarize our interview findings into key attributes of existing governance structures for digital projects within PSOs.

Legal Motivation: Well-established governance generally only exists where legally mandated. In Germany, AI systems as of today underlie no specific laws. As a result, in the surveyed organizations AI governance is largely handled through

structures established for traditional digital and process governance. Few organizations have AI-specific governance, such as logging and oversight requirements, in place; only one interviewee cited an AI-specific governance body with formal authority in their organization.

Dedicated Governance Units: Established digital governance responsibilities include data protection, cybersecurity, and accessibility. In the studied organizations, each of these are held by dedicated units or individuals, which frequently sit in standalone positions across the organization — in the case of data protection, this is legally required in the EU. These individuals wield significant authority, including veto rights on projects.

Event-triggered Involvement: Interviewees in implementing departments report *proactively* consulting governance teams to trigger compliance processes and obtain approval. These interactions are event-triggered when adapting projects in ways relevant to these teams. Examples of named events were new projects, processing new forms of data, entering new stages of deployment, e.g., from PoC to internal prototype, or adding new features to existing projects. Governance is not generally considered a continuous process after approval is given or a solution is deployed.

Adversarial Compliance Processes: Because compliance functions are rarely directly integrated in development processes, the relationship between compliance teams and project teams was often described as *adversarial*. Projects scrutinized by compliance teams frequently require time-consuming iterative processes to redefine scope and structure before compliance approval is granted. Some interviewees report "self-censorship" and reductions in the scope of projects before involving these units.

Limited Employee Capacity: Interviewees lamented that outside "digital" units, employee capacity for digital skills is generally low. Simultaneously, their workload and expected throughput is frequently very high, leading to the perception of additional duties, such as transition to new technologies, oversight of imperfectly functioning digital tools, or development of model evaluation metrics, as burdens rather than reliefs.

Success of Breaking Hierarchy: Interviewees partaking in models of collaboration that break these hierarchical and adversarial structures report higher project success and smoother governance integration. Examples include cross-functional expert steering teams that integrate governance functions in project management, and ad-hoc interest networks of AI specialists from many units across an organization.

7 Discussion

We find strong evidence for our hypothesis that current governance structures face severe challenges in adapting to agent governance. In particular, our interviews reveal agent governance requirements may produce *intensifications* of governance challenges already familiar to PSOs from existing digitalization projects and implementations of non-agentic AI systems. Here, we first aggregate three classes of shortcomings and conclude by highlighting promising paths forward for both agent architectures and PSO governance structures.

1. **Continous oversight** is required to translate mechanical visibility into accountability at the operational level. It likely cannot be guaranteed entirely by segmented structures that isolate governance requirements in separated teams, or by processes that are event-triggered: the frequency of events produced by agents exponentiates communication costs between operative and governance units, which are already prohibitive. As corroborated by most interviewees, governance responsibilities must therefore be diffused towards the end-users: the implementing operational departments whose work is augmented by agents.

2. **New governance capabilities** are required throughout the agent development cycle. Some specific expert capabilities, such as pre-deployment testing and occasional auditing, may be well covered by new departments within existing governance structures. Visibility and evaluation, however, require much deeper integration between subject knowledge and technical understanding — again necessitating upskilling and a redefined role for operational workers. This in turn may amplify a variety of risks related to how these workers, who now take on dual roles as overseers, are influenced by the technology in the exercise of their own discretion (de Boer and Raaphorst, 2023; Bullock et al., 2020). Some degree of central oversight must thus likely remain.

3. **Interdepartmental coordination** is required. Agents, which handle processes and connect to tools and databases across traditional departmental boundaries, intensify an existing digital governance challenge frequently named by interviewees: collaboration and responsibility attribution for increasingly complex, cross-cutting projects. Required adaptations include mechanisms for cross-departmental coordination, and sufficient in-department governance competency to avoid dependence on mediation and consultation by external units. An interviewee already involved in structures enabling this, both for governance and for project coordination, cited them as "unlocking the potential" for many projects.

Collectively, these developments imply successful agent governance is centrally coordinated, but diffused throughout PSOs. We visualize this transition in figure 1. This matrix organization (Turk, 2018) is evocative of governance models that iterate on strictly bureaucratic structures, such as those mentioned above (§4). A promising direction of future research is therefore to evaluate which features of these neo-Weberian governance approaches may be promising for agent governance — especially as the uptake of agentic systems drives more fundamental transformation of state structures (Ilves et al., 2025). We propose three design principles for the tailoring of technical work on agent oversight to public-sector contexts.

1. **Design observability tooling to be utilized by either technical external teams, or subject matter experts.**

   Existing agent observability tooling, including logging of actions and tools, control over autonomy levels, and evaluation metrics, frequently interweaves subject-specific online oversight and technical, abstract offline oversight in combined interfaces and applications (Dong et al., 2024). Our interviews show that PSOs frequently maintain organizational separation between these groups of tasks (§6). Oversight tooling should therefore be designed with delegation to these distinct groups in mind.

2. **Design oversight interfaces in collaboration with non-technical public servants where online supervision is required.**

   Non-technical public servants are usually ill-suited and not enabled by the organization to translate technical oversight metrics into those required for performance of their duties. Mechanisms by which they control and oversee agent actions must therefore reflect the requirements, logics, and language of the individual end user and the organization, rather than derive from the provided technology (§3). Beyond enabling non-expert oversight, this allows the attribution of responsibility for agent actions to humans.

3. **Design agent oversight that anticipates interaction with legacy systems.**

   While the potential of agents (§2), is enormous in the public sector, the ideal-state implementation relies on the prior end-to-end digitalization of existing services, and is unlikely to be realized in PSOs with fragmented, often partially manual, processes across legacy systems (§3). Agent oversight systems must therefore allow for this variability. This includes designing systems that do not assume complete process observability, and incorporating human-in-the-loop fallback mechanisms for scenarios where automated oversight is insufficient.

Limitations

This study has several limitations that should be acknowledged: *(1) Institutional scope:* Our analysis is limited to German public sector organizations with their specific administrative traditions. While the results are indicative of challenges faced by public sector organisations, they cannot be directly applied to other administrative contexts. *(2) Sample characteristics:* Our sample of six qualitative expert interviews, while providing consistent corroboration of our core hypotheses, may not capture the full spectrum of governance challenges across all public sector organizations. *(3) Implementation Challenges:* Our framework requires context-specific adaptation across diverse institutional structures. Our focus on PSOs as standalone entities does not capture their integration with private-sector institutions for LLM or agent technology, infrastructure, product development, auditing and oversight services.

Acknowledgements

JR was supported by the Engineering and Physical Sciences Research Council [Grant Number EP/W524311/1]. JB was supported by the Federal Ministry of Education and Research of Germany [Grant Number 16DII131]. CS and JB acknowledge funding by the Hertie School's program "AI and Data Science for the Public Sector", funded by the Dieter Schwartz Foundation.

References

Omolola A. Adeoye-Olatunde and Nicole L. Olenik. 2021. Research and scholarly methods: Semi-structured interviews. *Jaccp: Journal of the American College of Clinical Pharmacy*, 4(10):1358–1367.

Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989.

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. Blog post.

Peter M. Blau. 1970. A Formal Theory of Differentiation in Organizations. *American Sociological Review*, 35(2):201–218. Publisher: [American Sociological Association, Sage Publications, Inc.].

Mark Bovens and Stavros Zouridis. 2002. From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review*, 62(2):174–184. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0033-3352.00168.

Justin Bullock, Matthew M. Young, and Yi-Fan Wang. 2020. Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, 25(4):491–506.

Rodrigo Cetina Presuel and Jose M. Martinez Sierra. 2024. The adoption of artificial intelligence in bureaucratic decision-making: A weberian perspective. *Digital Government: Research and Practice*, 5(1):1–20.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. 2024. Visibility into AI agents. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 958–973, Rio de Janeiro Brazil. ACM.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 651–666, New York, NY, USA. Association for Computing Machinery.

Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE.

Stewart Clegg and Michael Lounsbury. 2009. Weber: Sintering the Iron Cage Translation, Domination, and Rationality Stewart Clegg. In Paul Adler, editor, *The Oxford Handbook of Sociology and Organization Studies: Classical Foundations*, page 0. Oxford University Press.

Noortje de Boer and Nadine Raaphorst. 2023. Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Management Review*, 25(1):42–62. Publisher: Routledge _eprint: https://doi.org/10.1080/14719037.2021.1937684.

Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. AI agents under threat: A survey of key security challenges and future pathways. *ACM Computing Surveys*, 57(7):1–36.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *Preprint*, arXiv:2303.18190.

Paul J. DiMaggio and Walter W. Powell. 1983. The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48(2):147–160. Publisher: [American Sociological Association, Sage Publications, Inc.].

Liming Dong, Qinghua Lu, and Liming Zhu. 2024. AgentOps: Enabling Observability of LLM Agents. *arXiv preprint*. ArXiv:2411.05285 [cs].

Patrick Dunleavy, Helen Margetts, Simon Bastow, and Jane Tinkler. 2006. *Digital era governance: IT corporations, the state, and e-Government*. Oxford University Press.

Stephen Goldsmith and William D. Eggers. 2004. *Governing by Network: The New Shape of the Public Sector*. Brookings Institution Press.

Shreya Gupta and Steven Furnell. 2022. From cybersecurity hygiene to cyber well-being. In *HCI for Cybersecurity, Privacy and Trust: 4th International Conference, HCI-CPT 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual*

Event, June 26 – July 1, 2022, Proceedings, pages 124–134, Berlin, Heidelberg. Springer-Verlag.

Izzeddin Gur, Hiroki Furuta, Austin V. Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world WebAgent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations*.

Christopher Hood. 1991. A Public Management for All Seasons? *Public Administration*, 69(1):3–19. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9299.1991.tb00779.x.

Luukas Ilves, Max Kilian, Tiago Peixoto, and Ott Velsberg. 2025. The agentic state: How agentic AI will revamp 10 functional layers of government and public administration. Tex.howpublished: Whitepaper, Global GovTech Centre.

Nanna Inie, Jonathan Stray, and Leon Derczynski. 2025. Summon a demon and bind it: A grounded theory of LLM red teaming. *PLOS One*, 20(1):e0314658.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2023. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI agents that matter. *Preprint*, arXiv:2407.01502.

Barbara Krumay, Edward W. N. Bernroider, and Roman Walser. 2018. Evaluation of cybersecurity management controls and metrics of critical infrastructures: A literature review considering the NIST cybersecurity framework. In *Secure IT Systems*, pages 369–384, Cham. Springer International Publishing.

Christie Lawrence, Isaac Cui, and Daniel Ho. 2023. The bureaucratic challenge to AI governance: An empirical assessment of implementation at U.S. federal agencies. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 606–652, New York, NY, USA. Association for Computing Machinery.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. AgentBench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2025. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1160–1183, Albuquerque, New Mexico. Association for Computational Linguistics.

Rohit Madan and Mona Ashok. 2023. AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, 40(1):101774.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Jakob Mökander and Ralph Schroeder. 2024. Artificial intelligence, rationalization, and the limits of control in the public sector: The case of tax policy optimization. *Social Science Computer Review*, 42(6):1359–1378.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115.

Oliver Neumann, Katharina Guirguis, and Reto Steiner. 2024. Exploring artificial intelligence adoption in public organizations: A comparative case study. *Public Management Review*, 26(1):114–141.

Joshua Newman, Michael Mintrom, and Deirdre O'Neill. 2022. Digital technologies, artificial intelligence, and bureaucratic transformation. *Futures*, 136:102886.

William A. Niskanen. 1971. *Bureaucracy and Representative Government*. Transaction Publishers. Google-Books-ID: dOYe1ld9F1QC.

OpenAI. 2025. Introducing operator. Blog post.

Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health*, 42(5):533–544.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2024. Introducing gemini 2.0: our new ai model for the agentic era. Blog post.

David E. Pozen. 2018. Transparency's Ideological Drift. *Yale Law Journal*.

Scott Rome, Tianwen Chen, Raphael Tang, Luwei Zhou, and Ferhan Ture. 2024. "ask me anything": How comcast uses LLMs to assist agents in real time. In *Proceedings of the 47th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '24, pages 2827–2831, New York, NY, USA. Association for Computing Machinery.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 22:4349–4357.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *Preprint*, arXiv:2501.04227.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA. ACM.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open problems in mechanistic interpretability. *Preprint*, arXiv:2501.16496.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, and Alan Hickey. 2023. Practices for governing agentic AI systems.

Vincent J. Straub, Youmna Hashem, Jonathan Bright, Satyam Bhagwanani, Deborah Morgan, John Francis, Saba Esnaashari, and Helen Margetts. 2024. AI for bureaucratic productivity: Measuring the potential of AI to help automate 143 million UK government transactions. *Preprint*, arXiv:2403.14712.

Vincent J. Straub, Deborah Morgan, Jonathan Bright, and Helen Margetts. 2023. Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly*, 40(4):101881.

Eva Sørensen and Jacob Torfing. 2017. Metagoverning Collaborative Innovation in Governance Networks. *The American Review of Public Administration*, 47(7):826–839.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. 2024. Evil geniuses: Delving into the safety of LLM-based agents. *Preprint*, arXiv:2311.11855.

Allison Tong, Peter Sainsbury, and Jonathan Craig. 2007. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6):349–357.

Thomas A. Turk. 2018. Matrix Organization. In Mie Augier and David J. Teece, editors, *The Palgrave Encyclopedia of Strategic Management*, pages 1030–1033. Palgrave Macmillan UK, London.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Max Weber. 1991. *From Max Weber: Essays in Sociology*. Psychology Press. Google-Books-ID: Y_pqZS5q72UC.

Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. ToolSword: Unveiling safety issues of large language models in tool learning across three stages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2181–2211, Bangkok, Thailand. Association for Computational Linguistics.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524, Bangkok, Thailand. Association for Computational Linguistics.

Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Breaking agents: Compromising autonomous LLM agents through malfunction amplification. *Preprint*, arXiv:2407.20859.

Zhendong Zhao, Xiaotian Yue, Jiexin Xie, Chuanhong Fang, Zhenzhou Shao, and Shijie Guo. 2025. A dual-agent collaboration framework based on llms for nursing robots to perform bimanual coordination tasks. *IEEE Robotics and Automation Letters*, 10(3):2942–2949.

A Interview Guide

A.1 Research Framework and Methodology

This guide is grounded in a literature review of LLM agent governance best practices. The interviews will be recorded, transcribed, and stored on secure University servers with access limited to project researchers. All data will be pseudonymized immediately after collection, retaining only minimal demographic information (agency and seniority level). Analysis will utilize coding based on our Agent governance readiness framework to identify gaps and potential, with support from secure LLMs for initial coding. Findings will be reported following COREQ guidelines.

A.2 Introduction for Interviewer

This interview guide is designed to assess current AI governance practices in German public sector organizations and identify gaps in preparedness for implementing LLM-based agents. The interview should take approximately 30 minutes. Begin by introducing yourself and providing context on LLM-based agents before proceeding with the questions.

Interviews will be conducted in German by a native German speaker familiar with the public sector context. Focus on creating a comfortable environment for honest discussion, and adapt questions based on the participant's role and familiarity with AI systems.

A.3 Introduction for Participants

[Read to participant in German]

Thank you for participating in this interview. Today we'll be discussing governance practices for digital technologies and AI in your organization, with a particular focus on how these practices might apply to LLM-based agents.

Your responses will help us understand your organization's governance structures in a neutral and non-opinionated way, focusing on how current governance might need to adapt for the responsible implementation of these technologies. While we are primarily interested in understanding the formal governance processes, we are also keen to hear about your personal experience with these structures, including any challenges or opportunities you have encountered. As such, your personal experience with these processes in your organization are highly valuable.

This interview will last around 30 minutes. It will be recorded and transcribed for research purposes. All information will be pseudonymized or aggregated in the paper, and only the research team will have access to the transcript. The findings will contribute to developing better governance frameworks for AI agents in public sector organizations.

(If relevant) Lastly, I want to acknowledge explicitly that we know each other from the [certificate program]. While this familiarity may shape some aspects of our discussion, I will be following an interview guide here that does not presuppose this.

Before we begin, let me briefly explain what we mean by "LLM-based agents." As we briefly touched upon in the course, these are AI systems built on large language models that can perform tasks with some degree of autonomy. Unlike simple chatbots that just respond to queries, these agents can:

- Execute multi-step processes
- Access and utilize various tools and systems
- Make certain decisions within defined parameters
- Potentially interact with other systems and databases

For example, an LLM-based agent might automatically draft responses to citizen queries about building regulations by accessing relevant databases, interpreting regulations, and formulating appropriate responses.

Do you have any questions before we begin?

A.4 Questions

1. To start, we'd like to gather some basic background information about your role and experience.

- What is your position/role in the organization?
- Seniority level (years in current role/public sector)?
- Do you have experience with digital transformation and/or AI projects in your organization? If yes, what role do you generally take in them?

2. Could you describe how NEW digital projects are currently managed in your organization?

- How are decisions about implementing new digital systems made?
- Are there dedicated teams or individuals focused on digital project governance?
- In what ways, if any, do AI-related projects differ from traditional digital projects in your organization?

3. When implementing a new digital system or tool, what approval processes or oversight mechanisms are typically involved?

- Where in the organization are these specialized departments located, and how do they interact with other units?
- What are the processes for bringing them into a digital project?
- At what stage of implementation do they typically get involved?
- Are there differences in governance processes, oversight, or approval mechanisms between AI and non-AI projects?

4. How does coordination work across departments when implementing digital systems that affect multiple units? Consider, for example, projects involving multiple *Fachabteilungen*, or projects implemented by *internal third parties*, such as centres of excellence.

- What challenges have you experienced with cross-departmental digital initiatives?
- How are responsibilities divided when a system spans multiple departments?
- Are there formal mechanisms for cross-departmental collaboration?
- Which governance skills/tasks are where; if you are in a *Fachabteilung*, are they in your team?

5. What capabilities does your organization have for evaluating the technical performance and safety of AI systems before deployment?

- Who conducts these evaluations?
- What metrics or standards are typically used?
- Do you have the technical expertise internally, or do you rely on external evaluators?
- How do you assess potential risks or failure modes?

6. How does your organization approach ongoing monitoring and security for AI systems after they've been deployed?

- What mechanisms exist for detecting and responding to potential misuse or system failures?
- How are incident response procedures structured?
- How often are deployed AI systems reviewed or reassessed?

7. What procedures exist for documenting and maintaining visibility into how AI systems function and make decisions?

- How is information about AI systems communicated to relevant stakeholders?
- Are there requirements for explainability or transparency?
- How do you ensure citizens understand when they're interacting with AI systems?

8. Considering LLM-based agents that can perform tasks with some autonomy, what aspects of your current governance approach do you think would need to change?

- What new challenges do you anticipate?
- What resources or capabilities would need to be developed?
- How would you approach the question of accountability for agent decisions?

9. How would you assess your organization's current capacity in terms of staff expertise and resources for governing more autonomous AI systems like LLM-based agents?

- Can you provide examples of any existing governance structures or processes that might already be applicable to autonomous AI agents?
- How do you think governance would need to change for AI agents capable of executing multi-step tasks independently, such as processing permit applications or responding to citizen inquiries based on real-time data?
- Where in the organization is this expertise currently?
- What training or skill development would be needed to manage these new governance requirements?
- Are there resource constraints that would affect this capacity?

10. Is there anything else you'd like to share about AI governance in your organization that we haven't covered?

Thank you for your time and insights. Your responses will help us better understand how public sector organizations can adapt governance structures for emerging AI agent technologies.

The findings from this research will be analyzed according to our governance readiness framework and may be used to develop recommendations for public sector organizations. We will follow up with a summary of our research findings once the study is complete. If you have any questions or additional thoughts later, please feel free to contact us.

A.5 Notes for Interviewer

- Adapt questions based on the participant's role and familiarity with AI systems
- If time is limited, prioritize questions 1, 2, 3, 8, and 9
- For participants with limited AI knowledge, provide additional context as needed
- Document any patterns, contradictions, or notable responses for cross-interview analysis
- Immediately after the interview, record your observations about non-verbal cues or contextual factors that might be relevant to interpretation
- Ensure all recordings are transferred to the secure server within 24 hours
- Maintain a research journal documenting methodological decisions and reflections to support rigor

A.6 Post-Interview Protocol

1. Transfer recording to secure University server
2. Complete interviewer reflection form
3. Initiate transcription process
4. Pseudonymize all identifying information
5. Begin preliminary coding using the governance readiness framework
6. Document any emerging themes or patterns to explore in subsequent interviews

B COREQ checklist

Domain 1: Research team and reflexivity

Personal Characteristics

1. **Interviewer/facilitator:** Which author/s conducted the interview or focus group?
   Answer: All interviews were conducted by the first author.
2. **Credentials:** What were the researcher's credentials? *(e.g. PhD, MD)*
   Answer: MSc.
3. **Occupation:** What was their occupation at the time of the study?
   Answer: Doctoral Researcher / PhD Student.
4. **Gender:** Was the researcher male or female?
   Answer: Male.
5. **Experience and training:** What experience or training did the researcher have?
   Answer: Completed course in qualitative fieldwork and effective interviewing.

Relationship with participants

6. **Relationship established:** Was a relationship established prior to study commencement?
   Answer: Yes, participants were recruited via a certificate program organized (but not lectured) by the first author.
7. **Participant knowledge of the interviewer:** What did the participants know about the researcher? *(e.g. personal goals, reasons for doing the research)*
   Answer: Outline of research area and interests.
8. **Interviewer characteristics:** What characteristics were reported about the interviewer/facilitator? *(e.g. bias, assumptions, reasons and interests in the research topic)*
   Answer: Assumptions, reasons, and interests were all presented in both the invitation and beginning of each interview (see Appendix A).

Domain 2: Study design

Theoretical framework

9. **Methodological orientation and Theory:** What methodological orientation was stated to underpin the study? *(e.g. grounded theory, discourse analysis, ethnography, phenomenology, content analysis)*
   Answer: We use content analysis and open-coding grounded in theory of public management to conduct and analyse the interviews.

Participant selection

10. **Sampling:** How were participants selected? *(e.g. purposive, convenience, consecutive, snowball)*
    Answer: We use purposive sampling to select highly relevant interviewees (Palinkas et al., 2015). Still, there was a degree of convenience sampling as the participants volunteered for interviews.
11. **Method of approach:** How were participants approached? *(e.g. face-to-face, telephone, mail, email)*?
    Answer: We invited for interviews both in a face-to-face class as well as in a follow-up email.
12. **Sample size:** How many participants were in the study?
    Answer: 6 participants.
13. **Non-participation:** How many people refused to participate or dropped out? Reasons?
    Answer: None of the volunteers dropped out of the study. Most participants of the certificate course did not volunteer, often due to information security and privacy reasons.

Setting

14. **Setting of data collection:** Where was the data collected? *e.g. home, clinic, workplace*
    Answer: The data was collected in auto-transcribed online meetings.
15. **Presence of non-participants:** Was anyone else present besides the participants and researchers?
    Answer: No.
16. **Description of sample:** What are the important characteristics of the sample? *e.g. demographic data, date*
    Answer: All participants sampled worked in bureaucratic institutions for more than a year and were interviewed between February and March 2025. Further characteristics listed in §5.

Data collection

17. **Interview guide:** Were questions, prompts, guides provided by the authors? Was it pilot tested?
    Answer: We use the interview guide from Appendix A. We piloted the questions internally between the authors.
18. **Repeat interviews:** Were repeat interviews carried out? If yes, how many?
    Answer: No.
19. **Audio/visual recording:** Did the research use audio or visual recording to collect the data?
    Answer: We only collected automatic transcripts from the interviews.
20. **Field notes:** Were field notes made during and/or after the interview or focus group?
    Answer: The interviewer wrote informal field notes during and after each interview.
21. **Duration:** What was the duration of the interviews or focus group?
    Answer: Each interview was approximately 30 minutes.
22. **Data saturation:** Was data saturation discussed?
    Answer: Due to the diversity in institutions surveyed and low N, data saturation was not discussed.
23. **Transcripts returned:** Were transcripts returned to participants for comment and/or correction?
    Answer: Each participant was presented with the transcript immediately after the call, and offered the opportunity to comment, redact, or correct. No correction was requested.

Domain 3: Analysis and findings

Data analysis

24. **Number of data coders:** How many data coders coded the data?
    Answer: The data were read and analyzed by all authors.
25. **Description of the coding tree:** Did authors provide a description of the coding tree?
    Answer: We use open coding based on the literature review in §3.
26. **Derivation of themes:** Were themes identified in advance or derived from the data?
    Answer: Themes were derived from the data informed by our literature review.
27. **Software:** What software, if applicable, was used to manage the data?
    Answer: No special software on the data.
28. **Participant checking:** Did participants provide feedback on the findings?
    Answer: The findings will be shared in future seminars with the participants.

Reporting

29. **Quotations presented:** Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? *e.g. participant number*
    Answer: No, we perform only aggregate analysis to maintain confidentiality.
30. **Data and findings consistent:** Was there consistency between the data presented and the findings?
    Answer: Yes, all presented findings originate from both the literature review and the interviews.
31. **Clarity of major themes:** Were major themes clearly presented in the findings?
    Answer: Yes, see §6.
32. **Clarity of minor themes:** Is there a description of diverse cases or discussion of minor themes?
    Answer: Yes, though elaborate discussion is constrained by the page limit.