Cascading Effects: A Multifaceted Governance Challenge in AI Systems

Anna Neumann
Research Center Trustworthy Data Science and Security
Germany
neumann@rc-trust.ai

Jatinder Singh
Research Center Trustworthy Data Science and Security, Germany
& University of Cambridge, United Kingdom
jat@rc-trust.ai

Abstract

AI-based systems entail combinations of foundation models and other mechanisms, like tools, memory systems, and surrounding prompt scaffolding. This can be described as a chain of interconnected different components, for example AI agents and multi-agent workflows. While these chains commodify building AI applications and systems, they present a key challenge: individual components may exhibit different behaviours when connected to each other than when separated, affecting the properties of the overall system. This means that isolated evaluations and audits of each part do not ensure a safe and reliable overall system. *This paper describes how effects of components interacting can cascade throughout systems and result in evaluation challenges, and a discusses on benefits of cascade-level analyses for (private) governance.*

1 AI Systems as Cascading Socio-Technical Chains

AI-based systems deployments increasingly rely on the combination of interconnected components, i.e. foundation models being combined with other components like guardrails and classifiers, databases being added through retrieval systems, browsing- or programming-based tools or increasingly complex prompt scaffolding techniques. These componentsinteract to form complex chains [1] where outputs from one component become inputs to one or more other ones [2; 3]. Analysing interactions between mechanisms and overall in these complex systems requires frameworks acknowledging these architectures and their effects.

We define *mechanisms* here as socio-technical components within AI-based system that perform specific functions, including RAG systems, system prompts, guardrails, and tool-use frameworks. 'Agents', for example, can be described as chains: they combine foundation models like large language models (LLMs) with 'pools' [4] of tools, examples, memories, and other mechanisms. Multi-agent workflows could also be seen as a kind of chain where singular agents make up one mechanism. Take OpenAI's 'AgentKit' [5], where customers define logic between agents and other tools (see Figure 1).

These chains grow more complex as different mechanisms are controlled by different stakeholders [6], e.g. foundation model developers, application developers, end-users, fitting into broader 'AI supply chains' [7; 1; 8]. An inherent problem in AI supply/value chains [9] is non-modularity: unlike traditional supply chains, AI system mechanisms cannot be evaluated **just** on their own, since the properties of mechanisms can change through connection to other mechanisms [7; 1]. When mechanisms are connected and interact (at *interaction points*), effects arise that neither mechanism would produce alone and that current evaluation paradigms of foundation models and deployed AI-

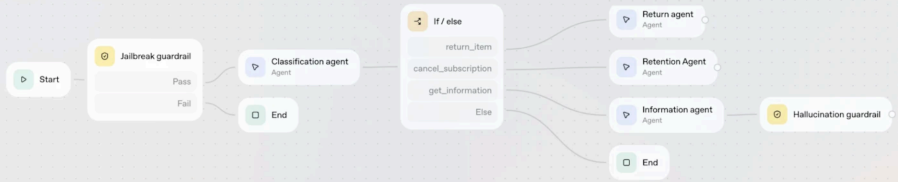

Figure 1: Example workflow of the 'AgentKit' by OpenAI [5]

based applications fail to capture. These effects could be transformations of data or model behaviour, but they could also be monitoring, logging [10], access controls, or reconfigurations that occur as data, instructions, requests, and control flow through the chain [3].

Subsequently, what constitutes 'one mechanism' depends on the observer's (so different stakeholders') level of access and abstraction: within an agent, multiple mechanisms connect, while in a multi-agent framework, one agent may function as one mechanism. Similarly, an external auditor might only have access to an API as one mechanism, while internal audits consider multiple interacting mechanisms.

When unexpected or harmful outcomes arise from these cascades, it can be difficult to assess responsibility, trace harms, or ensure safety; *this requires governance mechanisms that address core challenges of AI cascades*. As AI-based systems become ever more complex and widely deployed, the gap between component-level oversight and cascade-level behaviour will only widen, making cascade-aware governance increasingly urgent. We next set out some key considerations towards this.

2 Requirements and Limitations for Cascade-Level Insight and Oversight

Cascade-level governance needs to grapple with three core challenges: **(i) Non-modularity**, as properties emerge from mechanism interactions rather than summing of individual parts, meaning components cannot be evaluated in isolation [2], **(ii) distributed visibility**, as no single actor sees the full chain [7; 3] (foundation model providers, application developers, and end-users each have partial views and control [8; 11]), and **(iii) varying levels of abstraction**, since a mechanism audited at one access level may consist of several mechanisms that can be audited separately at another.

Current governance mechanisms tend to target either foundation models or specific applications, which can miss intermediate mechanisms and their interaction points.This creates attribution challenges when determining responsibility ('accountability horizon' [7], [11]), inspection limitations as regulators lack access to full mechanism stacks [12; 13], and incident reporting that does not capture cascade-level effects [14]. Generally, we lack tooling to trace cascading effects through multi-mechanism systems, be it through a flow of data or control of said data.

Understanding AI systems as cascades reveals governance opportunities at multiple levels. When organizations recognize how their components interact within larger chains [15], they gain better visibility into system behaviour and potential failure modes. This cascade-aware perspective benefits both public and private governance: regulators can better target intervention points, while industry actors can coordinate around shared interaction standards to better pre-empt and manage issues before they occur and incident response protocols when certain issues do inevitably occur [14].

Governance at any level could benefit from viewing AI-based systems through a 'cascading lens' as it supports understanding of how specific mechanisms interact, where failures originate or propagate, and which *interaction* points and thus *intervention* points matter most. Private governance is well situated here as it can operate closer to the speed and technical specificity cascade-level oversight requires. As a starting point, we flag several mechanisms that could profit from this view:

• **Technical standards and protocols**: Recognition of failure patterns emerging at interaction points,
• **Cooperative frameworks**: Clarity on stakeholder coordination needs based on cascading effects,
• **Contractual measures**: Anticipation of obligation gaps emerging from non-modular interactions,
• **Industry certifications and procurement policies**: Systematic identification, documentation and meaningful communication of mechanism interactions and their effects

3 Conclusion

For better understanding and audits of AI-based systems, *private governance and public governance will benefit from taking cascading effects into account*. We detail three challenges inherent therein that warrant attention: (i) non-modularity, (ii) distributed visibility, and (iii) abstraction levels. We urge the development of cascade-aware governance frameworks, both public and private.

References

[1] Aspen Hopkins, Sarah H. Cen, Andrew Ilyas, Isabella Struckman, Luis Videgaray, and Aleksander Mądry. Ai supply chains: An emerging ecosystem of ai actors, products, and services, 2025. URL https://arxiv.org/abs/2504.20185.

[2] Sarah Huiyi Cen, Aspen Hopkins, Andrew Ilyas, Aleksander Madry, Isabella Struckman, and Luis Videgaray Caso. Ai supply chains. 2023.

[3] Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, 7:6562–6574, 2019. doi: 10.1109/ACCESS.2018.2887201.

[4] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models, 2024. URL https://arxiv.org/abs/2401.02777.

[5] OpenAI. Introducing AgentKit — openai.com, 2025. URL https://openai.com/index/introducing-agentkit/. [Accessed 14-10-2025].

[6] Agathe Balayn, Lorenzo Corti, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. Understanding stakeholders' perceptions and needs across the llm supply chain, 2024. URL https://arxiv.org/abs/2405.16311.

[7] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1186–1197, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. doi: 10.1145/3593013.3594073. URL https://dl.acm.org/doi/10.1145/3593013.3594073.

[8] Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. Position is power: System prompts as a mechanism of bias in large language models (llms). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (Forthcoming)*, FAccT '25, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732038.

[9] Blair Attard-Frost and David Gray Widder. The ethics of ai value chains. *arXiv preprint arXiv:2307.16787*, 2023.

[10] Shreya Chappidi, Jennifer Cobbe, Chris Norval, Anjali Mazumder, and Jatinder Singh. Accountability capture: How record-keeping to support ai transparency and accountability (re)shapes algorithmic oversight, 2025. URL https://arxiv.org/abs/2510.04609.

[11] David Gray Widder and Dawn Nafus. Dislocated accountabilities in the "ai supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620, 2023. doi: 10.1177/20539517231177620. URL https://doi.org/10.1177/20539517231177620.

[12] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3659037. URL https://dl.acm.org/doi/10.1145/3630106.3659037.

[13] David Gray Widder, Meredith Whittaker, and Sarah Myers West. Why 'open' AI systems are actually closed, and why this matters. *Nature*, 635(8040):827–833, November 2024.

[14] Agathe Balayn, Yulu Pi, David Gray Widder, Kars Alfrink, Mireia Yurrita, Sohini Upadhyay, Naveena Karusala, Henrietta Lyons, Cagatay Turkay, Christelle Tessono, Blair Attard-Frost, and Ujwal Gadiraju. From stem to stern: Contestability along ai value chains. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, page 720–723, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711145. doi: 10.1145/3678884.3681831. URL https://doi.org/10.1145/3678884.3681831.

[15] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. Unpacking trust dynamics in the llm supply chain: An empirical exploration to foster trustworthy llm production & use. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2025.