Practices for Governing Agentic AI Systems

Shavit et al., OpenAI (2023)

Context

As AI systems become more autonomous, new accountability challenges arise. When an agentic AI causes harm, responsibility often gets diffused across stakeholders. Yet traditional oversight measures like one-off approvals and static regulations are insufficient, as agentic AI operates across long action sequences where errors compound.

While agentic AI offers many benefits, this OpenAI white paper argues for shared governance standards. It proposes 7 key practices for keeping agentic AI safe and accountable and highlights systemic risks from widespread adoption (e.g. correlated cross-organisational failures).

Key Insights

Systemic risks from widespread adoption

- *Adoption races*: competitive pressure may drive premature deployment of unreliable agents
- *Labour displacement*: agentic AI may automate a broader range of tasks than static AI, unevenly impacting workers and reducing human expertise over time
- *Offense-defence shifts*: AI lowers the cost of certain threats (e.g. fraud) faster than it strengthens defences, disrupting existing security equilibria
- *Correlated failures*: widespread reliance on the same underlying models means systems can fail simultaneously

Seven key governance practices for safe agentic AI

1. *Evaluate task suitability:* test reliability across expected conditions before deployment
2. *Constrain actions and require approval:* keep humans in the loop for high-stakes or irreversible decisions
3. *Set safe defaults:* build in common-sense behaviours *so* agents err toward least-disruptive actions when the user intent is unclear to the AI
4. *Ensure legibility*: expose reasoning traces and action logs so errors can be spotted early
5. *Deploy automatic monitoring*: use secondary AI to review agent actions at scale
6. *Enable attributability*: trace every agent action back to a responsible human party
7. *Maintain interruptibility*: users must always be able to shut an agent down gracefully

As agentic AI diffuses accountability, shared standards are needed to close governance gaps.

Implications

- Do you have a clear framework to classify which AI-driven decisions require human sign-off, based on impact and risk? How does this evolve as systems become more autonomous?
- If an AI agent acting on your behalf causes harm, is accountability clearly allocated and auditable, including when errors accumulate gradually?
- Can you always shut down any AI system operating on your behalf, with a clear fallback plan?
- How do you verify your AI monitoring systems catch errors and agents cannot circumvent their own restrictions?

Shavit, Y., Agarwal, S., Brundage, M. et al. (2023). Practices for Governing Agentic AI Systems. OpenAI White Paper. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf