Agentic AI System

agentic-system · safety · concept

Source: https://policywindow.org/wiki/agentic-system

Generated 2026-05-30T22:07:42 UTC

Summary

An AI system that takes actions in the world — calling tools, executing code, browsing the web, sending messages, planning multi-step sequences — rather than only generating text or images for a human reader.

At a glance

Used by

3 instrument(s)

Related concepts

tool-use-safety, scalable-oversight, alignment, deceptive-alignment, multi-turn-evaluation, prompt-injection

Primary source

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y. (2022), 'ReAct: Synergizing Reasoning and Acting in Language Models.'

Details

An agentic system, in the technical sense, is one whose outputs include actions with external effects (tool calls, API requests, code execution, file writes) and whose loop structure permits multi-step planning over those actions. The architecture pattern emerged with ReAct (Yao et al. 2022, 'ReAct: Synergizing Reasoning and Acting in Language Models'), AutoGPT and BabyAGI (2023, open-source), and is now the deployment substrate for Claude's tool use, GPT's function calling + assistants API, and Google DeepMind's Project Astra demos. The governance-relevant distinction from chat-only LLMs is that agentic systems can cause harm by acting (sending money, running attacks, exfiltrating data) rather than only by saying — Wittgenstein's 'words can wound' becomes 'words and actions can wound, and the actions are at machine speed.'

Regulatory vocabulary has not caught up. EU AI Act treats agentic systems as a sub-case of GPAI plus deployment context, with no agentic-specific obligations. Seoul Declaration (May 2024) and the 16 frontier-lab Frontier AI Safety Commitments mention 'advanced AI systems' but do not operationalise the agentic-vs-chat distinction. UK AISI's evaluations include agentic-capability tests (autonomous-replication, self-exfiltration) that imply the category but do not define it. The G7 Hiroshima Code §1 uses 'advanced AI' as the umbrella. Industry-side frameworks (Anthropic RSP, OpenAI Preparedness, DeepMind FSF) treat agentic capability as a tier-relevant signal: at sufficient action capability, capability-tier safeguards apply that wouldn't apply to a chat-only model with equal knowledge.

How to cite this article

APA

Policy Window. (n.d.). Agentic AI System [Wiki article — Concept]. https://policywindow.org/wiki/agentic-system

CHICAGO

Policy Window. n.d.. "Agentic AI System." Wiki article (Concept). https://policywindow.org/wiki/agentic-system.

HARVARD

Policy Window (n.d.) 'Agentic AI System', Wiki article — Concept, available at: https://policywindow.org/wiki/agentic-system.

OSCOLA

Policy Window, 'Agentic AI System' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/agentic-system> accessed [date].

BIBTEX

```
@misc{policywindow-agentic-system,
title  = {Agentic AI System},
author = {Policy Window},
year   = {n.d.},
howpublished = {agentic-system — safety},
url    = {https://policywindow.org/wiki/agentic-system},
note   = {Primary source: https://arxiv.org/abs/2210.03629}
}
```