Agent Workspaces Face Their First Friction Test as Multi-Agent Competition Broadens

AI News Digest

2026-07-11

Agent Workspaces Face Their First Friction Test as Multi-Agent Competition Broadens

By AI News Digest • July 11, 2026

OpenAI is refining ChatGPT Work after early user feedback while Microsoft brings hosted agent infrastructure to general availability. Elsewhere, model competition is moving toward multi-agent research and orchestrated workflows, alongside new evidence on creative exploration and biosafety testing.

OpenAI's agent workspace gets an early course correction

ChatGPT Work responds to usability and workflow feedback

OpenAI said it is revising the recent ChatGPT Work and Codex integration after users reported confusing usage limits, a desktop reorganization that made chats and projects harder to find, regressions in some multi-agent workflows, and plugin issues. The company reset limits twice, is changing defaults and the model picker, and plans further interface improvements next week; it also emphasized that Codex "is here to stay." [1]

Why it matters: The episode is an early reminder that bringing agents into a shared workspace is as much a product-design challenge as a model-capability challenge. OpenAI's stated goal remains a workspace where people and agents collaborate. [1]

Microsoft makes hosted agent compute generally available

Microsoft Foundry Hosted Agents is now generally available, offering agent-native compute across frameworks, languages, and models. Microsoft highlights an end-to-end setup spanning the GitHub Copilot App, Microsoft IQ, Foundry, Teams, Agent 365 governance, and continuous optimization for long-running agents. [2, 3]

Why it matters: Major platforms are moving beyond standalone model access toward managed environments for building, operating, and governing agents over time.

The capability race extends from coding to mathematics

OpenAI says GPT-5.6 Sol Ultra produced a conjecture proof with 64 subagents

OpenAI said GPT-5.6 Sol Ultra produced a proof of the 50-year-old Cycle Double Cover Conjecture using 64 subagents in just under an hour, and published the prompt and proof. [4]

Why it matters: The claim illustrates the growing focus on coordinated multi-agent systems for difficult research tasks—not only on a single model's response quality. The proof itself remains a claim from OpenAI and would need independent mathematical validation.

Grok 4.5 expands into third-party agent workflows

Perplexity made Grok 4.5 available as an orchestrator model for Consumer Pro and Max subscribers, later extending access to Enterprise organizations. Perplexity says Grok 4.5 scored highest among six tested orchestrator configurations on its internal WANDR agentic-research evaluation, at roughly half the cost of Claude Opus 4.8. [5, 6]

Why it matters: Competitive model comparisons are increasingly being made at the *agent harness* level, where orchestration quality and cost per completed task can matter as much as raw model benchmarks.

Research and safeguards: exploration remains difficult to automate

Sakana study finds diverse agents explore more creatively—but humans still lead

Sakana AI Labs, working with MIT and NYU, recreated the open-ended Picbreeder experiment with vision-language-model agents that evolve images without a target objective. The researchers found agents often revisited similar concepts and made smaller conceptual leaps than humans, while diverse agent personalities substantially improved exploration and, in some runs, approached human semantic diversity. [7]

Why it matters: The findings distinguish task completion from open-ended discovery: diversity in agent populations may help exploration, but the authors say humans remain better at recognizing and extending promising accidents. [7]

OpenAI doubles biosafety jailbreak rewards

OpenAI is turning its Bio Bug Bounty into an ongoing private program and doubling rewards to $50,000. It is inviting qualified researchers to attempt to find a universal jailbreak that defeats predefined biosafety challenges on its frontier models. [8]

Why it matters: As frontier systems are promoted for more capable reasoning and agentic work, testing whether biological safeguards hold up under adversarial pressure is becoming a continuing operational requirement rather than a one-off exercise.

Sources

1. X post by @thsottiaux
2. X post by @jeffhollan
3. X post by @satyanadella
4. X post by @___eknight___
5. X post by @perplexity_ai
6. X post by @AravSrinivas
7. X post by @SakanaAILabs
8. X post by @OpenAI