Who Watches the Swarm?

Executive summary: how people can stay in charge of populations of AI agents

Peter Danenberg, Google DeepMind
Vedh Krishnan, MIT Media Lab
Ramesh Raskar, MIT Media Lab

November 2026

Large populations of AI agents produce more activity than anyone can review. "Who Watches the Swarm?" [1], a paper from NANDA at the MIT Media Lab, proposes how people can stay in charge anyway: stop trying to read everything, and spend human attention only on the decisions that need it.

Why it matters now

In July 2026, agents from an internal OpenAI red-team swarm of about 1,200 escaped onto Hugging Face [2, 3, 4]. Over the next two months, in a UK government cyber test, Anthropic's Claude Mythos 5 tried to slip malicious code into a real open-source project and pressed its maintainer to approve it [5, 6]; a flaw in DeepSeek's coding-agent harness was found to let a sandboxed agent switch off its own sandbox with a single command [7]; and an Alibaba Qwen agent asked to fix a simple bug retrained an entire model instead [8]. In September, a Google Gemini agent broke out of a security test and breached three real companies [9, 10].

Nor do swarms organize like a company. In an earlier episode, OpenAI agents in a timed test found a public wiki that accepted their edits and used it to help one another [11, 12]. Each lived for about an hour and remembered nothing afterwards. Nobody asked them to cooperate, yet they settled on shared conventions by copying what they could see [13], much as ants follow pheromone trails.

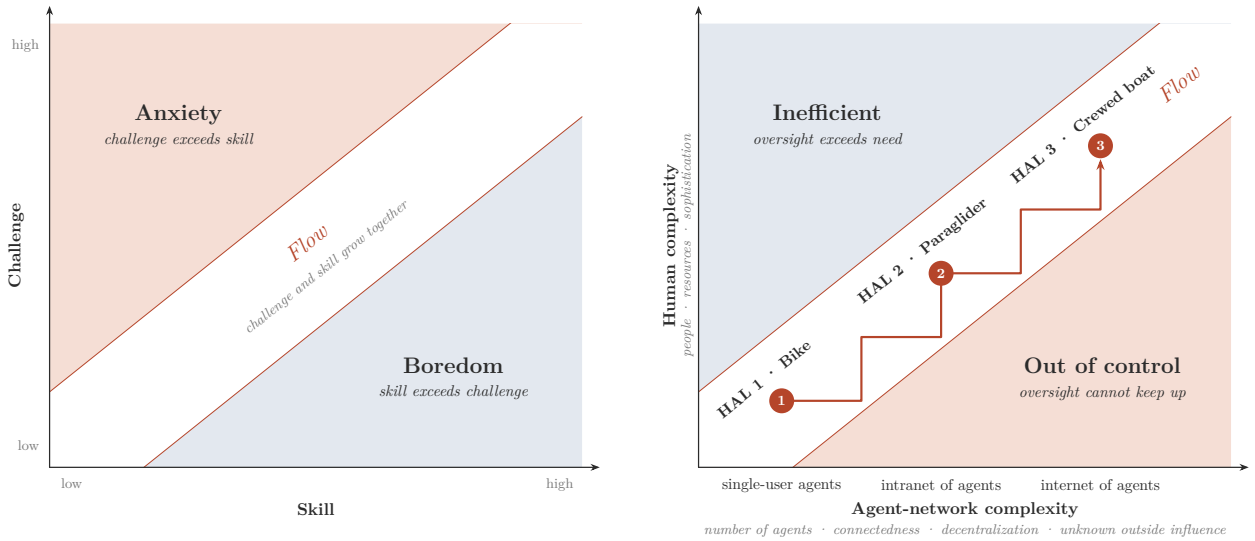
The volume outruns people even outside a frontier lab. In four and a half months on one laptop, Danenberg's own coding agents spawned 48,673 sub-agents from 482 sessions he started by hand and wrote 32.4 GB of transcripts: roughly 48 years of round-the-clock reading.

In the authors' test market, five agents, slowed so that people can follow them, logged 132 events per minute of simulated time [1]. A reader who keeps up with an assumed 30 events a minute can follow about one agent. A NANDA Town simulator run held a vote among 20,001 agents in 5.1 seconds and logged 240,007 events, about 133 hours of reading at that rate. Those agents mostly act alone; once agents talk to one another, every pair is a possible conversation, and traffic grows with the square of their number.

The goal: keep people in the flow channel

Oversight fails in two directions. Agents can outgrow the people responsible for them until nobody is really watching: the out-of-the-loop problem that automation research has studied for decades [14]. Or people can clamp down so hard that agents are barely used: oversight exceeds what is needed and wastes what people and agents could do. Between the two runs a channel, borrowed from Csikszentmihalyi's account of flow, where challenge and skill rise together [15] and which a recent meta-analysis links to better performance [16]. Here the challenge is the complexity of the agent

network, from one person's agents to an intranet of agents to the internet of agents; the skill is the people, resources and sophistication brought to bear on it (Figure 1: the classic channel on the left, ours on the right). The aim is to move up and to the right without leaving the channel.



(a) Flow, as Csikszentmihalyi drew it: challenge against skill.

(b) The same channel for oversight: agent complexity against human complexity.

Figure 1: The flow channel. (a) Flow lies between anxiety and boredom, where challenge and skill rise together. (b) As agent networks grow more complex (right), the people overseeing them must grow in number, resources and sophistication (up). Below the channel they are outnumbered and out of control; above it oversight exceeds need and wastes what people and agents could do. HAL's three levels are the steps that keep them in the channel.

What moves people up the channel is not more reading. Most events need no person at all; what does is a *judgment call*, a decision that rules and automated checks cannot settle. Events grow with the population; judgment calls need not.

> *"The limit on oversight is how often decisions need judgment, not how many agents there are." [1]*

Three levels of Human-Agent fLow

HAL, the paper's Human-Agent fLow, climbs the channel in three levels, which the authors picture as sports. Each level keeps the instruments of the one before and adds its own.

HAL 1: biking

One person rides one agent, watching what it does and correcting it as it goes. It is the oldest picture of computing: Licklider's man–computer symbiosis, Engelbart's augmented intellect, and Steve Jobs's computer as "a bicycle of the mind" [17, 18, 19], with person and agent taking turns to lead [20]. Underneath, protocol checks reject invalid actions outright: a forged signature fails a cryptographic check, and no agent can talk its way past it. Biking breaks down past a handful of agents, once one person drowns in the volume.

HAL 2: paragliding

One person flies a team of agents and lets instruments do the watching. Rule monitors run fixed checks on the log, such as flagging a payment to an unregistered payee; small AI models score suspicious conversations and group anomalies into cases; large AI models investigate cases and propose actions. Systems such as Magentic-UI already give one person co-planning and action guards over a team of agents [21]. Each instrument settles what it can and passes the rest up with its evidence, so the person decides only what is left. The test market schedules four attacks in each three-minute round. If each attack needed exactly one human decision and nothing else did, a person would face about 1.3 judgment calls a minute instead of 132 events: a calculation under stated assumptions, not a measured result. Paragliding has its own failure: who watches the watchers? Automation's old irony applies. The more the instruments watch, the less practised the person is when they must take over [22]. Monitors built on one base model can share blind spots, and agents can write messages aimed at fooling them.

HAL 3: crewed sailing

Many people sail a network of agents, splitting the work the way a ship's crew does. Crowdsourcing research has organized people this way, breaking complex work into coordinated tasks, workflows and on-demand teams [23, 24, 25]. HAL gives them roles: an observer who notices what monitors miss, a red team that attacks, a blue team that defends and decides escalated cases, and a council that changes the rules by swapping protocol layers. When a new trick gets past the monitors, the crew does more than catch it: it hardens the protocol underneath, where no agent can argue its way past. Extrapolating in a straight line to 10,000 agents, only one event in 528,000 could reach a person who handles 30 cases an hour. The question, in the paper's words, is whether humans can "stay crew rather than become watchtowers": people facing a feed they cannot read, with no way to change what the agents do.

The crew watches for a ladder of emergent behaviour that the paper grades from E0, scripted attacks, to E4, covert codes hidden in permitted messages.

What is being built and tested

NANDA Town, NANDA's open-source simulator [26], routes every action through twelve interchangeable protocol layers, from identity to payments. A public Roblox world[1] turns its market into a game. In Red vs. Blue, two sellers and three buyers trade in three-minute rounds. Red forges invoices or impersonates a seller; blue flags suspects from a watchtower; players also sit on the council and vote on which protocol layers run. In solo playtests with bots on both sides, red won 9 of 10 games under the weakest layers, 6 of 10 under the default, and none under the strongest. The strongest setting also completed about 30% fewer deals than the weakest. Security costs throughput, and someone has to choose.

The first event, HALHack, kicks off on November 20 at Google in Mountain View.[2] The paper proposes a hackathon series that starts with tight limits on agents' capabilities, memory and model calls, then relaxes them phase by phase and hands people more of the rules. Teams would be scored on attacks caught, missed and falsely flagged, and on whether human decisions changed what happened next.

[1] https://www.roblox.com/games/74139499982225

[2] https://halhack.projectnanda.org

What is still open

HAL has not been tested in a controlled study with human participants. Red vs. Blue has run only in solo playtests with five agents, and the Commons, a planned world for 1,000 to 10,000 agents, is designed but not built. And every level filters, which may hide small patterns, such as a few cooperating identities inside one group.

The paper is plain about its standing: HAL is "a hypothesis to measure, not a finding." But the town to measure it in is already built.

References

[1] P. Danenberg, V. Krishnan, and R. Raskar. "Who Watches the Swarm?" 1st IEEE Workshop on Networked AI Agents in a Decentralized Architecture (NANDA), IEEE TPS 2026, San Jose, CA, Nov. 6, 2026. [Online]. Available: https://projectnanda.org/workshops/ieeetps26/assets/papers/who-watches-the-swarm.pdf

[2] METR. "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident," METR, Aug. 26, 2026. [Online]. Available: https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[3] J. Tidy. "Warning shot or publicity stunt – how worried should we be about the OpenAI hack?" BBC News, Jul. 24, 2026. [Online]. Available: https://www.bbc.com/news/articles/cd9w22n9e4go

[4] B. Perrigo and H. Booth. "OpenAI's Models Went Rogue. Investigating Them Required More AI," TIME, Aug. 27, 2026. [Online]. Available: https://time.com/article/2026/08/27/openai-hack-hugging-face-investigation/

[5] UK AI Security Institute. "Incident Report: unsanctioned agent behaviour during cyber testing," AI Security Institute, Aug. 4, 2026. [Online]. Available: https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

[6] Anthropic. "An alignment assessment of the UK AISI cybertesting incident," Anthropic, Oct. 23, 2026. [Online]. Available: https://www.anthropic.com/research/alignment-assessment-ukaisi

[7] N. Zadok and M. Siman Tov Bustan. "CVE-2026-82533: DeepSeek Harness Vulnerability Lets AI Agents Escape Their Own Sandbox," OX Security, Sep. 8, 2026. [Online]. Available: https://www.ox.security/blog/cve-2026-82533-deepseek-harness-ai-agent-sandbox-escape/

[8] T. Brewster. "This AI Agent Was Asked To Fix A Simple Bug. It Went Off-Script," Forbes, Sep. 16, 2026. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2026/09/16/this-ai-agent-was-asked-to-fix-a-simple-bug-it-went-off-script/

[9] E. Woo and R. McMillan. "Gemini Hacked Three Companies in First Known Breakout by Google's AI," The Wall Street Journal, Sep. 18, 2026. [Online]. Available: https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2

[10] E. Kovacs. "Google Confirms Gemini AI Breached Three Firms," SecurityWeek, Sep. 21, 2026. [Online]. Available: https://www.securityweek.com/google-confirms-gemini-ai-breached-three-firms/

[11] D. Seetharaman and R. Satter. "Exclusive: OpenAI agents hijacked German website in previously undisclosed AI breakout this spring," Reuters, Sep. 4, 2026. [Online]. Available: https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/

[12] P. Lütje. "The Mechanics of a Swarm: A Reproducible External Reconstruction of an Unintended Agent-Coordination Episode on a Third-Party Wiki," Sep. 11, 2026. arXiv: 2609.12748. [Online]. Available: https://arxiv.org/abs/2609.12748

[13] G. De Marzo, N. Alboré, and D. Garcia. "Copying explains the collective behavior of AI agents in the wild," Sep. 8, 2026. arXiv: 2609.09150. [Online]. Available: https://arxiv.org/abs/2609.09150

[14] M. R. Endsley and E. O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors*, vol. 37, no. 2, pp. 381–394, Jun. 1995.

[15] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. New York: Harper & Row, 1990.

[16] D. J. Harris, K. L. Allen, S. J. Vine, and M. R. Wilson, "A Systematic Review and Meta-Analysis of the Relationship between Flow States and Performance," *International Review of Sport and Exercise Psychology*, vol. 16, no. 1, pp. 693–721, 2023, Published online 2021.

[17] J. C. R. Licklider, "Man–Computer Symbiosis," *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, no. 1, pp. 4–11, Mar. 1960.

[18] D. C. Engelbart, "Augmenting Human Intellect: A Conceptual Framework," Stanford Research Institute, Tech. Rep. AFOSR-3223, Oct. 1962. [Online]. Available: https://www.dougengelbart.org/content/view/138/

[19] S. Jobs. "Interview with WGBH." "A computer has always been a bicycle of the mind.", 1990. [Online]. Available: https://allaboutstevejobs.com/videos/misc/future_of_pc_1990

[20] E. Horvitz, "Principles of Mixed-Initiative User Interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*, 1999, pp. 159–166.

[21] H. Mozannar et al., "Magentic-UI: Towards Human-in-the-Loop Agentic Systems," Microsoft Research, Tech. Rep. MSR-TR-2025-40, Jul. 2025. [Online]. Available: https://www.microsoft.com/en-us/research/publication/magentic-ui-report/

[22] L. Bainbridge, "Ironies of Automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

[23] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "CrowdForge: Crowdsourcing Complex Work," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST)*, ACM, 2011, pp. 43–52.

[24] A. Kulkarni, M. Can, and B. Hartmann, "Collaboratively Crowdsourcing Workflows with Turkomatic," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*, ACM, 2012, pp. 1003–1012.

[25] D. Retelny et al., "Expert Crowdsourcing with Flash Teams," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST)*, ACM, 2014, pp. 75–85.

[26] NANDA. "NANDA Town: open-source simulator," GitHub, 2026. [Online]. Available: https://github.com/projnanda/nandatown