Work Agents Learn the Office

2026-07-10 / 00:23:23

> *"The agent product is becoming less like a chat box and more like a temporary colleague with file access, app access, memory, and a deadline."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

OpenAI and Meta both spent Thursday showing agents that can work across files, apps, tools, and long projects. Today's episode follows the product surface first, then the policy and research work trying to catch up to agents that no longer stay inside a single chat window.

- OpenAI's ChatGPT Work demo showed finance analysis, local files, browser tabs, app control, and shareable hosted sites as one work surface.
- OpenAI's GPT-5.6 model demo introduced programmatic tool calling, higher reasoning levels, and subagent delegation in Codex.
- Meta's Muse Spark 1.1 launch post framed its new model around a million-token context window, multi-agent orchestration, computer use, and the public Meta Model API.
- Techmeme's Financial Times summary reported OpenAI and Google serving advanced models to Singapore-based subsidiaries of Chinese tech companies, a concrete update to this week's model-access debate.

- The European Commission preliminarily found Facebook and Instagram in breach of the Digital Services Act over addictive design features including infinite scroll, autoplay, notifications, and recommender systems.
- The harness-engineering paper and the context-graph paper gave the research counterpoint: agents need contracts, source boundaries, monitors, and event-driven context before they can be trusted with enterprise work.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   OpenAI announced ChatGPT Work on Thursday as a new agent in ChatGPT powered by Codex and GPT-5.6. The company's tweet says it can take action across your apps and files, stay with a project for hours if needed, and turn a goal into finished work. That sentence is more literal than I expected. The demo centers on app access, file access, local desktop context, and a model that keeps moving after the first answer.

   - youtube.com
   - x.com
   - youtube.com
   - theverge.com
   - x.com
   - ai.meta.com

- siliconangle.com
- techmeme.com
- techmeme.com
- digital-strategy.ec.europa.eu
- theguardian.com
- theguardian.com
- arxiv.org
- arxiv.org
- aljazeera.com

2. Damra Vol 00:00:33

   The phrase that caught me was "stay with a project." That's a different promise from "answer this question." It says the model can hold a work object in its head long enough to gather context, make artifacts, revise them, and send them somewhere. The office app becomes the environment, not the destination.

3. Lenar Kess 00:00:52

   The launch stream opened with the model family: GPT-5.6 Sol rolling out to paid plans, with Terra and Luna coming to free users. OpenAI then moved quickly into product surfaces, including ChatGPT Work on web and mobile, a new desktop app, and hosted sites for paid users. In the short model video, OpenAI said GPT-5.6 is generally available and better at staying on track with ambiguous prompts. Then it showed Codex building a card game from a short prompt and adding levels, artwork, and a soundtrack.

4. Damra Vol 00:01:27

   And the model video gave the builder detail. Programmatic tool calling lets the model call sandbox tools without round-tripping every tool result through the context window. Then there are reasoning levels above Extra High, plus subagent delegation where the model decides for itself when parallel work helps. That claim is specific: subagents exist, and the model has been trained to choose when to use them.

5. Lenar Kess 00:01:54

   Simon Willison picked up the same detail in his note. His tweet said GPT-5.6 includes interesting API additions, especially programmatic tool calling and multi-agent support. Then, because Simon remains Simon, he also mentioned eighteen pelicans for six reasoning levels and three new models. [chuckle] I appreciate a launch taxonomy that requires bird accounting.

6. Damra Vol 00:02:18

<laugh-speak>The pelicans are helping morale.</laugh-speak> But the API addition matters because it turns the agent loop into something less chatty. If each tool result has to be copied back into the conversation, the model spends attention on bookkeeping. If the tool calls can be programmatic inside the sandbox, the project can feel more like a running process and less like a transcript with chores attached.

7. Lenar Kess 00:02:42

   The product demo stayed deliberately ordinary. OpenAI showed a finance workflow built around a forecast update. The agent looked across systems and reconciled an Excel model. It then proposed a new case, built a PowerPoint deck, made a hosted site with the same analysis, and sent the site link in Slack. The demo numbers were fake. The workflow wasn't. Someone in finance closes June, sees a forecast miss or beat, and needs to explain the drivers before the room moves on.

8. Damra Vol 00:03:12

   That choice was smarter than another tiny game demo. Finance has permission anxiety baked in. Spreadsheets, decks, Slack, calendar, forecast assumptions, and executive context all collide in one place. If ChatGPT Work can touch that surface, it's asking to be trusted with the messiest layer of office life: the layer where numbers become an explanation someone else can act on.

9. Lenar Kess 00:03:37

   Then the desktop app made the boundary more explicit. OpenAI showed it reading a local spreadsheet of support tickets and making an interactive visualization. It also looked at a folder with a launch-readiness PDF, user research interviews, a security review, and open Chrome tabs. After about ninety seconds, it produced a ready slide deck in a company template. The presenter said it used memory too, because some concerns had been discussed in ChatGPT but had not been shared through Slack or email.

10. Damra Vol 00:04:08

    That's the moment where the demo gets less cute. Memory is no longer a convenience feature. It becomes part of the evidence base. The agent is allowed to combine a PDF, browser tabs, a security review, and private conversational residue. The user has to ask something harder than "can it make slides?" It's "do I know which sources shaped this deck, and can I remove one?"

11. Lenar Kess 00:04:32

    The most startling demo for me was Apple Notes. The presenter dragged the agent into a messy notes app and asked it to make folders and move notes around. The agent got its own cursor and started operating the app in the background while the human kept using the computer. That's a

small domestic horror and a pretty plausible feature. Most people do have one app that is just a storage unit with a search bar.

12. Damra Vol 00:04:56

    It also makes the permission model visible in a way a text answer never does. A chat response can be wrong and still be contained. A cursor moving notes around changes the state of your machine. You can love that feature and still feel the new obligation: every background action needs an undo story, a preview story, or a record someone can inspect later.

13. Lenar Kess 00:05:19

    OpenAI also leaned into hosted sites. The demo showed dashboards, internal tools, launch pages, and little interactive educational scenes generated as answers. The model could make a visualization and publish it with one click so other people see the same interactive artifact. That part of ChatGPT Work looks less like a smarter assistant and more like an artifact factory.

14. Damra Vol 00:05:43

    The share button changes the standard. A generated chart inside your own chat is a draft. A hosted site shared with a team becomes a work product. The standard for correctness changes when the output leaves your private session and starts informing other people's decisions.

15. Lenar Kess 00:05:59

    The Verge report placed this release after the earlier political access drama around GPT-5.6. We covered model access this week, so the update here is narrower: the access question now attaches to a work agent, not just a model endpoint. Who can use Sol is one question. Who can give Sol local files, business apps, memory, and hours-long tasks is a much more operational question.

16. Damra Vol 00:06:24

    And the customer is no longer buying only intelligence per token. They're buying a working relationship with the software around the model. The file picker, the desktop app, the app connectors, the hosted-site surface, and the audit trail all become part of whether the model feels usable or reckless.

17. Lenar Kess 00:06:43

    Meta published Muse Spark 1.1 on Thursday and described it as a multimodal reasoning model built for agentic tasks. The launch post says it has major gains in tool and computer use, coding, and multimodal understanding, and it is available through a new Meta Model API in public preview.

That makes it a direct answer to the same demand OpenAI is chasing: agents that can plan, use tools, operate apps, and keep enough context to finish a project.

18. Damra Vol 00:07:12

    Meta's launch reads like an attempt to prove the whole stack at once. The model has a one million token context window, and Meta says it can retrieve information from much earlier in a task while compacting work so later steps keep the pieces they need. It can act as a main agent, make a plan, and delegate to parallel subagents. It can also serve as a subagent and know when to escalate back.

19. Lenar Kess 00:07:37

    That symmetry is important. A lot of agent demos treat subagents as disposable workers. Meta is saying Muse Spark can play both roles. The main agent gathers context and delegates. The subagent sticks to the job, uses tools, and returns when it should. If that works, the orchestration becomes less brittle because every participant has some sense of its own lane.

20. Damra Vol 00:08:00

    The SiliconANGLE piece put the coding claim in plainer terms. In Meta's internal demo, Muse Spark built a chat app and took automated screenshots. It found user-visible failures, traced them back to relevant code, implemented fixes, and validated the changes. That's a very agent-shaped coding loop: see the interface, connect the visual bug to source, patch, and check again.

21. Lenar Kess 00:08:24

    Meta also made the computer-use claim concrete outside coding. The launch post describes a Facebook Marketplace agent that takes smartphone video, extracts useful product photos, reasons about the item, operates the browser, and makes a listing. The example is mundane, but it gets at the product line Meta is trying to draw: multimodal perception plus browser action plus a user goal.

22. Damra Vol 00:08:50

    And then there is the compute story sitting next to it. Techmeme summarized Reuters reporting that Meta plans to start manufacturing its in-house AI chip, Iris, from September, as part of a plan to boost computing power to fourteen gigawatts in 2027. SiliconANGLE tied that to the Meta Model API and the possibility of broader enterprise offerings. Model, API, custom silicon, and data center capacity: Meta wants the agent story to look vertically supplied.

23. Lenar Kess 00:09:21

    My caution is simple: the capability story is still mostly launch material. Meta gives partner quotes, internal evaluations, and benchmark numbers, including a big jump on Vibe Code Bench and a

higher score on SWE-Atlas Codebase Q&A. That's useful evidence, but the independent reports will matter more: people running Muse Spark in Cline, OpenCode, Replit-style workflows, and their own app-and-file messes.

24. Damra Vol 00:09:50

    Yes. I am less interested in whether the launch chart beats another launch chart than whether context compaction works when the project has false starts. Long agent work is full of discarded approaches, renamed files, half-fixed tests, and instructions from an hour ago that are now wrong. A million-token window helps, but the model still has to know which memories have expired.

25. Lenar Kess 00:10:14

    Today's OpenAI and Meta stories rhyme without becoming the same story. Both companies are selling persistence. OpenAI shows it through ChatGPT Work, desktop context, local files, and hosted artifacts. Meta shows it through context management, multimodal computer use, subagent roles, and an API surface. The model race is now also a race to define the work environment around the model.

26. Damra Vol 00:10:41

    And the work environment is where switching costs live. A better model can arrive next month. A project history, a permission graph, saved sites, connectors, memories, and team habits are harder to move. That's why these product surfaces feel heavier than ordinary model announcements.

27. Lenar Kess 00:11:01

    Techmeme summarized a Financial Times report this morning saying OpenAI and Google are selling advanced AI models to Singapore-based subsidiaries of Alibaba, Baidu, and Tencent. The report says that exposes a gap in United States AI controls. After this week's earlier debate about GPT-5.6 access, this is the concrete update: restrictions aimed at national entities run into multinational corporate structure.

28. Damra Vol 00:11:29

    Subsidiaries are where policy language meets corporate reality. A model provider can ask where the account is incorporated. A government can ask where the parent company sits. The data, users, staff, and control rights may all point in different directions. That isn't an edge case in global tech. That's normal corporate architecture.

29. Lenar Kess 00:11:50

The same news set had China expanding its anti-sanctions toolkit, CXMT pushing into memory with state support and domestic sourcing, and Korea trying to bring overseas technical talent back through a K-Tech Pioneers program. I wouldn't collapse those into one mega-story. They're different levers. But together they show why AI controls keep escaping any single instrument. Model access, chip supply, sanctions response, memory manufacturing, and talent policy move through different channels.

30. Damra Vol 00:12:21

    The Singapore item is the one to hold onto because it names the mechanism. Controls can say "advanced models should not reach this category of actor." Companies then have to decide whether a subsidiary is that actor, whether use is local, and whether contractual limits survive the parent relationship. That's less cinematic than a chip ban and much closer to how access decisions get made every day.

31. Lenar Kess 00:12:46

    This also changes how to read OpenAI's and Meta's agent launches. A model that answers a prompt is already sensitive. A work agent that connects to files, apps, local desktops, and company memory carries more operational knowledge. Export control was already hard when the product was a model. It gets stranger when the product is an agent embedded in work systems.

32. Damra Vol 00:13:09

    And enforcement gets pulled down into product details. Which connector is allowed? Which file classes can be touched? Does the agent store project memory? Can a foreign subsidiary connect its internal docs? The policy fight becomes a settings page, a sales approval workflow, and a log review.

33. Lenar Kess 00:13:28

    The European Commission said today that it preliminarily finds the addictive design of Instagram and Facebook in breach of the Digital Services Act. The investigation names infinite scroll, autoplay, push notifications, and highly personalized recommender systems. It also says Meta didn't adequately assess risks to physical and mental wellbeing, including for minors and vulnerable adults.

34. Damra Vol 00:13:52

    That source is interesting because it treats design choices as regulated behavior. Infinite scroll and notifications aren't new technologies. They're interface decisions that shape attention. The Commission is saying those design mechanics have to be assessed for harm, and existing mitigation didn't satisfy the regulator.

35. Lenar Kess 00:14:12

    Henna Virkkunen, the Commission's executive vice-president for tech sovereignty, security, and democracy, said protecting the physical and mental health of Europeans must be a priority for social media platforms, and that the Digital Services Act provides a framework to hold platforms accountable for the addictive design and effects of their services. I am quoting that because it is plain about the target: content moderation and illegal material are only part of it; the mechanics that keep people moving through the product are in scope too.

36. Damra Vol 00:14:44

    It also sits awkwardly next to Meta's agent launch in a good way. Meta is saying its models can help you pursue goals and take action on what you value. European regulators are saying some Meta products have been too effective at pulling people through feeds. Both claims are about agency. One is personal agency helped by a model; the other is user agency weakened by interface design.

37. Lenar Kess 00:15:09

    In the United States, the Guardian reported that Senator Ed Markey unveiled an AI accountability agenda, including a coming bill for AI data centers. Companies that own or propose those facilities would need Federal Communications Commission certification that construction won't harm the public interest. The draft would examine air and water quality, noise, energy costs, grid reliability, local ecosystems, the local economy, and jobs.

38. Damra Vol 00:15:36

    That's a broad lane, and some of it will get stuck in Congress. But the list is useful because it refuses to treat AI accountability as model safety alone. Automated hiring, child chatbot dependence, bias audits, human override in healthcare, worker protections, and data center effects all sit in Markey's package. The regulator's unit of concern is the deployed system, not the leaderboard.

39. Lenar Kess 00:16:02

    The Guardian also reported that Facewatch, a facial recognition system used by more than one hundred UK businesses including Sainsbury's, B&M, and Spar, plans a feature that alerts police in an average of four seconds when a serious offender triggers a live match. Civil liberties groups warned that this moves far ahead of regulation, and the story notes false identifications and evidence that Black and Asian people are more likely to be incorrectly identified than white people.

40. Damra Vol 00:16:30

    That story is the deployed-AI version of "the settings page is policy." A shop, a private security vendor, and police can create a live intervention loop without the shopper doing anything in that

moment. Facewatch argues it is focused on the highest-risk repeat offenders and retail worker safety. Civil liberties groups see a private blacklist plugged into law enforcement. Both sides are talking about the same four-second path.

41. Lenar Kess 00:16:59

    I would keep the regulation block separate from the launch block rather than forcing one grand argument. OpenAI and Meta are showing agents that can act. Regulators are also looking at systems that act on people: feeds, hiring tools, data centers, and face recognition alerts. Those are neighboring stories, but the details carry more than a tidy umbrella.

42. Damra Vol 00:17:21

    The shared pressure is action. Who acts, with what permission, against which evidence, and how quickly can the affected person contest it? That question applies to a desktop agent moving your notes and to a store camera alerting police. The consequences are wildly different, so the rules can't be copy-pasted across them.

43. Lenar Kess 00:17:42

    The arXiv batch today had two papers that fit the product news almost too neatly. One is called From Prompts to Contracts, about harness engineering for auditable enterprise agents. The other argues for proactive agents built on context graphs. Neither paper proves the OpenAI or Meta products work. They help name the engineering problem those products create.

44. Damra Vol 00:18:06

    The harness paper is the one I would hand to anyone building the compliance layer of a work agent. It moves behavior out of prompts and into code-owned artifacts: source boundaries, entity routing, answer contracts, reproducible traces, schemas, and validation checks. The model composes language at a replaceable boundary; the surrounding system decides what claims are allowed to enter the answer.

45. Lenar Kess 00:18:32

    The abstract has a sharp result. Across three hosted models, the harness checks passed on all 270 composition-boundary runs. Failures were confined to the model-composed side and were caught and recorded. In an ablation, prompt instructions alone let recommendation-language and internal-trace leakage violations reach the reader. A bolt-on external filter blocked violations too, but over-refused and dropped utility, while the harness preserved full utility in that test.

46. Damra Vol 00:19:03

Work agents need exactly that distinction. You don't want the model to remember, by vibes, that it shouldn't expose internal trace fields or make a forbidden recommendation. You want the application to make those states unreachable or to stop the answer before it reaches a person. The model is talented. The model isn't the whole product.

47. Lenar Kess 00:19:25

    The context-graph paper starts from a different complaint: enterprise agents are reactive. They wait for a human to ask. The authors propose a live graph of enterprise entities, relationships, and state transitions. Around that graph, they add a delta detector, a proactivity scorer, and a surfacing layer that sends grounded notifications. Their evaluation claims Precision at five of 0.83, a false-positive rate of 0.11, and mean time to surface dropping from forty-seven minutes to under thirty seconds across three generic enterprise cases.

48. Damra Vol 00:20:01

    That paper is early, and the examples are generic, but the architecture fits the week. ChatGPT Work says an agent can stay with a project for hours. A context graph asks how the agent knows something changed while it was with the project. Contract status, incident dependency, idle sales deal, overdue ticket: those are graph events before they are chat prompts.

49. Lenar Kess 00:20:25

    Put the two papers together and you get a research answer to the product demo. If agents can touch files, apps, browser tabs, memory, and hosted artifacts, then the surrounding system needs source eligibility and event detection. It also needs persona fit, validation, trace records, and a way to suppress low-value notifications. Otherwise the agent becomes either too passive to matter or too active to trust.

50. Damra Vol 00:20:50

    And the most interesting work may be deciding when silence is correct. A proactive agent that surfaces everything becomes another inbox. A work agent that never interrupts misses the promise of knowing the project state. The craft is ranking the moment when the user would have wanted to know, before they knew to ask.

51. Lenar Kess 00:21:10

    One last update, and I am keeping it compact because we have covered compute constraints heavily this week. Al Jazeera reported that SK Hynix raised 26.5 billion dollars ahead of a Nasdaq listing, selling 177.9 million American depositary shares at 149 dollars each. The report says it is the largest United States debut by a foreign company, passing Alibaba's 2014 IPO.

52. Damra Vol 00:21:37

    This is finance news, and it is also memory news. SK Hynix is one of the companies sitting closest to the demand for advanced memory chips used in AI systems. Al Jazeera quoted a Pepperstone strategist saying the AI memory cycle, the earnings, and the global capital access are what the pricing revealed. The market is rewarding the supply chain, not only the model labs.

53. Lenar Kess 00:22:01

    The article also says Bloomberg sources described the listing as more than seven times oversubscribed, with about 171 billion dollars in orders for a 24 to 28 billion dollar deal. That came during a rough week for semiconductors and the Korean market. So I wouldn't read this as a blessing for every AI-adjacent stock. It's that investors are hunting for the pieces of the buildout that look like bottlenecks.

54. Damra Vol 00:22:26

    And memory is a lovely closing place for today because both meanings are in play. Agents need memory in the software sense: project history, context compaction, user concerns, and traces. They also need memory in the hardware sense: chips, supply chains, and capital. Friday's story is that the work agent wants both, and neither one is merely a model feature.

55. Lenar Kess 00:22:51

    Friday's release week ends in a fairly concrete place. GPT-5.6 and Muse Spark aren't only faster answer machines. They're attempts to put models inside ongoing work, with files, apps, context, and delegated action around them. The next evidence should be independent work logs where the agent touches a messy project, makes reversible changes, cites the sources it used, and leaves a trace someone else can audit. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic