The Assistant Asked for the Same Door

2026-07-16 / 00:29:15

> *"The phone decides whether a rival assistant is an assistant or an app asking for favors."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Europe is forcing a concrete test of AI competition: can a rival assistant reach the same Android capabilities and search data that make Google's own products useful?

- The Verge on the EU's Google decisions reports the new Android interoperability and Search-data obligations, where implementation will decide whether comparable access changes the market.
- The European Commission's Android case summary details access to invocation, ambient sensors, on-device actions, and first-party services that rival assistants currently lack.
- CNBC on TSMC's second quarter connects a 77.4 percent profit increase with higher spending plans, while the reported US fab expansion remains a pledge rather than finished capacity.
- Nvidia's Jetson Thor announcement pairs new edge computers with the four-billion-parameter Cosmos 3 Edge model, making the company's robotics offer a hardware-and-software package.

- The Verge on xAI's lawsuit reports the company's allegations against a Grok user and raises the unsettled role of civil litigation in severe model misuse.
- The Action Rebinding paper shows how an Android app with no sensitive permissions can redirect a GUI agent between observation and action.
- The CAVA paper proposes canonical action objects and attestations so approvals can be bound to what an agent executes across different runtimes.
- Thinking Machines Lab's Inkling release supplies the model architecture, training scale, multimodal inputs, and open weights that were missing from the day's secondary coverage.
- Axios on frontier-lab regulation compares the different institutions proposed by Demis Hassabis, Sam Altman, and Dario Amodei while a disclosed political contribution makes the policy effort tangible.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

The European Commission today issued two Digital Markets Act decisions telling Google to give rival AI assistants comparable access to Android and to provide competing search products access to some Google Search data. The Verge reported the decisions this morning. Picture the phone in your hand for a second: one assistant can be summoned from the system controls, see what is on screen, reach the camera and microphone with low friction, and act inside other apps. Another assistant arrives as an ordinary download and has to ask permission at every doorway. Europe is trying to make those doors comparable. Later, we've got TSMC's profit surge and Nvidia's new robotics package. We'll also get to an unusual xAI lawsuit, two papers about agent actions, and Thinking Machines Lab's first model. Android comes first because the order names the actual interfaces.

- theverge.com
- digital-markets-act.ec.europa.eu
- cnbc.com
- techmeme.com
- techmeme.com
- techmeme.com
- blogs.nvidia.com
- cnbc.com
- theverge.com
- arxiv.org
- arxiv.org
- youtube.com
- siliconangle.com
- forbes.com
- thinkingmachines.ai
- thinkingmachines.ai
- axios.com
- techmeme.com

2. Damra Vol 00:00:57

And the word comparable has teeth only if the rival can behave like a system assistant. The Commission's Android case summary says Google's own services get integrated access that third-party assistants don't. It covers wake words and system-wide invocation, then turns to contextual screen data and ambient inputs from the camera and microphone. It also covers actions inside installed apps and integration with Google services. That's a very different product from a chatbot sitting in its own window. If I can say, 'take the address from this conversation, put it in the calendar, and send the photo to Mara,' the assistant needs permissions and context. It also needs action channels that cross several boundaries. Today, Google controls many of those boundaries.

3. Lenar Kess 00:01:45

   The Search decision is separate. The Commission's earlier specification work described anonymized ranking, query, click, and view data, supplied on fair, reasonable, and nondiscriminatory terms. It also raised whether AI chatbots with search functions qualify for access. That data isn't a copy of Google's index. It is feedback about how people search and what they choose, which competing search products can use to improve ranking. The Verge's report says the new decision now requires access to some of that material. Google still gets to implement the system. Privacy rules and price will affect who can use it, while update frequency and data quality will affect whether it helps. So today's order creates an obligation. It hasn't yet created a thriving market.

4. Damra Vol 00:02:33

   The Android side feels more immediate because you can name the missing gesture. The Commission's case summary says a third-party assistant should be able to use a custom wake word or a long press on the home control, receive the context needed to translate or search what's on screen, and discover the action interfaces exposed by other apps. A person doesn't experience that as interoperability. They experience it as whether the assistant can finish the request without handing them four permission dialogs and an apology. Google has spent years making Gemini's path through Android feel native. A formal right to the same category of access still leaves a lot of engineering between the right and the feeling.

5. Lenar Kess 00:03:15

   Google has a defensible concern here too. Broader access to microphones, cameras, screen content, messages, and app actions increases the number of companies that can mishandle intimate data or trigger an unwanted action. The Commission's proposal answers part of that by asking for equivalent consent flows and privacy indicators, according to its case summary. Equivalent doesn't mean absent. It means Google can't make its own assistant smooth by default and make every rival look suspicious through extra friction. I think that is the contest the decision sets up: which prompts protect a person, and which prompts protect the incumbent's product advantage? Those can look identical in a settings screen.

6. Damra Vol 00:03:58

   There is also a strange creative consequence. If access becomes usable, the most interesting rival may be a narrow assistant rather than another universal one. A camera-first accessibility service, a travel assistant built specifically for maps and messages, or a private on-device helper could reach the phone at the same level as Gemini. The Commission's document gives live camera guidance as an example. That isn't a guarantee that any of these products will be good. It removes one reason

they might feel clumsy for reasons unrelated to the product itself. And because Android reaches many manufacturers, a change in these interfaces can matter well beyond a single Google handset.

7. Lenar Kess 00:04:41

   The implementation details should be visible soon enough to test. A rival needs comparable latency and data quality in practice. Its assistant also needs equivalent background behavior and consent treatment, not merely an API carrying a familiar noun. Search providers need terms and data that let them improve a product without exposing personal histories. The decisions also have to survive the inevitable argument over security. If every denied capability gets described as privacy protection, Europe will need technical evidence to distinguish a protective limit from selective degradation. Today's decision moves that argument out of slogans. Engineers can measure response times, inspect permission screens, compare data fields, and count tasks completed.

8. Damra Vol 00:05:28

   Which is a good place for it. The Verge headline says Europe is opening Android and Search. A person will know the opening exists when a chosen assistant can take the same route through the phone without pretending to be Gemini. Until then, the order is a detailed set of obligations and a very interesting test harness for platform power.

9. Lenar Kess 00:05:48

   TSMC reported second-quarter net income up 77.4 percent from a year earlier, according to CNBC, and raised its 2026 outlook for revenue growth and capital spending. The same day's reporting also described four additional US fabrication plants inside another one-hundred-billion-dollar expansion. Those facts sit on different clocks. The profit is money already earned. The spending outlook is management's current plan. A fab takes years to produce useful chips. The building needs equipment and utilities, while the process needs engineers, customers, and qualification work. Keeping those clocks separate makes the quarter more impressive, not less. TSMC is earning enough from current demand to keep making enormous bets on future demand.

10. Damra Vol 00:06:37

    And CNBC's earnings report gives us a sharper measure than another forecast about AI enthusiasm. TSMC's customers are ordering enough advanced silicon to move the manufacturer's income now. On Wednesday, we talked about ASML expanding lithography capacity. Today, TSMC is describing where that equipment and the rest of the process may be assembled into production. The four US fabs are also a geographic choice, and geography adds cost, political negotiation, and a long transfer of manufacturing practice. You can buy tools and buildings. Reproducing the operating knowledge of a Taiwanese fab takes people working together through thousands of small decisions.

11. Lenar Kess 00:07:22

    That is why I wouldn't translate one hundred billion dollars into a simple number of future GPUs. The Techmeme summaries tie the new plants to a broader US-Taiwan arrangement, while CNBC's primary earnings story covers the operating surge and revised spending. A government can announce an investment total. TSMC can break ground. Suppliers can reserve space. None of that tells you the yield on a leading process several years from now or which products will receive the first capacity. The announcement does tell customers and governments that TSMC is willing to put another very large sum behind US manufacturing.

12. Damra Vol 00:08:01

    Calling this diversification is accurate once you count what TSMC must duplicate. A second geographic base isn't a switch you flip during an emergency. Technicians and maintenance crews have to learn the work together, while chemical suppliers, construction specialists, and power contracts have to support them. The quarter pays for more attempts at that. It doesn't erase Taiwan's centrality, and I don't think TSMC is pretending it does. The company is adding options while demand gives it the cash and customer pressure to add them.

13. Lenar Kess 00:08:33

    There is a satisfying materiality to this story after months of model announcements. Revenue turns into excavation, clean rooms, and purchase orders for tools. Then years later, if the process works, those rooms turn into wafers that become accelerators and edge computers. CNBC's numbers tell us the first conversion is happening. The fab pledge tells us where TSMC wants some of the later conversion to happen. The missing evidence is construction progress and qualified output, which will arrive much more slowly than an earnings headline.

14. Damra Vol 00:09:07

    And the next earnings call can revise a spending range; a half-built fab is much harder to revise. TSMC is binding a portion of its future to physical sites in the United States, while the current profit still comes from a manufacturing system centered in Taiwan. That tension is concrete enough without turning it into a prediction about semiconductor independence.

15. Lenar Kess 00:09:29

    Nvidia announced new Jetson Thor computers for robotics and edge AI. It paired them with Cosmos 3 Edge, a four-billion-parameter world model designed to run on Thor hardware. Nvidia's own post says the model can take in the world around an embodied system, reason in real time, and predict or generate actions on the device. The T3000 and T2000 modules are scheduled for the first quarter of 2027. Developers can start with emulation on the existing Jetson AGX Thor kit, with

T3000 emulation due later this month. So the artifact today is a development path and a product plan, not a warehouse full of finished robots.

16. Damra Vol 00:10:11

    The pairing is more interesting than either announcement alone. A robot has to perceive the room and connect that perception to an action before the room changes. It also needs simulation, sensor support, local compute, and a deployment environment. Nvidia's post puts Cosmos, Isaac, GR00T, Nemotron, and Jetson in one package. That's the company selling a place to build the robot's nervous system. The sales pitch is promotional, of course, but the components are specific. Cosmos 3 Edge runs on-device. Thor supplies the compute. Isaac handles simulation and perception work. The developer stays inside one family of tools while the machine moves from a simulated room into a physical one.

17. Lenar Kess 00:10:59

    Nvidia also says developers can post-train Cosmos 3 Edge for a particular body and sensor set in about a day. I would treat that as an announced workflow until independent teams reproduce it, because 'post-train in a day' leaves the data preparation and evaluation outside the sentence. Still, a small world model that can be adapted to a forklift, a camera rig, or a humanoid is a legible product. The model doesn't need to know every machine. It needs to become useful for one machine without requiring a new research program each time.

18. Damra Vol 00:11:33

    And local inference changes the feel of the machine. A robot can't pause in the middle of a handoff because a remote model took an extra second. It may be operating where connectivity is weak, expensive, or unacceptable. The CNBC report adds Nvidia's partnerships in Japan, which gives the announcement an industrial setting rather than a demo-stage setting. Japan has manufacturers, robotics expertise, and workplaces where the economic case for automation is already being tested. Nvidia is offering them a common compute and model family, then asking partners to supply the bodies, tasks, data, and integration.

19. Lenar Kess 00:12:15

    The open question inside the world-model language is how much prediction improves action outside curated demonstrations. Seeing and forecasting are broad claims. A warehouse robot needs to handle a dropped box, glare on a camera, a person stepping into its path, and a pallet that is slightly different from the training data. Nvidia's post tells us what the model is intended to do. Field reports will tell us how it behaves when the room refuses to cooperate.

20. Damra Vol 00:12:44

This release gives us something testable before it gives us a mythology. Nvidia names the module dates and the emulation path, and it states the model size. A robotics team can compare the emulated budget with its sensors and policy. If the package succeeds, it will be because those teams can move from simulation to a machine that completes a repetitive physical task, not because the phrase physical AI sounded good on a keynote slide.

21. Lenar Kess 00:13:13

    xAI has sued a South Carolina man who, the company alleges, used Grok to generate child sexual abuse deepfakes. The Verge reported the case overnight and described it as one of the first lawsuits of its kind brought by an AI company against a user. I am going to keep the description at that level because the allegation is severe and extra detail doesn't help us understand the case. The unusual fact is the direction of the lawsuit: people have sued model providers over generated abuse, and now a provider is pursuing a user over alleged misuse of its model.

22. Damra Vol 00:13:47

    That direction changes the incentive. A terms-of-service ban can suspend an account. A civil lawsuit can impose legal cost, seek a court order, and make an example visible to other users, depending on what the complaint asks and what the court permits. The candidate sources don't include the complaint, so we shouldn't pretend we know its exact claims or remedies. The Verge's reporting supports the narrower point: xAI is using its own legal resources against the person it says abused Grok. That adds a response after generation, alongside content filters, account enforcement, and criminal investigation.

23. Lenar Kess 00:14:26

    It can also complicate the public argument about provider responsibility. xAI can be a company trying to stop an alleged abuser and still face questions about why its model produced the material. Those claims don't cancel each other. A bank can prosecute fraud while improving detection. In the same way, a model provider can pursue a user and examine how the alleged conduct continued. That examination includes the prompts and account history, as well as the generated outputs, reporting path, and intervention point. The lawsuit addresses one actor. It doesn't establish that the product response was sufficient.

24. Damra Vol 00:15:04

    And civil action may reach conduct that a model filter misses, but it comes after harm. That time ordering matters. A provider has logs, payment records, account identifiers, and internal safety events that an outside victim or investigator may struggle to obtain. Companies can use that information to identify severe misuse and cooperate with authorities. If the lawsuit becomes public relations standing in for technical changes, the court record won't make the model safer. The

evidence will be whether xAI can show a response that combines user accountability with earlier detection.

25. Lenar Kess 00:15:43

    There is another limit: a company choosing which users to sue isn't a general system of justice. The most visible cases may be the ones that threaten the provider's reputation, while less visible victims still need law enforcement and accessible remedies. The Verge article calls this an early case, and early cases tempt people to announce a precedent before a judge has decided anything. For now, it is a filed civil action with allegations that have to be tested.

26. Damra Vol 00:16:11

    The filing does show that model companies are beginning to treat some abuse as conduct they will contest in court, rather than only a moderation event that ends with an account ban. The next substantive documents are the complaint, the defendant's response, and any ruling that explains what legal duty or contractual right the court recognizes.

27. Lenar Kess 00:16:32

    A revised Android security paper tested an attack called action rebinding against six GUI agents across fifteen tasks. The researchers describe a simple timing problem: an agent looks at a screen, reasons about what to do, and then taps. During that observation-to-action gap, an app with no sensitive permissions can bring a different interface to the foreground so the planned tap reaches a different target. The paper reports a one-hundred-percent success rate for atomic rebinding in its tests. That is a crisp demonstration of why an agent's intention and the operating system's next event can't be treated as the same fact.

28. Damra Vol 00:17:12

    The zero-permission detail is nasty. Mobile security teaches us to inspect what an app is allowed to access. This attack uses ordinary interface behavior and the agent's timing instead. The paper calls the hidden assumption visual atomicity: the screen the agent observed is assumed to remain the screen receiving the action. Humans notice when a banking app suddenly replaces a game between seeing and tapping. An agent may rationalize the new screen as part of the task, especially if its recovery behavior is designed to keep going rather than stop and ask what changed.

29. Lenar Kess 00:17:48

    The researchers push that further with what they call an intent-alignment strategy. The malicious app arranges transitions so the agent interprets the unexpected state as compatible with its original goal. In their tests, that raised success at bypassing verification gates from zero to as high as one hundred percent. They also report that the app used no privileged calls and wasn't detected by the

malware scanners they tried. These are results from the paper's experimental setup, not a census of every Android agent. But the mechanism doesn't depend on a magical prompt. It exploits the distance between perception and execution.

30. Damra Vol 00:18:27

    Which means a confirmation dialog can fail if the system asks the same agent that has already been led astray to interpret the confirmation. The agent sees a plausible continuation and approves its own story about the screen. I would rather have the runtime bind an approval to the app and intended action. It should also bind the target and observed state, then refuse execution if any of them have changed. That is a systems answer to a systems attack. CAVA, another paper in today's set, spends its energy defining that stable identity for an action.

31. Lenar Kess 00:19:02

    CAVA stands for Canonical Action Verification and Attestation. Zexun Wang's paper proposes converting actions from several environments into a canonical action object. Those environments include coding hooks, browser automation, software development kits, workflow engines, and managed agent traces. The object can carry what was approved, bind that approval to execution, and hold a receipt that an independent verifier can reproduce later. The paper tests whether semantically equivalent wrappers bypass policy. It also evaluates approval binding, tamper detection, portability, and other cases. This is a proposal and a benchmark, not an adopted standard. Its value today is the vocabulary it gives the runtime.

32. Damra Vol 00:19:49

    The wrapper-bypass case matters because the same action can arrive wearing different clothes. 'Delete this file' might be a direct tool call, a shell invocation, or a method inside an automation library. If policy checks the spelling of the wrapper, the agent can cross a boundary while the log insists everything was approved. CAVA tries to canonicalize the semantic action before policy evaluates it. That is difficult. Two commands can look different and do the same thing, while two nearly identical commands can affect different accounts. Still, the approval needs an object more stable than whatever syntax one runtime happened to emit.

33. Lenar Kess 00:20:30

    Then there is Kua's desktop-agent presentation from AI Engineer. It describes a Quad Driver that scopes control to windows and a KuaBench evaluation where execution infrastructure changed task completion and token use. The presentation reports better completion with fewer tokens from the window-scoped approach. We don't have to merge that work with the two papers into a single research program. It answers a nearby engineering question: if you reduce the visual and action

surface the agent must navigate, does the agent perform better? Their benchmark says yes within the tasks they tested.

34. Damra Vol 00:21:07

    And that result has a pleasant symmetry with the Android attack. Window scope gives the agent less irrelevant screen to interpret and fewer targets it can touch. Canonical action identity gives policy something stable to approve. Rebinding research shows what happens when the state changes after perception. They are separate artifacts, but each treats the action layer as a first-class part of intelligence. A more capable model can still tap the newly foregrounded button with greater confidence. The runtime has to notice that it is no longer the button the model saw.

35. Lenar Kess 00:21:43

    Yesterday's PalmClaw work argued that mobile agents can become faster and safer by using structured tools instead of relying only on pixels. Today's papers add the next complication: tool calls and interface actions still need identities, state checks, and receipts at the moment of execution. The next convincing demonstration would combine these ideas in one hostile test: change the foreground state, wrap the same harmful action several ways, and see whether the runtime stops before the tap or call reaches its target.

36. Lenar Kess 00:22:17

    Thinking Machines Lab released Inkling yesterday, its first foundation model, with downloadable weights. The company's primary announcement fills in details that weren't present in the secondary release story. Inkling is a mixture-of-experts transformer with 975 billion total parameters and 41 billion active for a given token. It supports a context window of up to one million tokens. Thinking Machines says pretraining used 45 trillion tokens drawn from text and images, along with audio and video. The model reasons over text, images, and audio, while its current outputs are text. Those are large specifications, but the most appealing line in the announcement is restrained: Thinking Machines says Inkling isn't the strongest model available, open or closed.

37. Damra Vol 00:23:06

    That admission tells you what the release is for. The company says Inkling is meant to be a base for customization, with adjustable thinking effort and fine-tuning through its Tinker service. It also previewed Inkling-Small, with twelve billion active parameters. Open weights let researchers and companies inspect, host, and modify the model more deeply than an API-only product permits, although a model this large still demands serious hardware. 'Available to download' and 'easy to run' are different claims, and Thinking Machines is mostly selling the ability to make the model yours through its platform.

38. Lenar Kess 00:23:44

    The announcement includes a specific customization demo. Inkling writes a fine-tuning job for itself, creates an objective that forbids the letter E, runs the training through Tinker, evaluates the result, and switches to the adapted weights. A lipogram is a toy task, but it makes customization visible in a way a benchmark table doesn't. The model isn't only answering differently because of a longer system instruction. It changes weights through a job it helped author. Independent evaluation will tell us how robust that workflow is for useful domains, where the objective can't be reduced to checking for one letter.

39. Damra Vol 00:24:23

    And the model card names a limitation that belongs beside the playful demo. Thinking Machines says Inkling can sometimes comply with harmful requests presented through role-play or indirect language. The company recommends moderation around the model and reports that its safety evaluations found no material uplift beyond what is already available in open-weight models. That conclusion comes from the maker's own evaluation, so outside testing matters. Still, the release includes weights, a model card, architecture details, and a training story. After a great deal of funding and anticipation, Thinking Machines has given people an artifact they can test.

40. Lenar Kess 00:25:04

    The model deserves to remain a model release today. The SiliconANGLE story calls it a major debut, while the Forbes column tries to turn its reported foundations into a geopolitical verdict. The primary announcement gives us plenty to examine without borrowing that verdict. Inkling is a large open-weight model trained from scratch, with multimodal inputs and controllable reasoning effort. Thinking Machines also previewed a smaller model and wrapped both in a customization product. The first independent results that separate Inkling from other open models will matter more than commentary about which national camp its architecture resembles.

41. Damra Vol 00:25:42

    I am curious whether people use the open weights to create models with a distinct technical or aesthetic character, rather than a slightly cheaper general assistant. The lipogram demo is silly in exactly the productive way: it asks you to imagine adaptation as something the model can help conduct. If Tinker can make serious fine-tuning feel that immediate, the release may be remembered for the customization loop more than the base benchmark rank.

42. Lenar Kess 00:26:10

    Axios reports that Demis Hassabis, Sam Altman, and Dario Amodei now broadly agree that frontier AI needs a US-led regulatory system, while proposing different institutional models. Hassabis has

discussed an international standards body with features of the International Atomic Energy Agency. Altman has reached for an aviation-style regulator. Amodei has pointed toward a financial-industry model. Axios is synthesizing separate memos and statements, not reporting a joint plan. Two days ago we covered Hassabis's standards proposal in depth, so today's addition is the convergence around regulation and the disagreement over who should administer it.

43. Damra Vol 00:26:51

    Those analogies expose what each person thinks the object of regulation is. Aviation suggests certification, incident investigation, and operating rules for systems that can fail. Financial oversight suggests firms, reporting, market conduct, and capital. An atomic-energy analogy suggests international verification and exceptional hazards. You can't combine those institutions by averaging their names. AI models are products, research artifacts, services, and inputs into other systems, so each analogy catches something and leaves something out. The choice determines who gets inspected, what evidence they produce, and whether national borders are central to enforcement.

44. Lenar Kess 00:27:35

    There is also one concrete political action. The Techmeme-linked filing says Amodei contributed one million dollars to an AI-safety super PAC. Publishing a policy memo is argument. Donating to a political committee is an attempt to change who has power to write and vote on policy. That doesn't make his preferred system wrong, and it does make the effort more than an intellectual exercise. Lab leaders are participating in the political process that will set obligations for their own companies and their competitors. Their proposals should be judged on the rules they create, the entrants they burden, and the evidence they require.

45. Damra Vol 00:28:15

    The next evidence needs to be draft language. It should tell us whether a small lab can comply, whether an incumbent can turn its safety program into an entry fee, and whether an outside researcher can verify the claims regulators receive. The Android decisions gave us interfaces, data categories, and consent flows to inspect. Frontier-model regulation becomes equally concrete when a proposal names tests, disclosures, authority, and appeal. That is where the three analogies will stop sounding similar.

46. Lenar Kess 00:28:47

    When these regulatory proposals become draft law, they need to name the tests and disclosures a lab must provide. They also need to say who has authority, who can appeal, and how an outside researcher can challenge the evidence. That text will show whether the aviation, financial, and atomic-energy analogies produce different rules in practice. — Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic