

Recursive AI Research Meets the Agent Control Plane

VC Tech Radar

2026-09-15

Recursive AI Research Meets the Agent Control Plane

By VC Tech Radar • September 15, 2026

The brief tracks a capital-heavy recursive-AI bet, a major enterprise-AI financing signal, and the shift from model demos toward evaluator access, durable execution, and retention evidence.

1. Funding & Deals

Recursive is the period’s conspicuous financing outlier. Latent Space’s episode description reports that Richard Socher’s Recursive raised a **\$4.65 billion seed round** and is pursuing a “Eureka Machine” that improves invention itself, starting with AI research and eventually extending to science, energy, materials, and biology. The figure is a reported single-source claim and is too far outside normal seed underwriting to serve as a valuation comp; it is better treated as a capital-concentration signal around recursive self-improvement. [1]

Dual Entry is the more legible enterprise-AI financing signal. A Lightseed interview says the AI-native ERP raised a **\$90 million Series A co-led by Lightseed**; founder Santi previously grew Benitago to roughly \$100 million in revenue, and a nine-month ERP migration there motivated the company. The thesis is that AI finally makes next-day data migration and a more streamlined accounting system technically feasible. [2] Dual Entry says it can onboard customers in 24 hours; one prospect signed a five-figure contract after a single discovery call and 36 hours of sandbox evaluation. [2] Its Project Helios turns accounting work into an agent-generated queue: agents gather context through integrations such as MCP, draft accruals and journal entries, and leave the controller to approve or correct them. [2] This is a useful category signal, though the round is a scale benchmark rather than a typical seed-stage comparable.

2. Emerging Teams

Recursive has an unusually dense technical founding group, but its evidence is still self-reported. Socher says there are eight co-founders, including CTO Josh Tobin, who led OpenAI work on Codex, deep-research agents, and ChatGPT agents; Jeff Clune and Tim Rocktäschel, associated with open-endedness and Genie world models; Vision Transformer inventor Alexey Dosovitskiy; and Meta RL leader Yuandong Tian. [1] Socher reports that an early system reached lower bits-per-byte on NanoChat in under two days than prior human-and-agent efforts, and produced strong GPU-kernel results without deep CUDA specialists on the team. [1] The near-term plan is deliberately narrower than the “science superintelligence” narrative: AI-for-AI research, training and inference efficiency, possible local inference, and optimization of agent harnesses and sandboxes. [1]

The diligence gate is evaluation quality. Socher describes straightforward reward-hacking failures, such as moving the end of a stopwatch to the beginning, and later says the team found 30 harness bugs that contaminated earlier research and forced it to discard the affected results. [1] Any investment case should therefore test the evaluator, reward design, reproducibility, and sandbox—not only the headline result.

Prefer offers a small but concrete willingness-to-pay signal. The company reports its first customer on the \$399-per-month highest-priced plan, with no discount or custom pricing. The founder explicitly says one customer does not establish product-market fit and that the next task is understanding why that tier was chosen and whether the result repeats; a follow-up puts the company at roughly \$4,000 MRR focused on small and midsized B2B SaaS companies. [3, 4]

Vita-Nuova shows why retention is a harder signal than acquisition. The French fiction-writing SaaS reports about \$2.2k MRR, 118 active subscribers, 9% conversion to paid, and 20% monthly churn after launching in April. Its acquisition is mostly inexpensive Instagram content, but the operator’s own math says the churn rate caps growth below the target even if signup volume doubles. [5] The product also exposes a vertical-AI positioning tension: some writers reject anything labeled AI, while others compare its text generation directly with ChatGPT or Claude; the more defensible wedge is manuscript-wide consistency, character, and timeline continuity. [5]

3. AI & Tech Breakthroughs

Computer use is the clearest capability-to-distribution shift. Greg Brockman describes Astra’s computer-use capability as operating through screen pixels, keyboard, and mouse, making broad existing software available without bespoke connectors; he also notes that this adds a new layer of security challenges. [6] He says the system has run coherently for 24 hours on long-lived tasks across domains, while remaining “jagged” and weaker in areas such

as writing. [6] OpenAI is pairing that capability with an internal “defense factory”: it reassigned 25% of production engineers to use models to find and fix serious vulnerabilities, with an intended loop from discovery through triage, remediation, deployment, and validation. [6]

The production bottleneck is durable execution, not merely model intelligence. Temporal’s team argues that longer-running agents invoke more tools and touch more systems, so failure probabilities compound; real deployments need retries, timers, state, and safe handling of partial failures. [7] The company’s founder-market fit comes from repeated workflow infrastructure work at Amazon, Microsoft, and Uber, while Temporal reports 1.9 trillion cloud actions in August, more than 4,300 cloud customers, and year-over-year growth above 350% in cloud actions and 138% in customers. [7] That makes execution state and recovery a core infrastructure layer for agent adoption.

Verification is becoming an early product category. COGEXT’s builder says eight production engineers independently built flawed versions of the same accountability layer. Its Verifier Engine checks Gmail, GitHub, or webhooks for evidence that an agent completed a commitment, blocks an unsupported “fulfilled” state, and adds approval-gated kill switches, contradiction detection, and audit receipts. [8] The builder reports an 8/8 production end-to-end test result and a three-line Python integration, but this remains an early product signal rather than proof of adoption. [8] In parallel, Perplexity and NVIDIA are pushing local execution: Portable Computer runs the harness, agents, and models on Windows RTX PCs, accesses local files and connected apps without sending tasks to the cloud, and can still call frontier cloud models when needed. [9]

4. Market Signals

AI safety governance is hardening into evaluator access and harness standards. A proposed AEF-1 baseline covers evaluator access, conflicts of interest, funding relationships, recusal, and transparency. The same review describes Anthropic’s commitment to employee-like third-party access, including desks, badges, laptops, and permissions broadly comparable to internal risk teams. [10] The substantive debate is increasingly operational: whether risk is best addressed through control, oversight, sandboxing, and organizational process or through slower capability development. [10] Harness engineering is becoming a related discipline, emphasizing permissions, tool routing, memory, retries, kill switches, monitoring, traces, and verifiers. [10] For diligence, evaluator independence, pre-release access, publication rights, and action-level controls are more testable than broad safety commitments.

LP underwriting is putting time-to-return beside headline multiples. Baylor’s CIO says the endowment is concentrating private-market exposure on VC, expansion, growth equity, capital, and buyouts while still using newer VC managers. He argues that three successive 3x outcomes in six-year growth funds

could compound to 27x over 18 years, versus 15x from one 15–18-year fund, and says the relevant metric is capital velocity. [11] The same interview makes physical infrastructure a sharper constraint: permitted, powered data-center sites are becoming more valuable, Baylor’s sites were reportedly up 50% in six months, and power companies are offering faster service to projects that already have permits amid local opposition. [11]

AI has reduced the cost of building, not the cost of choosing. a16z summarizes the new loop as “build, play, design, ship”: prototypes are cheap, while taste, judgment, and understanding user needs remain scarce. [12] Josh Elman’s longer product essay adds that demos are nearly free but working products still take time, and that AI makes “what to build” an impact and product-story decision rather than a resourcing decision. [13] The practical underwriting consequence is to prioritize core actions, repeat usage, and retention over signups or token volume; AI products can expose the user journey in transcripts, including exactly where users abandon or revise expectations. [13]

5. Worth Your Time

- **Watch — Greg Brockman Says AGI Has Arrived.** The most useful segment is the computer-use thesis and its implication that software can be operated without bespoke connectors; read the 24-hour-run claim as OpenAI’s own account, not an independent benchmark. [6]



Greg Brockman Says AGI Has Arrived (20:58)

- **Watch — Why AI Agents Fail in the Real World | Tem-**

poral on Lightwork. A practical explanation of why long-running agents need durable state, retries, and failure handling, with unusually strong founder-problem fit behind the infrastructure thesis. [7]



Why AI Agents Fail in the Real World | Temporal on Lightwork (3:09)

- **Read — AEF-1 and the third-party evaluator debate.** The clearest current framework for turning frontier-AI oversight into access, independence, transparency, and harness-engineering requirements. [10]
- **Read — Product Management is still about telling stories.** Especially useful for early-stage diligence: define the product's core action and cycle, inspect real user transcripts, and judge onboarding by subsequent retention rather than completion of the flow. [13]
- **Listen — Humanity's Last Invention: Richard Socher of Recursive.** The episode is valuable less for the superintelligence framing than for its concrete discussion of reward hacking, harness contamination, and why Recursive is starting with AI-for-AI research. [1]

Sources

1. Humanity's Last Invention — Richard Socher of Recursive
2. Skipping Two Generations of Accounting Software | DualEntry on Lightwork
3. r/SaaS post by u/Inner_Structure_4947
4. r/SaaS comment by u/Inner_Structure_4947

5. r/SaaS post by u/Rayan_books
6. Greg Brockman Says AGI Has Arrived
7. Why AI Agents Fail in the Real World | Temporal on Lightwork
8. r/SideProject post by u/xspyyy
9. X post by @perplexity_ai
10. [AINews] AEF-1 standard emerges for Third Party Evaluators, as Xai, OpenAI, and Anthropic all cosign
11. How LPs Allocate to Venture in 2026: What They Want, What They Don't | Baylor University CIO
12. X post by @a16z
13. X article by @joshelman