

Decision Models Move Into Production as Open Weights and Agent Permissions Tighten

VC Tech Radar

2026-09-22

Decision Models Move Into Production as Open Weights and Agent Permissions Tighten

By VC Tech Radar • September 22, 2026

A concise investor brief on decision models moving into production, a deep-tech inference bet, accelerating open-weight adoption, and the workflow, distribution, and permission constraints shaping AI deployment.

1. Funding & Deals

Corridor Brokers announced a \$25M seed led by BainCapVC to serve small-business employees with health insurance. The company frames the opportunity as changing the economics of a segment it says has been underserved because it was not economically viable; the investor describes the team as transforming how health insurance is sold. For the radar, this is a seed-stage distribution and market-structure bet. The disclosed evidence supports an insurance thesis, not yet an AI-specific moat. [1, 2]

2. Emerging Teams

iKawn Mirror is moving a live AI try-on demo into a retailer workflow. The team is working with its first wig client after demos reached 700,000+ Reddit views. The client engagement is exposing the real product requirements: hair density, hairlines, colour accuracy, preserving the customer's face during movement, and a cost for every minute of use. The useful pilot metric is conversion lift with and without try-on—not engagement or visual impressiveness alone. [3, 4]

SpacePlanner is a sharp distribution signal, but not yet a monetization signal. The free, no-signup floor-plan tool went from a few hundred daily users to 15,000+ unique users, 70,000+ pageviews, and thousands of registrations. Its founder attributes the surge largely to unpaid creator exposure,

including a SetupsAI video from an account with 6M+ followers, followed by smaller creators across several countries; the original Indonesian source remains unidentified. The founder had to redeploy to cheaper infrastructure as costs rose, and a feedback popup generated almost 5,000 responses in two days. This is evidence that dark-social distribution can overwhelm a technically modest product, not evidence of durable revenue or retention. [5, 6]

3. AI & Tech Breakthroughs

Unconventional AI is betting that the next inference breakthrough is a new computer, not a larger model. Co-founder and CEO Naveen Ralph has electrical-engineering training and a neuroscience PhD; he says he founded Nirvana Systems, sold it to Intel, ran Intel’s AI group, and later built GPU-scaling infrastructure that joined Databricks. Unconventional AI’s stated goal is $1,000\times$ power efficiency, with the original five-year target revised to three and a half years. Ralph’s energy case—using reported token volumes and an assumed 10 joules per token to estimate 12 GW for one company’s AI services and an energy shortfall in roughly three years—is explicitly his estimate, not an independently validated forecast. [7]

The technical claim is a dynamical computer built from interconnected oscillators, combining memory and computation in the same elements. Ralph says the team open-sourced UNO, an oscillator-based image-generation model, then taped out the first physical dynamical computer in five months and generated images at roughly 500 nanojoules each; the architecture uses physical die-stacking plus time dynamics, which the company calls “4D computing.” The proposed commercial product is a managed VM/rack within about two years, but existing models will require model-layer porting and the company is coordinating theorists, biologists, physicists, and chip engineers. That makes independent energy/quality benchmarks, fabrication, porting friction, and cross-disciplinary execution the key diligence gates. [7]

Decision models are becoming both a model category and a control primitive. SemIf, an open-source Qwen3.5-based alternative, is being served free through LangSmith Gateway; it is reported at 74.7 versus Jev’s 75.4 on JevBench, while an open-source multilingual System 1 decision model became Hugging Face’s top trending model shortly after Jev began trending. LangSmith’s Jev-as-a-judge integration is designed to score every production trace, evaluate multiple criteria, and trigger webhooks for issues such as PII leakage, prompt injection, or toxicity. Its setup documentation also notes that TypeSafe does not yet offer zero data retention, a relevant enterprise diligence flag. [8, 9, 10, 11]

The adoption pattern is less about flashy general intelligence than cheap, repeated semantic decisions: a non-affiliated directory collected 607 Jev builds, with examples including 500 emails classified for \$0.035 and 3,282 posts scored for \$0.13. ThruWire’s accompanying architecture shows where value may accrue

around the model: checkpoints require evidence-bearing artifacts and stable upstream relationships, while Jev supplies probabilistic judgments about whether evidence supports a claim or an implementation satisfies a requirement. The likely investable layer is therefore the connective tissue—checkpoints, provenance, policies, and feedback loops—not an isolated classifier. [12, 13]

4. Market Signals

Open-weight models are becoming a strategic deployment layer. Interconnects argues that Chinese models such as GLM-5.2 and Kimi K3 have crossed a commercially meaningful agentic-capability threshold; it estimates Chinese open-weight models are 2–5 months behind the closed U.S. frontier, while U.S. open models remain 6–9 months behind. OpenRouter usage reportedly rose from about 1T to 80T open-model tokens per week, with Chinese models exceeding 80% of usage; OpenCode reports roughly 95% or more of inference volume from Chinese models. The same analysis identifies inference platforms, fine-tuning APIs, and purchased agent-training environments as early layers of an open-model economy, while warning that open software is difficult to keep from bad actors and that U.S. products are increasingly dependent on Chinese models. [14]

Corporate AI language is moving faster than measured business impact. In the June 2026 earnings season, 33% of S&P 500 companies making calls quantified AI use, but only 15% quantified AI’s business impact. Cost/productivity claims outnumbered revenue claims 26% to 16%; deployment language reached 12% while pilot language stayed below 4%; and “agentic” appeared in 24% of calls versus 5% for “generative AI.” These are earnings-call signals, not proof of production value, and the source warns that pilot activity is difficult to infer from the language. [15]

That gap strengthens the workflow-first investment thesis. UiPath CEO Daniel Dines says models are interchangeable while the defensible asset is the enterprise “map of work”—workflows, exceptions, procedures, and systems. A vibe-coded procurement tool still required human intervention for connectors, permissions, audit, security, testing, maintenance, and a faulty database schema. The implication is to underwrite process capture, exception handling, and deterministic controls rather than API access alone. [16]

Agentic commerce is meeting a permission wall. Amazon cut off Meta’s Muse from shopping on Amazon.com after saying Meta had not obtained permission, Muse did not identify itself as an automated agent, and it appeared to capture and store customer credentials. Amazon says an agent that can move through account pages, order history, transactions, and sensitive data must operate openly and have the merchant’s agreement; Muse’s browser fallback makes services without public APIs reachable in precisely the way that raises this dispute. Agent identity, delegated consent, credential isolation, and merchant opt-out are becoming distribution requirements, not post-launch compliance details.

[17]

The governance problem is also reaching capital markets. Harry Stebbings asks how a frontier provider can pursue an IPO without answering who is liable for harmful acts by rogue agents, while Eric Schmidt argues that a development pause conflicts with participants' incentives and would be difficult to verify. [18, 19]

5. Worth Your Time

- **Watch — Naveen Ralph on Unconventional AI.** The strongest segment explains oscillator synchronization, UNO, the physical prototype, and the memory-compute architecture behind the “4D computing” claim.



[7]

Naveen Rao: 4D Computing, AI's Energy Wall & Beating Biology (12:04)

- **Watch — Daniel Dines on the enterprise “map of work.”** The clip is useful for understanding why process discovery, exceptions, and one-process-at-a-time adoption may matter more than access to a frontier



model. [16]

Daniel Dines on Why Work Processes Not Models Will Be The Most Valuable Asset in AI (27:53)

- **Read** — *The current balance of power in open models.* The article is the best current synthesis here of Chinese open-weight capability, usage concentration, infrastructure winners, and the associated security and policy risk. [14]
- **Read** — *Jev and AI Checkpoints: Using Decision Models to Wrangle Agent Work.* Its practical value is the architecture: let agents execute flexibly, but require evidence-bearing checkpoints, semantic validation, provenance, and feedback loops. [13]

Sources

1. X post by @realnikhy
2. X post by @ajay_bcv
3. r/SaaS post by u/ivineets
4. r/SaaS comment by u/kr1sck
5. r/SideProject post by u/One_Progress_1044
6. r/SideProject comment by u/One_Progress_1044
7. Naveen Rao: 4D Computing, AI's Energy Wall & Beating Biology
8. X post by @hwchase17
9. X post by @hwchase17
10. X post by @ClementDelangue

11. Jev is now available in LangSmith Evals
12. r/SaaS post by u/Sarthak999gupta
13. X article by @JoshARosen
14. The current balance of power in open models
15. Monday data: More AI numbers, more clarity?
16. Daniel Dines on Why Work Processes Not Models Will Be The Most Valuable Asset in AI
17. Amazon blocks Meta's Muse AI assistant in new standoff over agentic shopping
18. X post by @HarryStebbing
19. X post by @DavidSacks