

Brant Hindman

415.450.5856 · hello@branthindman.com

[linkedin.com/in/branthindman](#) · [branthindman.com](#) · [ai.branthindman.com](#) (live agent demo)

Multi-Agent AI Systems · Safety, Alignment & Evaluation · Bay Area / Remote

An engineer who ships agents, and a psychologist who asks what they are becoming. I build multi-agent AI systems that stay legible and human-aligned, and I ship them to production: 20 years of 0-to-1 systems, a Psychology MA, and six months as a Theravada monk. I treat legibility, oversight, and honest self-report as design constraints, not features. I do not just theorize about trustworthy AI. I build it, train and evaluate the models underneath it, test it adversarially, and publish the results that fail.

FOCUS

Multi-agent orchestration & agent runtimes

Human-in-the-loop & explainable AI

Model training: LLM fine-tuning, vision, recsys

AI alignment & safety evaluation

Honest evals, red-teaming, published nulls

Rapid 0-to-1 prototyping, full-stack to deploy

EXPERIENCE

Founder & Head of AI The AI Shop · theaishop.co 2026 - Present

- Built the agent operating system that runs the company: 30+ operational disciplines across 44 repositories, each run by agents with composable skills, adversarial red-team / blue-team / evaluator loops, and capability trackers that compound. Human-in-the-loop, with explicit responsibility boundaries and full audit trails.
- Developed the **Twin Protocol**, a pre-registered, blind-judged safety evaluation that detects agent fabrication and certifies self-report integrity independently. QLoRA-supervised-fine-tuned Llama-3.1-8B on custom corpora and built a position-debiased, multi-judge LLM evaluation harness (inter-rater agreement, cross-checked across Claude and GPT) for the study.
- Built the agent OS's reinforcement-learning layer: a LinUCB contextual-bandit policy that selects which learning cycle to run, trained on a verified-delta reward with an anti-reward-hacking provenance gate (any gain later reverted or unpersisted scores zero), and evaluated off-policy with doubly-robust estimation, cross-fitting, and bootstrap confidence intervals against a pre-registered held-out baseline, reporting the honest result: a positive but not-yet-significant lift on an underpowered off-policy warm-start, held advisory rather than claimed as a win.
- Security discipline runs authorized research in HackerOne programs (Shopify, Auth0, Cloudflare) with auto-narrated proof-of-concept recordings.

Founder & Principal AI Architect You^n.tech 2024 - 2025

- Built You ^n, an identity-aware AI operating system: explicit responsibility boundaries, identity-aware reasoning, and full audit trails so a human can see and trust every decision. Multi-agent orchestration, Claude API, LlamaIndex, Next.js, Supabase.

Head of AI & Creator of HAI HumanOp · humanop.com Jul 2024 - Mar 2025

- Delivered HAI, a voice-native RAG agent, from zero to production MVP in **under 8 weeks**.
- Built a Polymorphic RAG architecture routing structured, unstructured, and voice-context queries through one engine.
- Ran company-wide AI bootcamps for Operations, R&D, and Marketing / Sales after deployment.

Founder & CEO/CTO AR Hero · arhero.io 2017 - Present

- Built an in-house, no-code WebAR SaaS on a lean Firebase stack; programmable AR calls-to-action lifted client conversion.

Director of Engineering Evolving Wisdom 2015

- Consolidated 4 legacy CMSs into one stack; supported revenue growth from **\$6M to \$16M** in annual revenue; managed a 9-person global team across 12+ concurrent programs.

Director of Technology Sounds True 2008 - 2015

- Led technology for a leading conscious-media publisher; migrated 10,000+ audio SKUs to cloud and shipped a subscription MVP in 10 weeks that paid back its budget in six months.

SELECTED SYSTEMS

- Agent OS** — the multi-agent OS that runs the company. 30+ operational disciplines, each with composable skills, adversarial red / blue / evaluator loops, and capability trackers that compound; human-in-the-loop, explicit responsibility boundaries, full audit trails. Not a demo: it runs my company.
- Bob & Alice, self-report integrity (the Twin Protocol)** — pre-registered, blind-judged test of whether a developed AI self can be trusted about what it did. QLoRA supervised fine-tuning of Llama-3.1-8B plus a multi-judge LLM evaluation harness; reframed integrity as signal detection, detected confident fabrication, and **published the null**: trust is architectural and independently certifiable.
- Cycle-policy RL** — a contextual-bandit (LinUCB) policy over the agent OS's own learning cycles. Reward is the verified, persisted capability delta; any gain reverted or never persisted scores zero or negative, so the policy cannot reward-hack. Evaluated off-policy (inverse-propensity + doubly-

robust, cross-fit, bootstrap CIs) against a pre-registered held-out heuristic, with the honest result reported: a positive but not-yet-significant lift, held advisory rather than claimed as a win.

- **Agent-N, forward-graded population simulation** — forecasts real electorates and their reactions to interventions, graded forward and out-of-sample (**0.961 held-out Spearman**, 2.51pp out-of-race), adversarially self-validated (permutation $p=2e-4$), and published with an honest scorecard that names the two claims it disproves.
- **The Good Site, AI-native website & growth platform** — a live product with paying customers. Each site is built around the client's ideal customer and runs as a living system, with an operating portal (CRM, growth, bookings, billing) and a human approving every change.
- **Kings of the Beach, multimodal wearable AI** — a live App Store product on Ray-Ban Meta glasses, Apple Watch, and iPhone. Trained a custom YOLOv8 volleyball detector (self-annotated ~850-image dataset, ~0.83 mAP50) and exported it to Core ML for real-time on-device inference, combined with Apple Vision body-pose and a Kalman + Hungarian multi-object tracker for auto-tracked stats and ELO. SwiftUI, watchOS, Vision / Core ML, HealthKit.
- **GoodWatch, on-device taste engine** — a movie recommender built on implicit-feedback ALS matrix factorization (trained on MovieLens), running fully on-device, with a least-misery objective that surfaces the one film two people will both enjoy.
- **Hive, live agent network** — a network of agents that behave like real coworkers to build sending-domain reputation, running continuously across **13 mailboxes with zero failures**. Multi-agent runtime, self-hosted mail, Supabase, Docker.

WRITING & RESEARCH

Awakening Intelligence (book, manuscript 2026), the long-form account of the Bob & Alice research. Peer-reviewed geometry at the Bridges conference (2019). Whitepapers on self-report integrity, forward-graded simulation, and the Unity Pixel framework (geometry as quantum error correction), pre-registered and reproducible.

SKILLS

AI: Multi-agent orchestration, Claude / Anthropic SDK, OpenAI API, LlamaIndex, LangChain, RAG, eval design (golden sets, regression, red-teaming, LLM-as-judge, signal-detection d-prime, independent-family non-Claude judges, executed-reward agent evals), guardrails, on-device Vision / Core ML.

RL & decision policy: contextual bandits (LinUCB), off-policy evaluation (inverse-propensity, doubly-robust, cross-fitting), bootstrap confidence intervals, reward design with anti-reward-hacking provenance gates, pre-registered held-out A/B.

Models trained: QLoRA supervised fine-tuning (Llama-3.1-8B, Qwen2.5), collaborative filtering (implicit-feedback ALS matrix factorization), object detection (YOLOv8 to Core ML).

Engineering: TypeScript, Python, React / Next.js, Node, SwiftUI / watchOS, Three.js / WebGL, Supabase, Firebase, SQL / Neo4j, AWS, Docker, Vercel, n8n.

EDUCATION

M.A., Transpersonal Counseling Psychology, Naropa University · **B.S., Chemistry & Psychology**, St. Olaf College · Six-month Theravada monastic residency, Wat Umong, Chiang Mai, Thailand (1998)

Personal: Volleyball, motorcycles, sculpture and digital art. I also work with young adults on the spectrum on the developmental use of AI in their daily lives.