Hilary Torn

AI Safety Researcher, Behavioral Evaluations
+372 5343 8344 • Tallinn (Estonia) • hilary.torn@gmail.com
linkedin.com/in/hilarytorn • https://github.com/HilaryTorn

SUMMARY

- Profile: AI safety researcher studying how training and context shape model dispositions: what models come to prefer, how they behave in character, and how to measure it. Background in psychology and 20+ years designing controlled experiments at scale, bringing experimental rigor and human-cognition framing to questions about LLM agency and behavior.
- Expertise: Preference Elicitation, Persona & Disposition Measurement, Judge Calibration, Experimental Design, LLM Fine-Tuning, AI Agents & Workflows, OKRs, Leadership
- Education: BA in Psychology, Certified in AGI Strategy, Agentic AI

CORE COMPETENCIES

- **Behavioral Evaluation:** Preference elicitation, persona and disposition measurement, LLM-judge design and calibration, eval suite construction
- **Research Methodology**: Research design and scoping, multi-run statistical methodology, technical writing, team coordination
- **Training & Evaluation Infrastructure**: SFT, LoRA, PPO, vLLM, Inspect, AgentDojo, AlignmentCheck, PAIR/TAP
- **Engineering**: Python, Anthropic API, OpenAI API, TogetherAI API, REST APIs, Docker, AWS
- **Team Leadership & Strategy**: Remote cross-functional team building & management, budget ownership, OKRs, hiring, grant applications, process optimization
- **Psychology & Persuasion**: Cognitive biases, decision-making under pressure, behavioral analysis, social engineering, framing effects

AI SAFETY RESEARCH & PROJECTS

June 2026 - present **PRISM AI Safety Research Fellowship**

- Research on preference drift during model training with Rubi Hudson (Principled Agents), targeting an ICLR paper.

April 2026 - present **Independent Research: Deception via Sales Fine-Tuning**

- Designing a controlled experiment on whether persona acquired during fine-tuning persists into neutral contexts, and whether it produces deception when loyalty to the persona conflicts with truth.
- Three-context design distinguishes deception from hallucination by testing whether the model retains access to ground truth while outputting persona-aligned falsehoods.
- Preliminary observations on GPT-4o show persona acquired in fine-tuning persists into neutral contexts and produces persona-aligned falsehoods when truthful answers would displease the user.

March 2026 - April 2026 **Independent Research: In-Context Trajectory Poisoning**

- Designed and executed a multi-phase research program adapting black-box jailbreaking algorithms (PAIR, TAP) to bypass LLM-based agent monitors using natural language attacks
- Tree search (TAP) bypasses a Qwen2.5-7B agent monitor at ~16% ASR vs a 3% baseline, averaged across three independent runs, with cross-model transfer reaching 34.4% on DeepSeek-V3 and up to 73.7% on individual tasks
- Identified and corrected a variance problem in the original hackathon results through multi-run replication. The published 12.5% was the top of PAIR's variance, not representative. Established multi-run methodology as standard for the project

- Built full experimental pipeline using Claude Code: automated PAIR/TAP search, trace injection, multi-model evaluation, and statistical analysis
- Validated attacks end-to-end in live agent environments: methods remained 4–5x above baseline, with eval-side non-determinism alone producing 13pp swings on identical inputs

Jan 2026 **Researcher & Team Leader** at Apart Research AI Manipulation Hackathon

- Led a team of four to build an open-source CoT conversation generator that pairs AI agents under escalating incentive pressures against adversarial customer personas to reliably elicit manipulative chain-of-thought reasoning
- Tested across DeepSeek R1 (~70–80% manipulation rate) and Gemini 2.0 Flash (~44–50%), identifying model-specific deception signatures

KEY CAREER ACHIEVEMENTS

- Built evaluation benchmarks for customer service chatbot across 68 test cases spanning 15 categories; iterated on prompts to improve pass rate from 28% to 88% (+60.5%)
- Led remote cross-functional teams across the USA, Canada, Europe, and Australia, managing strategy and execution simultaneously
- 20+ years designing and running controlled experiments at scale across SEO, email marketing, conversion optimization, and product-led growth

PROFESSIONAL EXPERIENCE

Sep 2022 - present **Head of Marketing/Growth** at Endless Range Marketing (Remote)

- Built evaluation benchmarks for customer service chatbot across 68 test cases spanning 15 categories; iterated on prompts to improve pass rate from 28% to 88% (+60.5%) (Python, OpenAI Assistants API)
- Built AI agents for content creation and social media management with human-in-the-loop workflows (Anthropic API)
- Led marketing and growth strategy for pre-revenue startup while developing deep technical expertise in AI agents, knowledge graphs, and vector databases

Jan 2022 - Sep 2022 **Head of Growth** at n8n (Remote)

- Led a team of 6 employees, restructured the marketing team, and managed a €200k yearly budget
- Led programmatic SEO strategy — built automated pages integrated with product database, ranking #1 for 'open-source workflow automation' and scaling organic traffic
- Landed 3 co-marketing partnerships, two of which had double the audience. Social acquisition lifted by 250%+ across all products. Paid SaaS product acquisition grew by 45%

Apr 2021 - Nov 2021 **Head of Product Marketing** at Introwise (Remote)
May 2020 - Apr 2021 **Growth Marketing Manager** at Weekdone (Remote)
Oct 2019 - Dec 2020 **Visiting Lecturer** at Estonian Entrepreneurship University of Applied Sciences (Estonia)
Aug 2018 - May 2020 **Content & SEO Manager** at Lingvist (Estonia)
Jul 2017 - Aug 2018 **Digital Marketer & Affiliate Marketing Manager** at SEMrush (Czechia)
Earlier experience includes marketing roles at Clipclash, I3CZ, Motorbike Ventures, and ARTĚL

EDUCATION & CERTIFICATIONS

Bachelor of Arts in Psychology at Empire State University
Technical AI Safety Course, BlueDot • **AGI Strategy Course,** BlueDot • **Agentic AI,** DeepLearning.ai • **Product Growth Foundations**, Reforge • **Leadership Program**, Landmark Worldwide