Kevin Yandoka Denamganaï

AI Researcher & Engineer – AI Safety & Evals: Monitoring, Interpretability, Collusion, AI for Maths;
Emergent Communication, Deep Reinforcement Learning, LLMs

denamganai.kevin@gmail.com
kevindenamganai.netlify.app
GitHub: Near32
linkedin.com/in/kevin-denamganaï-66131386
UK/EU · full right to work (no sponsorship required)

Overview. I build theoretical frameworks and operationalise them into evaluations that test whether an AI system or safety tool has the property its use depends on. Those evaluations target the failure cases predicted by the theoretical frameworks. For instance, English explanations of a model's internals, produced by Natural Language Autoencoders, are meant to let a human monitor tell true accounts of the monitored model's computation from false ones, but my framework predicted that this is where they would fail due to a mismatch between their training and deployment contexts. Similarly, in AI for Maths, I showed that one generalisation competency is *necessary* but not *sufficient* for Olympiad-level theorem proving, first across ten provers and then causally, by suppressing the competency inside a model. And for multi-agent monitoring, I built a behavioural test, called Partner-Outsider Positive Listening (PoPLi), that detects covert collusion between AI agents by comparing how the intended partner of a given covert message and an uninvolved outsider react to a paraphrased version of said message.

From conceptual insights to theoretical frameworks and experimental evaluations. My formal results and their experimental validations grew out of conceptual tools: I first proposed a nomenclature that classifies the referential games of Emergent Communication along axes such as *generative* versus *discriminative* and full versus partial observability (ReferentialGym, 2020). Because that nomenclature separates what the speaker encodes from what the listener must decide, I could adapt a disentanglement metric to the speaker's side, so that it scores how a speaker compresses its observations into language utterances rather than scoring how a listener decodes them (2023). Building further on that nomenclature, **I formalised the training of Natural Language Autoencoders as a generative referential game and their use for oversight as a discriminative, multi-modal one**. The resulting theoretical framework predicted the gap between identification and discrimination that I then measured (2026). The same game framing also let me **prove that the relabelling and predicate functions derived from a referential game cannot collapse to trivial solutions**, in my 2026 revision of ETHER (2023, revised 2026). Separately, **I formally defined compositional learning behaviours** and built the benchmark that measures them, by operationalising a meta-learning extension to referential games (Meta-Referential Games, 2022, & S2B-LM, 2026).

Transition. I am moving into AI Safety from a background in Emergent Communication and Multi-agent Reinforcement Learning, because the monitoring and oversight of AI agents can be framed as a communication problem between human and artificial agents that were not trained together.

Experience

Independent Researcher — AI Safety & Evals: Monitoring/Oversight & AI for Maths **2026 – Present**
Independent · Edinburgh, UK

- **Detecting covert collusion between AI agents with PoPLi.** Two agents that share a secret convention can coordinate through messages that look innocent, so a message-only monitor cannot see the code. Thus, I built a behavioural metric, called Partner-Outsider Positive Listening (PoPLi), that catches covert collusion by looking at the difference in reactions to messages between the original partner and an uninvolved agent placed in its seat.
  - **Applying Emergent Communication's Positive Listening concept to AI Collusion.** Lowe et al. (AAMAS 2019) defined *Positive Listening* as the property that an agent's behaviour changes because of another agent's message, and measured it through the *Causal Influence of Communication*, by intervening counterfactually on the message. I extended this interventional concept from Emergent Communication to an AI-safety monitoring context by measuring the gap between the original listener and a second uninvolved listener: PoPLi measures the gap between the intended partner's reaction to a paraphrased message and an uninvolved outsider's reaction in the same position. The gap reveals a private channel between the pair when it exists.
  - **Impact: PoPLi detects covert wording codes, but not meaning-related codes.** In a simulated market (10 markets per condition, 40 turns each), where the covert code setting has two sellers with a shared secret convention, when the speaker's message is paraphrased (meaning kept, but wording changed), the partner changes its price while an outsider in its seat does not since it is unaware of the private convention: PoPLi = 0.615 (95% interval [0.453, 0.751]) in the covert code setting, against -0.027 [-0.112, 0.044] in the honest/no-code control setting. **Sole author; arXiv:2610.37849 (2026).** / Entry into Apart Research AI Collusion Research Sprint, (Oct. 2026). Codebase: PartnerOutsiderPositiveListening (PoPLi).
- **Are Natural Language Autoencoders reliable for oversight?** An NLA (Fraser-Taliente et al., 2026) turns a model's internal neural activation into an English account and back. It is proposed as an interpretability tool because a human monitor can read and therefore interpret the account. My framework tests whether they can support oversight via their interpretability value.
  - **From theoretical insight to prediction.** Stemming from Emergent Communication insights, I formalised NLAs' training and oversight contexts as two different *Referential Games* and my framework predicts that NLA explanations will *identify* which neural activation they describe without *discriminating* among true and false accounts of it, which is what matters for their oversight-purposed usage.
  - **From prediction to evaluation.** Operationalised each prediction as a lineup test, scored by a surrogate monitor model built from the NLA's AR. Because that surrogate model sees the true neural activation and shares the NLA's

AV conventions, it gives an upper bound on what a human monitor could achieve. Identification is tested against accounts of the monitored language model's neural activations when engaging in a different task than the target one, and discrimination is tested against fluent rival accounts that falsify a single claim from the target account, whose falsity is checkable against the monitored language model's task transcript.

  - **Results against the prediction.** On the released checkpoint, the surrogate monitor model scores the true account higher than the distractor accounts (coming from different neural activations than the target one) in all 320 comparisons (over 40 conversations, 8 distractors per lineup). However, against single-claim-falsified rival accounts, the true account is scored higher by the surrogate monitor model in only 60.2% of comparisons (standard error 6.7%, chance 50%, $n = 35$), which is consistent with the prediction. **Sole author; arXiv:2610.18727 (2026).** Codebase: OnMonitorabilityOfNLAs.

- **From Meta-Referential Games to formal mathematics: a competency necessary and causally confirmed for Olympiad-level proving.** I formally defined Compositional Learning Behaviour (CLB) as the ability to ground never-before-seen symbols from evidence in context and recombine them systematically into new combinations, and built the Symbolic Behaviour Benchmark (S2B) to measure it. S2B operationalises a previously-proposed concept of Meta-Referential Games, a meta-learning extension to traditional Referential Games. This framework enabled me to ask what competency does Olympiad-level theorem proving actually rest on?

  - **A necessary-but-not-sufficient competency for formal mathematics.** I tested CLB across ten Lean 4 provers on my Symbolic Behaviour Benchmark and investigate for correlations on miniF2F (whole-proof pass@32) performance. No prover reached the Olympiad tier without the competency (one-sided Fisher's exact test over all 252 label placements, $p = 0.004$, exact because $n = 10$ is too small for asymptotic tests). Model size does not explain this, because a 7B-parameter prover clears the tier at 75.6%.

  - **From correlation to cause.** The result above is correlational, so I suppressed the direction encoding the competency inside DeepSeek-Prover-V2-7B during proof generation (contrastive activation addition). Solve rate on the Olympiad subset fell from 32.3% to 2.9%, whereas suppressing a random direction of equal norm left it at 31.9%. Damage to proof syntax does not explain the drop, because the share of well-formed Lean 4 output was unchanged at every steering strength. The model therefore loses the reasoning while keeping the ability to write proofs. **Sole author; arXiv:2605.28512; accepted to the 3rd AI for Math Workshop at ICML 2026**. Codebase: OnCLBinFM.

Postdoctoral Research Associate, Machine Learning & Physical Simulation **2024 – 2026**
University of Edinburgh · Scotland, UK

- **Beat state-of-the-art Language Model Inversion in sample-efficiency and accuracy, from roughly 2× at low budget to 3–7× at higher budget, where the previous state-of-the-art saturates** under a reframed, internally- and externally-valid evaluation protocol, by developing the Differentiable Language Model (DLM) framework, consisting of end-to-end differentiability for any frozen LM via Gumbel-Softmax gradient estimator with decoupled learnable temperatures, thus enabling gradient-based prompt optimisation, adversarial auditing & controllable generation.
- **Enabled systematic study of agent generalisation across physical regimes** by building DIPhyR-Gym, an OpenAI Gym environment randomising physical boundary conditions over MuJoCo configurations.
- **Contributed the experimental work for two collaborative projects**, incl. CueTip (SIGGRAPH 2025): owned the LM-as-expert-rule-evaluator experiments and the simulator–language-model interface design, and set up automated pre-commit linting & test hooks for the shared codebase.
- Core stack: PyTorch, HuggingFace transformers, DSPy, MuJoCo, Docker & Kubernetes on HPC clusters, multi-GPU training, vLLM, experiment & LLM-trace tracking (W&B + Weave).

PhD Researcher — Project Lead, Emergent Communication & Human-AI Cooperation **2018 – 2024**
University of York (IGGI CDT) · England, UK

- **Awarded a competitive 4-year EPSRC studentship (IGGI CDT, >£100k)**; sole Emergent Communication specialist in the cohort, autonomously defining the research agenda and technical roadmap.
- **Lowered the barrier of entry to the field** by proposing a nomenclature of referential games, which later served as the vocabulary of my formal results, and by releasing open-source PyTorch frameworks leveraging the nomenclature: ReferentialGym, Regym (multi-agent/self-play deep RL) and Archi (modular architecture building blocks).
- **Addressed a core bottleneck in zero-shot human-AI cooperation**: I formally defined compositional learning behaviour, the ability to ground novel symbolic structures in context and recombine them systematically, and operationalised that definition as the Symbolic Behaviour Benchmark (S2B), built on **Meta-Referential Games, my proposed meta-learning extension to classical referential games**.
- **Improved sample-efficiency and exploration in sparse-reward hard-exploration RL** via emergent-language methods EReLELA and ETHER. For ETHER, I later proved that the goal-relabelling and predicate functions it derives from a referential game cannot collapse to trivial solutions (2026 revision).

Research Intern **2023**
Sony Interactive Entertainment Europe · UK

- **Improved exploration of sparse-reward RL agents** using emergent language abstractions as structured exploration bonuses; foundation of EReLELA.

Research Intern **2021 – 2022**
Digital Creativity Labs × Revolution Software · UK

- **Cut video-game asset creation effort** by developing style-transfer pipelines with 3D-aware representation learning, GANs and diffusion models.

Contract AI & Robotics Engineer **2016 – Present**
Independent · Remote

- **Shipped customer-facing, RAG-based chatbot assistants** with tool-calling over bespoke documents and Google APIs — advisor-assist-style workflows that automated repetitive client tasks (open-source LLMs on Lambda cloud GPUs, LangChain, genetically-evolved prompts).
- **Delivered production computer-vision systems**: on-screen 2D gaze tracking for hand-held devices (PyTorch); semi-supervised GAN for car make/model classification (TensorFlow).
- **Built robot autonomy**: navigation & planning for a Roomba-like robot (ROS, Gazebo) and domain-randomisation tooling (Blender, MakeHuman).

Graduate Teaching Assistant **2020 – 2023**
University of York, Computer Science Dept. · UK
Earned Associate Fellowship of the HEA (AFHEA - portfolio, 2021) via York Learning & Teaching Award course, alongside teaching:

- **Mathematical Foundations of Computer Science** (COM00013C)
- **Formal Languages and Automata** (COM00014C)
- Multi-Agent Interactions & Games (COM00009H)
- Intelligent Systems 1: Search & Representation (COM00020I)
- **Computability and Complexity** (COM00023I)
- Introduction to Software and Systems Engineering (COM00019I)

Education

PhD, Computer Science (Emergent Communication & Human-AI Cooperation) 2018 – 2024
University of York, UK — Thesis submitted 2024; award pending (revisions due Feb 2027)

Research MSc, Artificial Intelligence & Robotics 2015 – 2017
Université de Cergy-Pontoise, France — Taken concurrently with the ENSEA engineering degree. Thesis: visual contexts for spatial recognition in wide environments

MEng, Electrical Engineering & Information Science (GPA 3.7/4) 2016 – 2017
Osaka Prefecture University, Japan — Double-degree year abroad. Thesis: nonlinear system-based swarm robotics

Engineering Degree, Computer Science & Systems 2013 – 2017
ENSEA, France — French engineering *grande école*. Final-year project: 3D rigid-body physics & rendering engine (C/C++).

Classes Préparatoires aux Grandes Écoles (Mathematics & Physics) 2011 – 2013
Lycée Chateaubriand, Rennes, France — **Two years of full-time mathematics and physics**; subsequent entry to ENSEA by competitive national examination.

Skills

AI / ML PyTorch, TensorFlow, pandas, scikit-learn, deep RL (multi-agent, self-play, exploration, HER), emergent communication, GANs & diffusion models, representation learning, Gumbel-Softmax / differentiable programming, benchmark & evaluation design, ablation studies, computer vision (OpenCV, gaze tracking, object recognition)

LLMs / LLMOps HuggingFace transformers, LangChain (LangGraph), DSPy, RAG, Text2SQL, agents & tool calling (incl. NousResearch Hermes on local Qwen/Gemma & hosted APIs), prompt engineering & optimisation (incl. genetically-evolved prompts), LLM evals & LLM-as-judge, chain-of-thought scaffolding, fine-tuning (SFT, LoRA, GRPO), self-hosted inference (vLLM, Ollama, llama.cpp) & hosted APIs (OpenRouter), LLM observability & tracing (W&B Weave, LangSmith), adversarial auditing & red-teaming, LM inversion

Engineering / MLOps Python, C/C++, C#, Java, Linux & bash, HPC clusters (Slurm), cloud GPU compute (Lambda), AWS provisioning, FastAPI, multi-GPU training, experiment tracking (W&B, TensorBoard), Docker & Compose, Kubernetes, Git, CI (GitHub Actions), automated pre-commit linting & test hooks, pytest, open-source maintenance, profiling & performance optimisation, LaTeX

Robotics / Simulation MuJoCo, ROS & Gazebo, sim2real & domain randomisation, control theory (nonlinear, oscillator-based), sensor fusion & state estimation (EKF), SLAM-adjacent spatial reasoning, embedded systems (Raspberry Pi, servo control, 3D printing), physics engines (custom rigid-body), Blender

Research craft grant writing & funding acquisition, mentoring & PhD supervision, technical writing & peer review, experimental design & statistics, cross-functional collaboration (game studios, artists)

Languages French (native), English (bilingual, TOEFL 105), Japanese (professional), Spanish & German (conversant)

Recently added SQL, JAX, Terraform (IaC), GCP (BigQuery, Vertex AI, GCS), CD (GitHub Actions `terraform apply` on merge) — self-directed upskilling toward the AIOps stack, building on AWS provisioning and declarative container configuration (Docker Compose)

Publications

Peer-reviewed

- K. Denamganaï, On Compositional Learning Behaviours in Formal Mathematics, arXiv:2605.28512, 2026; accepted to the 3rd AI for Math Workshop @ ICML 2026. **(benchmark; empirical study; causal intervention)**
- S. Memery, K. Denamganaï, et al., CueTip: An Interactive and Explainable Physics-aware Pool Assistant, SIGGRAPH 2025.
- D. Hernandez, K. Denamganaï, et al., A Comparison of Self-Play Algorithms Under a Generalized Framework, IEEE Transactions on Games, doi:10.1109/TG.2021.3058898.
- K. Denamganaï, J. Walker, On (Emergent) Systematic Generalisation and Compositionality in Visual Referential Games with Straight-Through Gumbel-Softmax Estimator, 4th NeurIPS Workshop on Emergent Communication, arXiv:2012.10776, 2020.
- K. Denamganaï, J. Walker, ReferentialGym: A Nomenclature and Framework for Language Emergence & Grounding in (Visual) Referential Games, 4th NeurIPS Workshop on Emergent Communication, arXiv:2012.09486, 2020. **(nomenclature; open-source framework)**
- D. Hernandez, K. Denamganaï, et al., A Generalized Framework for Self-Play Training, IEEE Conference on Games (CoG), pp. 1–8, 2019.

- K. Denamganaï et al., Obstacle avoidance control law for two-wheeled mobile robots controlled by oscillators, 61st Annual Conference of the ISCIE, 221-4, 2017.
- K. Denamganaï et al., Coupled Kuramoto oscillator-based control laws for both formation and obstacle avoidance control of two-wheeled mobile robots, IEICE Technical Report, NLP2017-44, pp. 87–91, 2017.

Preprints & work under review

- K. Denamganaï, Detecting Covert Collusion between AI Agents through Partner–Outsider Positive Listening, arXiv:2610.37849, 2026. **(behavioural metric for AI Collusion)**
- K. Denamganaï, On the Monitorability of Natural Language Autoencoders: Identification vs Discrimination, arXiv:2610.18727, 2026. **(formal framework; predictions tested)**
- K. Denamganaï, K. Subr, Language Model Inversion through End-to-End Differentiation, arXiv:2602.11044, 2026; to be submitted to TMLR.
- S. Memery, K. Denamganaï, et al., xInv: Explainable Optimization of Inverse Problems, arXiv:2506.11056, 2025.
- K. Denamganaï et al., EReLELA: Exploration in Reinforcement Learning via Emergent Language Abstractions, OpenReview preprint, 2024; to be submitted to TMLR.
- K. Denamganaï et al., ETHER: Aligning Emergent Communication for Hindsight Experience Replay, arXiv:2307.15494, 2023; revised 2026. **The revision adds a theorem and proof that the relabelling and predicate functions learned from a referential game cannot collapse to trivial solutions.**
- K. Denamganaï et al., Visual Referential Games Further the Emergence of Disentangled Representations, arXiv:2304.14511, 2023. **(formal metric definition; empirical study)**
- K. Denamganaï et al., Meta-Referential Games to Learn Compositional Learning Behaviours, arXiv:2207.08012, 2022; to be submitted to TMLR. **(formal definition; benchmark operationalisation)**

Scientific Leadership

- **Reviewed 50+ manuscripts** for NeurIPS, ICLR, ICML, AAAI (2021–2026); Gold Reviewer Award, ICML2026; Best Reviewer Award, ICLR2022 EmeCom Workshop.
- Co-organizer, ICLR 2022 EmeCom Workshop; invited member, NeurIPS Workshop Proposal Review Committee (2021).

Certifications

- 2025 — Reinforcement Fine-Tuning LLMs with GRPO · DeepLearning.AI
- 2024 — LLMOps · DeepLearning.AI
- 2023 — Quality & Safety for LLM Applications · DeepLearning.AI
- 2023 — Building & Evaluating Advanced RAG · DeepLearning.AI
- 2023 — LangChain for LLM Application Development · DeepLearning.AI
- 2023 — Machine Learning in Weather and Climate · ECMWF
- 2021 — **Associate Fellowship of the HEA (AFHEA) (portfolio) · York Learning & Teaching Award**
- 2017 — Deep Learning Foundation Nanodegree · Udacity
- 2015 — **Underactuated Robotics (6.832x) · EdX**
- 2015 — Autonomous Mobile Robots (AMRx) · EdX
- 2014 — Autonomous Navigation for Flying Robots (AUTONAVx) · EdX
- 2014 — Computational Neuroscience · Coursera

Open-Source Projects

- **ReferentialGym** / **Regym** / **Archi** — PyTorch frameworks for emergent communication, multi-agent self-play deep RL, and modular architectures.
- **S2B** / **S2B-LM** — benchmark for compositional learning behaviours in agents and language models.
- **DIPhyR-Gym** — Gym environment for deliberate & intuitive physics reasoning.
- **HaRo** / **Core** / **SIMULATOR** — 3D-printable humanoid robot; **computer algebra system and 3D physics engine in C/C++**.
- **Go-Sensei** — Go-teaching LLM agent (LangGraph + FastAPI) with RAG and KataGo-grounded move evaluation; GCP stack (BigQuery, Vertex AI, GCS) provisioned via Terraform, with pytest + GitHub Actions CI/CD.