Call for PhD candidates

Autotelic Exploration for Automatic Evaluation of Large Generative AI Models: Adaptive Discovery and Mapping of Capabilities and Vulnerabilities

Supervision: Pierre-Yves Oudeyer (Flowers AI & CogSci Project Team, Inria, http://www.pyoudeyer.com
Location: Flowers AI & CogSci Lab, Inria, Bordeaux, France
Duration: 3 years
Deadline for application: Applications received before 30th june will receive priority, and will be reviewed in order of arrival. Start date September/October.
Language of work: English
Funding: The position is funded through an Inria doctoral contract (~2,100 €/month gross, ~1,750 € net).

Context: The Flowers team (Inria) has pioneered fundamental research in curiosity-driven autotelic learning and open-ended AI (Oudeyer et al., 2007; Colas et al., 2022; Gaven et al., 2025), and applications in domains ranging from robotics (Forestier et al., 2022) to educational technologies, with publications at NeurIPS, ICLR, and Nature Reviews. This project aims to extend the fundamental research in autotelic and open-ended AI and apply it to the societally important challenge of AI safety and discovery of emergent failures and capabilities.

Evaluating large language models (LLMs) represents a major challenge given their rapid evolution. Traditional benchmarks, often saturated and present in training data, struggle to reveal the true diversity of behaviors. This project proposes using autotelic curiosity algorithms to drive exploration by genAI models (Pourcel et al., 2024), inspired by human curiosity (Gottlieb & Oudeyer, 2018), to automatically generate adaptive evaluations that co-evolve with model capabilities.

Objectives and Methods. We aim to develop a unified framework combining automatic benchmark generation and vulnerability discovery. The approach relies on autotelic curiosity algorithms, of which quality-diversity (QD) algorithms form a particular family. These algorithms systematically explore a space of goals/problems by simultaneously optimizing local quality and global diversity according to semantic descriptors. We will generalize the ACES (Pourcel et al., 2024) and ACD (Lu et al., 2025) approaches to create adaptive benchmarks covering multiple domains and generating problems that are both diverse and unlikely to appear in training data. The iterative process maintains an evolving archive of problems/attacks: at each iteration, the system samples a target niche, selects examples from the archive, then uses a generator LLM to create new tests conditioned on those examples. We will adapt these principles to red

teaming approaches (Samvelyan et al., 2024; Lee et al., 2024), e.g. by using the recently developed methods for characterizing biases in large models (Kovac et al., 2024; Perez et al., 2025).

A key contribution is the use of LLMs to automatically cluster vulnerability findings into human-interpretable categories and produce synthetic reports detailing the strengths and weaknesses of the evaluated models. We will integrate a meta-learning dimension allowing the system to learn to predict the most informative niches based on past history, thereby optimizing its exploration strategy. On the methodological side, this project extends ACES and ACD to new domains, unifies benchmark generation and red teaming, and develops self-adaptive methods that evolve through meta-learning. Practical impact includes reducing evaluation costs, automatically discovering vulnerabilities before deployment, and proactively identifying dangerous emergent behaviors. This project thus proposes a paradigm shift: evaluation systems that continuously explore the space of possible behaviors (and their diversity), remaining relevant in the face of the rapid evolution of generative AI.

The project affords national and international collaborations with other research teams (both academic and industrial).

Expected output: publications in major AI conferences (Neurips, ICLR, ICML, etc), impactful open-source code, and societal impact through collaboration with (inter)national public agencies on AI safety.

Required background: Strong mathematical foundations (probability, optimization, linear algebra); solid Python programming and experience training/fine-tuning neural networks/genAI; Solid working knowledge in post-training, reinforcement learning, NLP. To be eligible, candidate must hold a master-level (or equivalent) diploma.

A plus: Familiarity with LLMs (transformers, RLHF, prompting strategies), quality-diversity algorithms, or red teaming literature; prior research experience (internships).

References: Gottlieb & Oudeyer (2018), Pourcel et al. (2024), Lu et al. (2025), Samvelyan et al. (2024), Lee et al. (2024), Perez et al. (2025), Kovač et al. (2024)

How to apply

Send to pierre-yves.oudeyer@inria.fr with [application] in the subject line:
1. CV
2. Motivation letter: what draws you to this project and which scientific directions interest you most, and describe your most relevant ML project or research experience in detail.
3. Links to code repositories: especially training code you have written.
4. Reports or write-ups of previous research projects (not necessarily on this topic).
5. Recent academic transcripts.

Shortlisted candidates will be invited to interviews.

Full References:

Colas, C., Karch, T., Sigaud, O., & Oudeyer, P. Y. (2022). Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. Journal of Artificial Intelligence Research, 74, 1159-1199.

Forestier, S., Portelas, R., Mollard, Y., & Oudeyer, P. Y. (2022). Intrinsically motivated goal exploration processes with automatic curriculum learning. Journal of Machine Learning Research, 23(152), 1-41.

Gaven, L., Carta, T., Romac, C., Colas, C., Lamprier, S., Sigaud, O., & Oudeyer, P. Y. (2025). MAGELLAN: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces. ICML.

Gottlieb, J., & Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12), 758–770.

Kovač, G., Portelas, R., Sawayama, M., Dominey, P. F., & Oudeyer, P. Y. (2024). Stick to your role! Stability of personal values expressed in large language models. *Plos One*, 19(8), e0309114.

Lee, S., Kim, M., Cherif, L., Dobre, D., Lee, J., Hwang, S. J., ... & Jain, M. (2024). Learning diverse attacks on large language models for robust red-teaming and safety tuning. *arXiv preprint* arXiv:2405.18540.

Lu, C., Hu, S., & Clune, J. (2025). Automated capability discovery via model self-exploration. In *Scaling Self-Improving Foundation Models without Human Supervision*.

Oudeyer, P. Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. IEEE transactions on evolutionary computation, 11(2), 265-286.

Perez, J., Kovac, G., Léger, C., Colas, C., Molinaro, G., Derex, M., Oudeyer, P., & Moulin-Frier, C. (2025). When LLMs Play the Telephone Game: Cultural Attractors as Conceptual Tools to Evaluate LLMs in Multi-turn Settings. *ICLR*.

Pourcel, J., Colas, C., Molinaro, G., Oudeyer, P. Y., & Teodorescu, L. (2024). ACES: Generating a diversity of challenging programming puzzles with autotelic generative models. *Advances in Neural Information Processing Systems*, 37, 67627–67662.

Samvelyan, M., Raparthy, S. C., Lupu, A., Hambro, E., Markosyan, A., Bhatt, M., ... & Raileanu, R. (2024). Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37, 69747–69786.