Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (SD TIB PAN)

SUBJECT:
Investigating Covert Capabilities and Latent Behaviours in AI Models

SUPERVISOR:
dr hab. Inż. Szymon Łukasik, szymon.lukasik@nask.pl
NASK National Research Institute, ul. Kolska 12, 01-045 Warsaw

ASSISTANT SUPERVISOR:
dr inż. Sebastian Cygert, sebastian.cygert@nask.pl
NASK National Research Institute, ul. Kolska 12, 01-045 Warsaw

DESCRIPTION:
Recent advances in generative models (especially large language models) have enabled unprecedented performance across a wide range of tasks. At the same time, increasingly capable models exhibit unexpected behaviours that are difficult to monitor and control. These might include reward hacking, deceptive responses, alignment faking, evaluation awareness and other latent capabilities that emerge under specific conditions.
The goal of this PhD thesis is to investigate methods for understanding and control of the hidden behaviours of generative models. The research might encompass theoretical and empirical approaches, including mechanistic interpretability, latent variable modelling, safety evaluation.

Potential research directions:

- Detection of reward hacking and goal misgeneralization in language models.
- Identification of alignment-faking.
- Evaluation awareness, monitoring awareness, behaviour shifts between training, benchmarking, and deployment environments.
- Methods for controllable generation and latent conditioning.
- Development of monitoring and auditing tools for advanced AI systems.
- Reliability evaluation under rare-event and high-assurance settings.
- Safety-focused training and alignment techniques.

REQUIREMENTS:
- Solid background in deep learning and familiarity with transformer-based language and vision-language models
- Programming proficiency in Python and the PyTorch / Hugging Face ecosystem
- Experience with, or willingness to learn, large-scale model training and evaluation on HPC/GPU clusters
- Reading knowledge of the AI safety and multimodal learning literature
- Good command of written and spoken English
- Independence and motivation to pursue open-ended research

BIBLIOGRAPHY:

Latent Variable Models:

1. *Ruan, Yangjun, Neil Band, Chris J. Maddison, Tatsunori Hashimoto. "Reasoning to Learn from Latent Thoughts." arXiv:2503.18866, 2025.*
2. *Wu, Chen Henry, Sachin Goyal, Aditi Raghunathan. "Mode-Conditioning Unlocks Superior Test-Time Scaling." arXiv:2512.01127, 2025.*

Evaluation awareness:

3. *Hua, Tim Tian, Andrew Qin, Samuel Marks, Neel Nanda. "Steering Evaluation-Aware Language Models to Act Like They Are Deployed." arXiv:2510.20487, 2025.*
4. *Nayan, Nilesh, Aishwarya Sampath Kumar, Rishiraj Girmal, Shivani Anilkumar, Sankaran Vaidyanathan, David A. Nader Palacio, Reshmi Ghosh, Soundararajan Srinivasan. "Evaluation Awareness Is Not One Capability: Evidence from Open Language Models." arXiv:2606.23583, 2026.*
5. *Li, Changling, Terry Jingchen Zhang, Jie Zhang, Zhijing Jin, Sahar Abdelnabi, Maksym Andriushchenko. "Decomposing and Measuring Evaluation Awareness." arXiv:2605.23055, 2026.*

Alignment degeneracies:

6. *Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Evan Hubinger. "Alignment faking in large language models." arXiv:2412.14093, 2024.*
7. *Betley, Jan, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, Owain Evans. "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs." arXiv:2502.17424, 2025.*

Reward hacking:

8. *Wang, Xinpeng, Nitish Joshi, Barbara Plank, Rico Angell, He He. "Is It Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort." arXiv:2510.01367, 2025*

Evaluations:

9. *UK AI Safety Institute (2024–2026) Evaluation Reports*
10. *Anthropic Alignment Science Reports (2024–2026)*