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

SUBJECT:
Detection and mitigation of safety failures in vision-language 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:
Vision-language models (VLMs) exhibit safety failures absent in text-only models: combining image and text creates attack surfaces where harmful intent is split across modalities, and safety alignment effective for text often fails to transfer to the multimodal setting. Current approaches address this mainly through data-centric safety fine-tuning. VLGuard (Zong et al., 2024), for instance, reports near-complete mitigation by fine-tuning on a curated safety dataset. Such methods, however, only patch the harmful distributions present in their training data, offering no guarantee of generalization to unseen attacks or novel image-text combinations, and they treat safety as a behavioural input-output property without explaining why a model complies or refuses, leaving failures observable only after the fact. This doctoral research takes a complementary, mechanism-level approach. It will investigate the internal representations behind cross-modal safety failures in VLMs and use that understanding to develop detection and mitigation methods that generalize beyond a fixed training distribution and remain robust across modalities and languages.

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:
1. *Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J. (2021). Unsolved Problems in ML Safety. arXiv:2109.13916.*
2. *Lee, Y., Kim, K., Park, K., Jung, I., Jang, S., Lee, S., Lee, Y.-J., Hwang, S. J. (2025). HoliSafe: Holistic Safety Benchmarking and Modeling for Vision-Language Model. arXiv:2506.04704.*

3. *Liu, X., et al. (2024). MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. ECCV.*
4. *Zong, Y., Bohdal, O., Yu, T., Yang, Y., Hospedales, T. (2024). Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models (VLGuard). ICML.*
5. *Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.*