Sønderborg, Denmark
phongsakon@outlook.dk
phomarkon.github.io
phongsakonmarkkonrad
phomarkon
0009-0004-2521-7879
Google Scholar

Phongsakon Mark Konrad

Curriculum Vitae

Summary

I work on the foundations of how models learn, on whether the representations a model ends up with capture the world's real causal structure or only imitate it. This sits at the intersection of representation learning, causal representation learning, and the theory of learning, with counterfactual reasoning as the test I trust most for telling understanding from pattern matching. I work phenomenon first, building small settings with known causal ground truth so I can see what learning actually recovers, and I run the full research lifecycle end to end, from problem formulation and experimental design through engineering, writing, and venue selection. I am finishing a BSc at SDU and begin the MPhil in Machine Learning and Machine Intelligence at the University of Cambridge in October 2026, after an unconventional path through military service, startups, and a semester at HKUST.

Research Interests

Foundations of deep learning, representation learning, causal representation learning, identifiability, counterfactual reasoning, learning dynamics, and the theory of learning.

Education

2026– **MPhil, Machine Learning and Machine Intelligence**, *University of Cambridge*, Cambridge, UK, *admitted, begins Oct 2026*

2025 **Exchange Semester, Computer Science & Engineering**, *HKUST*, Hong Kong SAR
Specialisation courses in Machine Learning (COMP4211), Deep Learning in Computer Vision (COMP4471), Large Language Models (COMP4901B), and Reinforcement Learning (COMP4901Z), plus the postgraduate Data Visualisation (COMP6411D).

2023–2026 **BSc in Engineering (Software Engineering)**, *University of Southern Denmark*, Sønderborg, DK, GPA $\approx$ 10.6/12 ($\approx$ 3.7/4.0)
Expected June 2026. BSc thesis with T. L. Adam, *Heimdall: Only the Safe Shall Pass*, a conformal verifier between autonomous bidders and the Nordic balancing market, with code and model weights released.

Selected Publications

2026 **P. M. Konrad**, T. Tanyel, S. Ayvaz. *Self-Reports Do Not Identify Self-Models: An Identifiability Test for Counterfactual Reports*. Accepted (**oral**), *PhilML Workshop, ICML 2026*. [PDF]
In short. Self-reports are evidence about behaviour in a prompt, not proof of a self-model. Wrong-source demonstrations pull affect-state reports toward the source family unless mechanism binding holds.

2026 **P. M. Konrad**, T. Tanyel, S. Ayvaz. *Decoded but Unused: Instruction Tuning Routes Moral Framing into the Judgment Readout*. Accepted (poster), *ICML 2026 Workshop on Mechanistic Interpretability*. [OpenReview]
In short. Moral framing is linearly decodable in pretrained models but only causally routed to judgment after instruction tuning.

2026 **P. M. Konrad**, T. Tanyel, S. Ayvaz. *A Path Already Walked: On Inheriting Network-Neuroscience Tools for Mechanistic Interpretability*. Accepted (virtual poster), *ICML 2026 Workshop on Mechanistic Interpretability*. [OpenReview]
In short. Mech-interp should import the graph vocabulary of network neuroscience (modularity, rich clubs, motifs) under explicit contracts.

2026 N.-T. Thinh, Y. Du, **P. M. Konrad**, A. Narechania. *Fact-check Your Information (FYI): A Design Probe to Understand How People Actually Fact-check Data-Driven Articles*. Accepted, *IEEE VIS 2026*. [PDF]
In short. A browser-extension design probe with 22 readers finds three human and AI workflow archetypes for fact-checking data-driven journalism, with visualization as the main way people audit AI conclusions.

Research Experience

Jan 2026– **Research Collaborator**, *HKUST DataVISards group*, Hong Kong SAR
Machine learning and data-visualisation research.

Jan 2026– **Research Collaborator**, *SDU Data & Intelligence Lab*, Sønderborg, DK
Independently initiate and drive ML research from end to end, from problem formulation and experimental design to pipeline development, manuscript writing, and venue selection. Co-author with Assoc. Prof. Serkan Ayvaz across interpretability, NLP, medical imaging, and remote sensing.

Sep 2024–Dec 2025 **Research Assistant**, *SDU Data & Intelligence Lab*, Sønderborg, DK
Built ML pipelines end to end and contributed to study design, literature reviews, model development, and manuscript preparation. Supported grant proposals and multimodal dataset management.

Full Publication List

Peer-reviewed and accepted

Interpretability and safety (workshop)

1. **P. M. Konrad**, T. Tanyel, S. Ayvaz (2026). *Self-Reports Do Not Identify Self-Models: An Identifiability Test for Counterfactual Reports*. Accepted (**oral**), *PhilML Workshop, ICML 2026*. [PDF]
2. **P. M. Konrad**, T. Tanyel, S. Ayvaz (2026). *Decoded but Unused: Instruction Tuning Routes Moral Framing into the Judgment Readout*. Accepted (poster), *ICML 2026 Workshop on Mechanistic Interpretability*. [OpenReview]
3. **P. M. Konrad**, T. Tanyel, S. Ayvaz (2026). *A Path Already Walked: On Inheriting Network-Neuroscience Tools for Mechanistic Interpretability*. Accepted (virtual poster), *ICML 2026 Workshop on Mechanistic Interpretability*. [OpenReview]

AI for software engineering (workshop)

4. **P. M. Konrad**, T. L. Adam, R. Terrenzi, S. Ayvaz (2026). *Architecture Without Architects: How AI Coding Agents Shape Software Architecture*. Accepted, SAGAI Workshop, IEEE ICSA 2026. arXiv:2604.04990
5. T. L. Adam, **P. M. Konrad**, R. Terrenzi, F. G. Lukas, R. Yilmaz, K. Sierszecki, S. Ayvaz (2026). *CAKE: Cloud Architecture Knowledge Evaluation of Large Language Models*. Accepted, KDA-AI Workshop, IEEE ICSA 2026. arXiv:2604.05755
6. R. Terrenzi, **P. M. Konrad**, T. L. Adam, S. Ayvaz (2026). *A Reference Architecture for Agentic Hybrid Retrieval in Dataset Search*. Accepted, SAML Workshop, IEEE ICSA 2026. arXiv:2604.16394

Conference proceedings

7. N.-T. Thinh, Y. Du, **P. M. Konrad**, A. Narechania (2026). *Fact-check Your Information (FYI): A Design Probe to Understand How People Actually Fact-check Data-Driven Articles*. Accepted, *IEEE VIS 2026*. [PDF]
8. **P. M. Konrad**, T. Tanyel, S. Ayvaz (2025). *Beyond Major Floods: Deep Learning for Detecting Shallow Water Inundation in Agricultural Areas*. *KES 2025*, Procedia Computer Science, 270, 301–310. [open access]

Preprints and works in progress

Interpretability and safety

9. **P. M. Konrad**, T. Tanyel, S. Ayvaz (2026). *Acceptance Cards: A Four-Diagnostic Standard for Safe Fine-Tuning Defense Claims*. Preprint. arXiv:2605.10575
10. **P. M. Konrad**, T. L. Adam, A. C. H. Merrild, R. de Rosa, R. Terrenzi, T. Tanyel, S. Ayvaz (2026). *The Open-Box Fallacy: Why AI Deployment Needs a Calibrated Verification Regime*. Preprint. arXiv:2605.10601

Applied ML and other

11. **P. M. Konrad**, A.-A. Popa, Y. Sabzehmeidani, L. Zhong, M. Tripathy, A. Constantinescu, E. A. Liehn, S. Ayvaz (2025). *Challenges in Deep Learning-Based Small Organ Segmentation: A Benchmarking Perspective for Medical Research with Limited Datasets*. Preprint. arXiv:2509.05892
12. **P. M. Konrad**, C. H. Kunstmann-Olsen, J. Fiutowski, S. Ayvaz (2025). *Non-Destructive Prediction of Fruit Ripeness and Firmness Using Hyperspectral Imaging and Lightweight Machine Learning Models*. Preprint. arXiv:2604.22788

Teaching

2026 **Teaching Assistant, Artificial Intelligence (BSc)**, *University of Southern Denmark*
Lead exercise sessions and student supervision, and design hands-on lab assignments. Built an interactive AI course with step-by-step animations, used as supplementary material.

Invited Talks

2026 **Agentic Engineering in Practice**, House of Software Sønderborg (inaugural event). Invited to introduce agent skills, the Model Context Protocol, and plugin-based development to an audience of 100+ software engineers, engineering leaders, and founders.

Academic Service

2025– **Student Member**, Educational Committee, Software Engineering Programme, SDU. Represent students in curriculum and quality assurance discussions on course content, assessment, and study environment.

Honors and Awards

2026 **Nominated, Best Thesis Award**, Faculty of Engineering, University of Southern Denmark.

2026 **William Demant Fonden Grant** supporting graduate study.

2026 **Nominated, Best Startup Award** for DreamBear (SaturoLabs).

2024, 2025 **Top 10**, Danish National Championship in AI (DM i AI), competing solo against 50+ teams.

2023 **1st place**, SDU Case Competition (48-hour sustainability challenge, Danfoss and Linak cases).

2021 **Formal Recognition for Exemplary Service**, Bundeswehr, for setting a benchmark in dedication and professionalism.

2021 **Performance Bonus for Outstanding Achievement**, Bundeswehr, for sustained excellence and independent handling of complex tasks.

Professional Experience

2026– **Founder**, *SaturoLabs*, Denmark
Products at the intersection of AI and software engineering. Live products include claudeboyz.com, getproofz.com, and dreambear.app.

2024 **CTO & Co-Founder**, *Tutora ApS*, Sønderborg, DK
Led development of the company's web application from end to end.

2022–2024 **Co-CEO & Co-Founder**, *Yeager GmbH*, Remote
Built *stabil.ai*, an AI powerlifting training app with personalised plans via MRV and MEV and real-time feedback.

2017–2021 **Staff Duty Soldier**, *Bundeswehr (German Armed Forces)*, Glücksburg, DE
Led a small HR team administering 600+ soldiers. Twice decorated.

Technical Skills

Languages Python, JavaScript, SQL

ML PyTorch, Hugging Face, TransformerLens, scikit-learn, pandas, NumPy, Optuna, Weights & Biases

Tooling Git, Docker, Google Cloud Platform, React, React Native, Node.js

Languages

Fluent English (professional), German (native), Thai (native)

Basic Danish