Jan-Philipp Fränken

jphilipp@stanford.edu
janphilippfranken.github.io

Education

2023– **Postdoc**, Stanford University
- Advisors: Tobias Gerstenberg, Noah Goodman

2019–2022 **Ph.D. in Psychology**, The University of Edinburgh
- Advisors: Neil Bramley, Chris Lucas
- 4-month research visit to UC Berkeley with Steven Piantadosi

2018–2019 **M.S. in Cognitive Science**, University College London
- Advisor: David Lagnado

2015–2018 **B.S. in Psychology**, Maastricht University
- Advisor: Henry Otgaar

Recent Representative Works

* Equal Contribution

2024 **Self-Supervised Alignment with Mutual Information: Learning to Follow Principles without Preference Labels**
Jan-Philipp Fränken, Eric Zelikman, Rafael Rafailov, Kanishk Gandhi, Tobias Gerstenberg, Noah Goodman. *Preprint*

2024 **STaR-GATE: Teaching Language Models to Ask Clarifying Questions**
Chinmaya Andukuri*, **Jan-Philipp Fränken***, Tobias Gerstenberg, Noah Goodman. *Preprint*

2024 **Procedural Dilemma Generation for Evaluating Moral Reasoning in Humans and Language Models**
Jan-Philipp Fränken, Kanishk Gandhi, Tori Qiu, Ayesha Kawhaja, Noah Goodman, Tobias Gerstenberg. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Oral)

2023 **Understanding Social Reasoning in Language Models with Language Models**
Kanishk Gandhi*, **Jan-Philipp Fränken***, Tobias Gerstenberg, Noah Goodman. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track* (Spotlight)

2023 **Social Contract AI: Aligning AI Assistants with Implicit Group Norms**
Jan-Philipp Fränken, Sam Kwok*, Peixuan Ye*, Kanishk Gandhi, Dilip Amurugam, Jared Moore, Alex Tamkin, Tobias Gerstenberg, Noah Goodman. In *NeurIPS Socially Responsible Language Modelling Research Workshop* (Oral)

Grants and Scholarships

2023 Stanford HAI-Google Funding
2022 Visiting Researcher Scholarship, UC Berkeley
2019 PhD Scholarship, German Academic Scholarship Foundation
2019 ESRC Studentship, Scottish Graduate School of Social Science

Teaching

2021–2022 Teaching Assistant, Python Programming, The University of Edinburgh
2019–2022 Teaching Assistant, Statistics, The University of Edinburgh
2016–2018 Teaching Assistant, Statistics, Maastricht University