Trisevgeni Papakonstantinou

(+44) 7907517806
p.trisevgeni@gmail.com
https://trisevgenip.github.io/personal/

Profile

Cognitive scientist and AI safety researcher working on socio-technical evaluations of frontier AI systems, with a focus on how large-scale human-AI interaction shapes attitudes, beliefs, behaviour, and downstream societal impacts through experiments, randomised trials, and computational methods.

Technical expertise in Python and R for experimental design, NLP, LLM evaluation, and advanced statistical modelling (multilevel/hierarchical/mixed-effects models). Published work on persuasion and belief change, belief inference in LLMs, social influence, behavioural interventions, and human-AI interaction in long-context settings.

Currently PhD candidate at UCL (intern at DeepMind, Turing Institute) studying belief revision, mental models, and social learning in human and machine minds, with an emphasis on model evaluations that are action-relevant for safety, robustness, and deployment decisions.

Education

University College London (UCL) London, UK

Ph.D. in Experimental Psychology (Cognitive Science, Computational Social Science) 2021 - 2026 (expected)

- Thesis: *Belief revision in human and machine minds: the effects of accuracy, coherence, and resource-rationality*. Supervised by David Lagnado and Nichola Raihani.
- Research focus: belief updating, mental models, theory of mind, social learning, persuasion, and socio-technical impacts of AI systems.
- Methods: large-scale online experiments, computational modelling, NLP, LLM evaluation, and human-AI interaction studies.

MSc in Cognitive and Decision Sciences (Distinction) 2019 - 2020

National and Kapodistrian University of Athens Athens, Greece

MSc in Science and Technology Studies 2018 - 2019

BSc in Philosophy of Science 2014 - 2018

Research Experience

Google DeepMind | London, UK Sept 2025 - Feb 2026

Student Researcher (Sociotechnical Evaluations)

- Designed and executed research on delusion reinforcement and affective trajectories in long-context (100+ turn) conversational AI interactions, evaluating when frontier models amplify harmful beliefs or behaviours through extended context.
- Developed multi-turn safety evaluation frameworks to isolate how context features (memory, agreement patterns, conversational history, sycophancy) shape model behaviour across extended human-AI interactions, informing mitigations for high-stakes deployments.
- Built synthetic datasets and experimental protocols for evaluating Gemini's impact on user beliefs with a focus on actionable signals for model design and policy-relevant safeguards.
- Collaborated with research, red-teaming, and product teams to translate evaluation findings into safety strategies, interventions, and deployment guidance for frontier models.

Digital Trust Council | Cambridge, US Sept 2025 - Present
Co-Founder & Head of Research

- Co-founded initiative focused on AI governance, safety, and certification; leading research strategy on third-party evaluation, assurance mechanisms, and external checks for trustworthy AI systems.
- Directing large-scale consensus research project with 20+ researchers to develop governance frameworks, external validation protocols, and independent oversight mechanisms for safe AI deployment.

The Alan Turing Institute | London, UK Oct 2023 - Jul 2024
Research Fellow (PhD Enrichment Scheme)

- Developed a benchmark and evaluation framework for GPT-4 and Llama 2 to assess frontier models' capacity to infer human beliefs and implicit mental models from text.
- Published findings in EMNLP 2025 (Findings track): a systematic framework for evaluating belief inference in LLMs at scale, connecting cognitive science theory to AI safety and trustworthiness evaluations.
- Won first place at AI UK Research Pitch Competition (2024) for work on belief systems, AI influence, and implications for safe deployment of powerful language models.

University of Edinburgh | Edinburgh, UK May - Jun 2024
Visiting Researcher

- Designed and conducted large-scale computational social science project to infer beliefs and intuitive theories from social media data using NLP, probabilistic modelling, and LLM-based methods in R and Python.
- Built research pipelines for extracting and analysing behavioural text data at scale, linking unstructured online discourse to psychologically meaningful belief structures, value systems, and persuasion patterns.
- Developed methods for mapping belief networks and attitude change from naturalistic human communication, informing how public input and real-world discourse can shape model values and evaluations.

ShiftPartner | London, UK Aug 2022 - June 2026
Data Science Lead

- Lead data science strategy across analytics, machine learning, and recommender systems, directing technical work in Python, R, Cypher, and SQL for products with real-world behavioural impact.
- Oversee cross-functional collaboration across engineering, BI, and UX teams to deliver analytical systems and ML models, ensuring that evaluation, monitoring, and user-level outcomes inform design and deployment decisions.

Department of Health and Social Care / Public Health England | London, UK Jan 2021 - Jul 2022
Behavioural Science Advisor

- Led quantitative analysis for multiple large-scale randomised controlled trials (RCTs) and behavioural interventions using advanced statistical methods to evaluate intervention effectiveness and policy-relevant outcomes.
- Applied behavioural science frameworks (COM-B, TDF, social norms theory) and analysed national RCTs to generate evidence on behaviour change, persuasion, and social influence, translating nebulous policy problems into measurable constructs.
- Designed and executed behavioural data science projects integrating experimental evidence with computational analysis to inform public health policy during the COVID-19 pandemic.
- Co-authored systematic review and meta-analysis of social norms interventions on health behaviours (published in *Nature Human Behaviour*, 2024), analysing 89 studies with multilevel meta-analytic methods.

Technical Skills & Methods

Programming: **Python (advanced)**: research pipelines, NLP (transformers, spaCy, NLTK), LLM evaluation (APIs, HuggingFace), probabilistic modelling, data analysis (pandas, numpy, scikit-learn), synthetic data generation; **R (advanced)**: statistical modelling, experimental data analysis, meta-analysis, visualisation, multilevel/hierarchical/mixed-effects models; **SQL, Cypher** (working knowledge).

AI/ML Methods: Frontier model evaluation, LLM prompting and evaluation, NLP, synthetic data generation, fine-tuning, robustness and behaviour analysis of large-scale language models.

Experimental Methods: Large-scale online experiments (oTree, Gorilla, Qualtrics), randomised controlled trials, longitudinal studies, human-AI interaction studies, survey design, behavioural interventions, experimental design and power analysis.

Statistical Methods: Multilevel/hierarchical/mixed-effects regression models (lme4, brms, lmerTest), meta-analysis (multilevel models, metafor), causal inference, Bayesian modelling, computational modelling of belief revision and social learning.

Tools: Git, Jupyter, Neo4j, Colab, LaTeX, SPSS, AWS.

Front-end: JavaScript, Python, oTree, HTML.

Publications

Conference Papers

1. **Papakonstantinou, T.**, Zhiteneva, A., Ma, A.Y., Powell, D., and Horne, Z. (2025). *Evaluating Large Language Models for Belief Inference: Mapping Belief Networks at Scale*. Findings of the Association for Computational Linguistics: EMNLP 2025, 20787.
2. **Papakonstantinou, T.** and Lagnado, D. (2025). *Cognitive coherence and resource rationality: Rethinking resistance to belief change*. Proceedings of the 47th Annual Meeting of the Cognitive Science Society.
3. **Papakonstantinou, T.**, Leong, K.I., and Lagnado, D. (2024). *Humans generate auxiliary hypotheses to resolve conflicts in observational data*. Proceedings of the 46th Annual Meeting of the Cognitive Science Society.
4. **Papakonstantinou, T.**, Raihani, N.J., and Lagnado, D. (2024). *Belief updating patterns and social learning in stable and dynamic environments*. Proceedings of the 46th Annual Meeting of the Cognitive Science Society.
5. Franklin, M., **Papakonstantinou, T.**, Chen, T., Fernandez-Basso, C., and Lagnado, D. (2023). *Blame attribution in human-AI and human-only systems: Crowdsourcing judgments from Twitter*. Proceedings of the 45th Annual Meeting of the Cognitive Science Society, 45(45).
6. Franklin, M., **Papakonstantinou, T.**, Chen, T., Fernandez-Basso, C., and Lagnado, D. (2023). *An Unsupervised Approach to Extracting Knowledge from the Relationships Between Blame Attribution on Twitter*. International Conference on Flexible Query Answering Systems, 221–233.

ArXiv

1. Gutoreva, A., Tsim, F., and **Papakonstantinou, T.** *Position: AI as Part of Self–Extending the Mind Requires Cognitive Co-Regulation*. arXiv preprint arXiv:2605.16197.
2. **Papakonstantinou, T.** and Horne, Z. (2023). *Characteristics of persuasive deltaboard members on Reddit's r/ChangeMyView*. PsyArXiv.

Journal Articles

1. **Papakonstantinou, T.**, Flecke, S.L., Edmunds, C.E.R., Cross, R., Tran, A., and Gold, N. (2025). *A systematic review and meta-analysis of the effectiveness of social norms messaging approaches for improving health behaviours in developed countries*. Nature Human Behaviour, 1–19.
2. Player, L., Hughes, R., Mitev, K., Whitmarsh, L., Demski, C., Nash, N., **Papakonstantinou, T.**, et al. (2025). *The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA)*. Psychological Methods.
3. Towler, L.*, Bondaronek, P.*, **Papakonstantinou, T.**, Amlôt, R., Chadborn, T., Ainsworth, B., and Yardley, L. (2023). *Applying machine-learning to rapidly analyze large qualitative text datasets to inform the COVID-19 pandemic response: Comparing human and machine-assisted topic analysis techniques*. Frontiers in Public Health, 11, 1268223.
4. Bondaronek, P., **Papakonstantinou, T.**, Stefanidou, C., and Chadborn, T. (2023). *User feedback on the NHS test & Trace Service during COVID-19: The use of machine learning to analyse free-text data from 37,914 England adults*. Public Health in Practice, 6, 100401.
5. van Leeuwen, F., Van Lissa, C.J., **Papakonstantinou, T.**, Petersen, M.B., et al. (2024). *Morality as Cooperation, Politics as Conflict*. Social Psychological Bulletin, 19, 1–22.
6. Thom, J., Bowen, S., Yang, Y., Devarajan, S., Doran, H., Zampetis, M., **Papakonstantinou, T.**, et al. (2024). *The effect of timers and precommitments on handwashing: A randomised controlled trial in a kitchen laboratory*. Behavioural Public Policy, 1–25.

Awards & Fellowships

- OpenAI Researcher Access Program (2 grants of $10,000 in API credits) - OpenAI
- First Place, Research Pitch Competition - AI UK, The Alan Turing Institute (2024)
- PhD Enrichment Fellow - The Alan Turing Institute (2023-2024)
- EPS Study Visit Grant (£3,500) - Experimental Psychology Society
- UCL CDI Impact Accelerator (with ShiftPartner) - Centre for Digital Innovation, UCL
- Associate Fellowship - Higher Education Academy

Teaching & Supervision

Teaching: Lead PGTA for statistical methods in R (PALS0045, PALS0046, UCL); Module Co-Organiser for 280-student statistics module (2022-2023); PGTA for Behaviour Change, Advanced Multivariate Statistical Methods, HCI.

Supervision: Supervised 6 undergraduate and postgraduate students (BSc, MSc, MSci dissertations in experimental psychology, cognitive science, social cognition).

Professional Service

Reviewer: *Cognitive Science*, *British Journal of Psychology*, *Journal of Medical Internet Research*, Annual Meeting of the Cognitive Science Society, International Conference on Computational Social Science.

Affiliations: London Initiative for Safe AI, Cognitive Science Society, Experimental Psychology Society, AI Qual Network.

Referees

Prof. David Lagnado, Professor, UCL
d.lagnado@ucl.ac.uk

Dr. Matija Franklin, Research Scientist, Google DeepMind
matija.franklin@google.com

Canfer Akbulut, Research Scientist, Google DeepMind
akbulut@google.com

Dr. Zachary Horne, Lecturer, University of Edinburgh
zachary.horne@ed.ac.uk

Dr. Paulina Bondaronek, Senior Research Fellow, UCL Institute of Health Informatics
Lead Behavioural Scientist, ShiftPartner
p.bondaronek@ucl.ac.uk