Luke Hewitt
luke@transluce.org — lukehewitt.info
Updated: June 2026

Summary

I study the impacts and risks from AI models influencing humans' beliefs, using a combination of large scale randomized human experiments and simulation-based AI evaluations.

Professional

2025- ***Research Fellow, Transluce***
I currently lead persuasion and manipulation research projects conducted for the UK AI Security Institute and the EU AI Office.

2024-25 *AI safety consulting, UK AISI*
Design and analysis of experiments evaluating the human influence of frontier LLMs.

2024 *AI safety consulting, OpenAI Preparedness Team*
Design and analysis of OpenAI's first human persuasion evaluations (GPT-4o).

2022-24 *Co-founder/Director, Rhetorical Labs*
Conducting RCT experiments in the US, South America and Europe, to inform clients on the impact of messaging for issues such as vaccination and climate change.

2019-20 *Research data scientist, Swayable*
Experiment design and analysis for persuasion measurement and national polling.

2013 *Microsoft Swiftkey, Research intern*
Developed language model for emoji prediction (US patent: US10664657B2)

Education / academia

2025- ***Stanford University, Affiliate (AI for Public Benefit Lab)***
2023-25 *Stanford University, Senior Research Fellow (PACS)*
Research on using AI models to simulate human experimental effects in the social and behavioral sciences.

2025 *AI safety research consulting, London School of Economics*
Designing experiments to evaluate whether frontier AI models can persuade users to donate their time, money and name to a political cause.

2022-24 *SSRC Mercury Project, Co-Principal Investigator*
Co-Principal Investigator on research developing scalable online methods for generating targeted public-health messaging in Brazil, Mexico and the U.S.

2015-22 *MIT, PhD. (Computational Cognitive Science) Advisor: Josh Tenenbaum*
PhD thesis focused on scalable methods for explainable AI that integrate deep learning into structured Bayesian models, and collaborations with cognitive & social scientists to build models of human perception, emotion, learning and persuasion.

2014-15 *UCL Machine Reading Group.* Developer of WOLFE language

2010-15 *UCL (Computer Science Dept.), MEng.*
My program was *Mathematical Computation.* I graduated First Class with Honors, and received a Dean's List Commendation for Outstanding Academic Performance.

Research publications

Artificial intelligence can persuade people to take political actions
(Hewitt, Hackenburg* et al, in review at PNAS)*

Predicting results of social science experiments using large language models
(Hewitt, Ashokkumar* et al., Nature 2026)*

AI systems out-persuade expert humans
(Hackenburg et al., in review)

The AI Epistemic Deference Index: A Continuous Measure of Sycophancy
(Botas et al., in review)

AI Epistemic Risks: Emerging Mechanisms & Evidence
(Yang et al., in review)

DeliberationBench: A normative benchmark for the influence of LLMs on users' views
(Hewitt et al, IASEAI 2026)

The levers of political persuasion with conversational AI
(Hackenburg et al., Science 2025)

Encouraging vaccination using the creativity and wisdom of crowds
(Tappin et al., working paper)

Outcome-based Reinforcement Learning to Predict the Future
(Turtel et al., TMLR 2025)

Large language models are more persuasive than incentivized human persuaders
(Schoenegger et al., in review at PNAS Nexus)

How will advanced AI systems impact democracy?
(Summerfield et al., Nature Human Behavior, 2025)

The impact of AI message-testing on public discourse
(Hewitt, IASEAI 2025)

Quantifying the returns to persuasive message-targeting using a large archive of campaigns own experiments
(Tappin, Hewitt, Coppock, APSA 2024)

GPT-4o System Card: Persuasion
(OpenAI, 2024)

How experiments help campaigns persuade voters: evidence from a large archive of campaigns own experiments
(Hewitt et al., APSR 2024)

Using survey experiment pre-testing to support future pandemic response
(Tappin & Hewitt, PNAS Nexus 2024)

Listening with generative models
(Cusimano et al., Cognition 2024)

Quantifying the persuasive returns to political microtargeting

(Tappin et al., PNAS 2023)

Emotion prediction as computation over a generative Theory of Mind
(Houlihan et al., Phil. Trans. A, 2023)

DreamCoder: growing generalizable, interpretable knowledge with wake-sleep bayesian program learning
(Ellis et al., Phil. Trans. A, 2023)

Rank-heterogeneous effects of political messages: Evidence from randomized survey experiments testing 59 video treatments
(Hewitt et al., working paper)

Hybrid memoised wake-sleep: Approximate inference at the discrete-continuous interface
(Le et al., ICLR 2022)

DreamCoder: Bootstrapping Inductive Program Synthesis with Wake-Sleep Library Learning
(Ellis et al., PLDI 2021)

Estimating the Persistence of Party Cue Influence in a Panel Survey Experiment
(Tappin et al., JEPS 2021)

Learning to learn generative programs with memoised wake-sleep
(Hewitt et al., UAI 2020)

Inferring structured visual concepts from minimal data
(Qian et al., CogSci 2019)

Learning to infer program sketches
(Nye et al., ICML 2019)

The Variational Homoencoder: Learning to learn high capacity generative models from few examples
(Hewitt et al., UAI 2018)

Auditory scene analysis as Bayesian inference in sound source models
(Cusimano et al., CogSci 2017)

Workshops organized

Workshop on *Risk modeling for AI Harmful Manipulation*	EU AI Office 2026
First workshop on *AI Manipulation and Info Integrity*	IASEAI 2026
AI Manipulation Hackathon	Apart Research, 2026
Symposium on *LLMs and the Future of Social Psychology*	SPSP 2025

Funding awards

I acknowledge research funding from:

- Tender for Technical Support on GPAI Safety, EU AI Office 2025
- Challenge Grant, UK AI Security Institute (AISI) 2025
- Future of Life Foundation 2025
- Phauna Foundation 2023
- The Mercury Project, SSRC 2022-23

- Future Fund (FTX) 2022
- Sandbox Innovation Fund, MIT 2021
- Multidisciplinary University Research Initiatives (MURI) 2020-21
- MIT-IBM Watson AI Lab 2018-19
- Henry E. Singleton (1940) Fellowship 2015-17
- Engineering and Physical Sciences Research Council 2014

Other recognition of talent

- Fellowship on AI for Human Reasoning, FLF 2025
- Alien of Extraordinary Ability (O-1) 2025
- Member, South Park Commons 2024
- Teaching awards:
  - Walle Nauta Award for Continued Dedication to Teaching, MIT 2019
  - Angus MacDonald Award, MIT 2018
  - Angus MacDonald Award, MIT 2017
  - Flame challenge finalist, Alan Alda Center 2016
- Dean's List Commendation for Outstanding Academic Performance, UCL 2015