Nicholas Clarke, PhD

Independent Researcher — AI Alignment & Moral Philosophy

Cambridge, UK
Email: nclarke431@gmail.com

RESEARCH FOCUS

AI alignment and safety; moral discovery and value idealisation; idealised human preferences as alignment targets; naturalistic moral theory (Goal Theory); epistemology and the theory of reasons under conditions of epistemic non-ideality.

EDUCATION

PhD in Philosophy — University of London
Awarded with no corrections (rare distinction)
Thesis: *Goal Theory* — a naturalistic account of idealised human desire

MA in Philosophy (Distinction) — University of London
Dissertation applying Bayesian reasoning to complex evidential claims under uncertainty

---

RESEARCH OUTPUTS (SELECTED)

- *Beyond Behavioural Control: Idealised Human Desire as an Alignment Target in AI* — submitted to *Minds and Machines*
- A programme of academic work developing *Goal Theory*, a naturalistic framework for morality, with applications to moral disagreement, moral luck, non-identity, and metaethical semantics
- Independent work in epistemology and the theory of reasons, developing non-factive operational accounts of reasons and knowledge suited to deliberation under epistemic uncertainty
- Working papers on empirical preference modelling, alignment target robustness, and human-in-the-loop idealisation

---

ACADEMIC & TEACHING EXPERIENCE

Associate Lecturer — AI Ethics & Technology
Anglia Ruskin University
Research-informed teaching on AI risk, alignment challenges, and governance

---

PROFESSIONAL EXPERIENCE (COMPRESSED)

University of Cambridge — Reporting & Information Analyst
Empirical analysis, structured evaluation, and decision-focused reporting across large organisational datasets.

Prior roles in technical research, documentation, analytics, and software development, including senior positions in enterprise systems and SaaS environments.

METHODS & SKILLS

Research design • conceptual analysis • evidence synthesis • experimental design • human-in-the-loop studies • quantitative analysis • systems thinking • failure-mode analysis • research communication • interdisciplinary collaboration

---

PUBLIC WRITING

Essays on AI alignment, ethics, and moral theory (Substack).