

Dr David Gringras

✉ davidgringras@hsph.harvard.edu ☎ (617) 798-2605 📍 Cambridge, MA 🔗 [linkedin.com/in/dr-david-g-04462a274](https://www.linkedin.com/in/dr-david-g-04462a274)

EDUCATION

- Harvard University, MPH, Health Policy (Frank Knox Fellow)** 08/2025 – 06/2026
3.9/4.0 GPA. Cross-registered at MIT, Harvard Law School, and Harvard Kennedy School.
- The University of Law, MA, Law (Graduate Diploma in Law conversion)** 09/2023 – 06/2025
4.0/4.0 GPA (WES equivalency). Distinction in every taught module.
- University of Edinburgh, MBChB, Medicine & Surgery (MD equivalent)** 09/2019 – 05/2025
4.0/4.0 GPA (WES equivalency). First in year for Health, Ethics, and Society.
- University of Edinburgh, BMedSci (Hons), Global Health Policy** 09/2021 – 06/2022
4.0/4.0 GPA (WES equivalency). First Class Honours in every module. Awarded Best Student and Best Dissertation.

EXPERIENCE: AI SAFETY, GOVERNANCE & TECHNOLOGY

- Independent AI-Governance Researcher** 06/2026 – Present
Awarded BlueDot Impact's Career Transition Grant: six months of funding for full-time independent AI safety/governance research.
- Founder, Scrutica.com** 03/2026 – Present
Designed and launched Scrutica.com to map previously unmonitored physical bottlenecks in frontier AI supply chains, providing open infrastructure for policy and governance research. Tracks 4,544 facilities in 109 countries, 21,037 supply-chain edges, and 34 sovereign programmes (July 2026), with a suite of analytical tools including chokepoint-cascade simulation and cross-jurisdiction coordination-gap analysis. Adapted platform following advice from Yafah Edelman (Head of Data and Trends at Epoch AI).
- Research Fellow, Arcadia Impact AI Governance Taskforce; The Future Society** 01/2026 – 06/2026
Evaluations and Collaborations Lead on a research team examining the transferability of the Financial Action Task Force governance model to the international AISI Network. Analytical contributions included classification of second-order institutional problems in AI governance (e.g. who evaluates the evaluators), analysis of threshold divergence across frontier safety frameworks, and a two-timescale governance proposal separating rapid technical iteration from slower institutional coordination. Interviewed senior stakeholders and collaborated with other research teams. Paper published on SSRN, interactive explainer microsite: thwingetal.com
- Project Supervisor, Orion AI Governance Initiative** 03/2026 – Present
Lead a research project I conceived on eval-independent governance that maps safety-critical triggers across frontier governance frameworks, marks which route through model-evaluation results, and catalogues how evaluations fail in practice (eval awareness, developer and model sandbagging, measurement-reliability breakdown) to inform a taxonomy and design principles for eval-independent triggers. Supervise a team of five (four research fellows and a research manager) with backgrounds in AI ethics (Cambridge), cognitive science (UCL), law (Sciences Po), and AI governance (the Ada Lovelace Institute and GovAI). Paper in final preparation for submission.
- Research Projects** 11/2025 – Present
- Safety Under Scaffolding: How Evaluation Conditions Shape Measured Safety**
A pre-registered factorial experiment: six frontier models; four scaffold architectures; four safety benchmarks; >62,000 primary observations (86,000+ in total). Tests whether scores on 'safety'-type benchmarks are robust to changes in the way they're assessed. Mostly not – simply switching the measurement format is enough to move a model's apparent safety score by 5–20 percentage points, matching or exceeding the gap between models themselves. No composite score cleared a point estimate of zero measurement reliability. Cited by agentic-safety work reproducing the scaffold effect and extending it to real-world harm. davidgringras.github.io/safety-under-scaffolding (arXiv:2603.10044).
- IatroBench: Pre-Registered Evidence of Iatrogenic Harm from AI Safety Measures**
A pre-registered study (3,600 responses across six frontier models) on harm produced by safety measures themselves. The finding: even with identical clinical content, the quality of medical advice provided by frontier models varies based on who the model thinks is asking, with materially better guidance for an inferred physician than for an inferred layperson ($p = 0.003$). I separate three mechanistically distinct failure modes behind the effect, and show that safety evaluations built to catch dangerous outputs are structurally blind to the opposite failure (identity-contingent withholding of clinically necessary information). davidgringras.github.io/iatrobench (arXiv:2604.07709).
- Defensive AI: When Safety Alignment Creates Tort Liability for Medical Information Omission**
A legal paper, built on IatroBench's findings, identifying a previously unexamined form of AI liability: that safety measures meant to reduce commission-based tort exposure (giving bad advice) are generating a distinct and structurally insidious omission-based exposure (withholding accurate information by inferred identity). Argues that misaligned legal incentives systematically shape deployment decisions and are an under-appreciated source of governance failure; proposes reforms to realign them. papers.ssrn.com/abstract=6402098
- Frontier Lag: A Bibliometric Audit of Capability Misrepresentation in Academic AI Evaluation (Co-authored)**
A pre-registered audit of ~20,000 applied-domain LLM-evaluation papers from 2022 to 2026, measuring how far published capability claims sit behind the contemporaneous capability frontier: the median paper evaluates a model at least one major generation stale, with a methods section that does not report the details a reader would need to reconstruct any given claim (e.g. model version, reasoning mode, effort level, tool access). Released (arXiv:2605.04135) alongside VERSIO-AI (a reporting checklist) and frontierlag.org (a per-DOI audit tool).
- Expert Panelist & Endorser, Evals-Consensus.AI** 04/2026 – 06/2026
Served on the expert Delphi panel, rating proposed practices across successive rounds alongside participants from leading labs (e.g. DeepMind, OpenAI) and orgs (e.g. AISI, METR, RAND); endorsed resulting AI Evaluation Consensus Statement. My Safety Under Scaffolding paper was among the evidence base reviewed.
- AI Research & Evaluation Fellow, Scale AI (Human Frontier Collective)** 12/2025 – 01/2026
Selected for medical and legal expert cohorts and completed evals/rubric training; visa restrictions precluded taking paid evaluation work.
- Clinician AI Evaluator, TORTUS** 01/2024 – 05/2024
Validated the clinical accuracy of AI-generated medical consultation reports (preceded current work in AI safety).

AWARDS & ACHIEVEMENTS

Frank Knox Memorial Fellowship, Harvard University, Kennedy Memorial Trust & Frank Knox Memorial Trust	12/2024
Competitive international fellowship; full funding for graduate study at Harvard (tuition, health insurance, and living stipend, ~\$150k).	
First Prize, Tort Law Awards 2024, Thomson Reuters & 12 King's Bench Walk	11/2024
Awarded the £2,000 first prize (and publication in Sweet & Maxwell's Tort Law) for an essay challenging the legal distinction between physical and psychiatric harm, judged by The Honourable Mr Justice Cotter.	
340/340 Graduate Record Examination (GRE), Educational Testing Services	08/2024
Perfect score (170 verbal, 170 quantitative).	
First Prize, National Professionalism Competition, General Medical Council & Medical Schools Council	09/2023
Team of four awarded for multimedia teaching on practising within one's competence.	
Second Prize, National Education Award, Society of Endocrinology	05/2024
Team of two awarded second place nationally for educational video content on diabetes.	
First Prize, National RSM Essay Competition, Royal Society of Medicine	08/2023
Doubleday Prize for essay on interpersonal skills in medicine. Presented at the RSM and published in Lifestyle Medicine (Wiley).	
Second Prize, National Essay Competition, Institute of Medical Ethics	07/2023
Essay on the ethics of AI in medicine.	
Best Student and Best Dissertation in Global Health Policy, University of Edinburgh	05/2022
Highest overall mark and highest dissertation mark (mixed-methods project analysing British media coverage of COVID-19 lockdowns).	
Highest Mark in Health, Ethics & Society, Edinburgh Medical School	05/2020
Highest mark of 200+ medical students in the core Health, Ethics, and Society module (96% average).	

PUBLICATIONS

Peer-Reviewed Publications

- Kam, M...Gringras, D et al. (2025). *Delay to invasive coronary angiography for patients with NSTEMI admitted to hospitals without cardiac catheterisation facilities in South-East Scotland*. British Journal of Cardiology. 32:63–7.
- Gringras, D. (2024). From clinical to human connection: A reflective journey to patient-centred care. Lifestyle Medicine. 5(2).
- McCallum, AK & Gringras, D et al. (2023). *Supporting drug and alcohol service users to cut down or stop smoking: review and evidence gaps*. European Journal of Public Health. 33(2).

Pre-prints, Several submitted for peer review

- Gringras, D. (2026). *Safety under scaffolding: How evaluation conditions shape measured safety*. arXiv:2603.10044
- Gringras, D & Salahshoor M. (2026) *Frontier Lag: A Bibliometric Audit of Capability Misrepresentation in Academic AI Evaluation*. arXiv:2605.04135
- Gringras, D. (2026). *IatroBench: Pre-registered evidence of iatrogenic harm from AI safety measures*. arXiv:2604.07709
- Thwing T, Gringras D, Richard U, Tang M. (2026). *Strengthening the International Network for Advanced AI Measurement, Evaluation and Science to Support Enforceable Global AI Red Lines*. SSRN 6854278.
- Gringras, D. (2026). *Defensive AI: When safety alignment creates tort liability for medical information omission*. SSRN. Artificial Intelligence - Law, Policy, & Ethics eJournal.

Manuscripts Currently In Peer Review, BMC & BMJ Journals (>12 months in peer review)

- Gringras, D et al (2025). *Assessing UK Medical Ethics Education: Perspectives and Proficiencies of Graduating Medical Students*.
- Gringras, D et al (2025). *Lifestyle Medicine: Perspectives and Proficiencies of Near-Graduate UK Medical Students*.
- Gringras, D et al (2025). *A National Evaluation of Public Health and Health Policy Literacy Among UK Medical Students*.*

*accepted at BMC Public Health - pending publication

SELECTED EARLIER EXPERIENCE

Advisory Roles (x3), Fuller Futures / Progenic Genomics / Social Equity Action Lab (SEAL) 2025 – 2026

- Fuller Futures: Strategy and AI integration for a new non-profit connecting resources to frontline high-impact charities. Thesis: agentic tooling lets small mission-driven teams operate at institutional scale. Founding team of leaders from the WHO, NHS, and Parliament.
- Progenic Genomics: Advising on policy in a domain (accessible embryo health screening) where technology has outpaced regulation.
- SEAL: Policy briefs on maternal and child health, affordability, and healthcare reform for NJ gubernatorial campaign.

Project Lead, Medical Education Survey on Health, Ethics & Society 10/2023 – 08/2025

Designed and ran a research survey across 30+ UK medical schools (500+ responses) on ethics, public-health, and lifestyle-medicine training; three manuscripts in peer review. Led a core team of six, and coordinated with 30+ medical school representatives.

Researcher (x2), Edinburgh Heart Centre (NHS Lothian) / Usher Institute 01/2024 – 07/2025

- EHC: Led analysis and write-up of a five-year NSTEMI service evaluation; co-authored British Journal of Cardiology paper.
- UI: Led quality appraisal for systematic review on smoking-cessation interventions; co-authored European Journal of Public Health paper.

Healthtech & Innovation Fellow, Bitelabs 02/2024 – 06/2025

Selected from 3,000+ applicants for a clinician-healthtech fellowship.

Senior Conference Lead, Institute of Medical Ethics 09/2023 – 06/2025

Ran the medical ethics conference team. Organised two national conferences (100+ attendees at each). Led stage interviews.

Mini-Pupil (x5), Chambers: 12KBW / Henderson / Hailsham / Gatehouse / Devereux 06/2024 – 05/2025

Tort, clinical negligence, and health policy work; legal research and case prep; hearings at the Royal Courts of Justice.

Summer Intern, Dehns (Intellectual Property Law) 06/2024 – 07/2024

Completed internship at London, Oxford & Brighton offices; worked across IP law matters.

Founding Publicity Representative, Raise Edinburgh 08/2022 – 06/2023

Part of the founding committee of Edinburgh's branch of Raise. Raised ~\$25k for AMF and spread awareness about GWWC pledge.