

Dr Chris Hicks

London, UK  chrishicks@mailbox.org  +44 7938875365 
[Google Scholar](#)  [Homepage](#)  [LinkedIn](#) 

Research scientist specialising in **reinforcement learning**, **language model agents** and the **security of machine learning systems**. I focus on two key areas, both at the intersection of AI and cyber security: **applying agents to security problems** including autonomous cyber defence, malware detection and autonomous vulnerability discovery, and **the security and robustness of ML systems** themselves including **data poisoning**, **backdoors** and **red teaming**. My work spans **50+ publications** at venues including **USENIX Security**, **PoPETs**, **NeurIPS**, **ICML**, **AAAI** and **TMLR**, with recent research alongside the **UK AI Security Institute** and **Anthropic** on data-poisoning vulnerabilities in LLMs. I have a decade of hands-on experience in computer security and move quickly from idea to validated, reproducible result.

Experience

Principal Research Scientist

The Alan Turing Institute | October 2022 – Present

- Develop **LLM and RL methods for autonomous cyber security**. Entity-based agents, deep RL reward design, robustness to distribution shift, and environments for training and evaluation that are aligned to real-world conditions.
- Research the **security of ML systems and training pipelines: data-poisoning and backdoor** attacks and defences for LLMs and deep RL agents.
- **50+ peer-reviewed papers** spanning deep RL, LLMs, AI agent benchmarking, and security for software, networks and ML systems.
- Co-authored, with **Anthropic and the UK AI Security Institute**, “Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples”, the largest pretraining-poisoning study as of late 2025.
- Awarded **~£4M** in research funding from government and defence partners; built and co-supervised the delivery teams (nine researchers, three engineers, eight PhD interns).

Research Associate: Privacy-Enhancing Technologies

The Alan Turing Institute | March 2020 – October 2022

- Won competitive funding to investigate **deep RL** for systems security and computer-network defence: the start of my current RL research programme (£180k combined).
- Designed **SIMple ID**, a privacy-preserving offline authentication scheme for low-cost feature phones (published at STM 2022; presented to UIDAI, Government of India).
- Advised the **NHS COVID-19 app** data-protection impact assessment (DPIA); work featured twice in WIRED.

Internships, Contracting & Advisory Roles

Various | 2012 – Present

- Research and engineering prototypes (vehicular-network authentication, mesh networking, embedded systems); expert advice to an ML/security SME (2023–24); expert advice to frontier labs (2025); advisory board member, **Black Hat Europe** (2025–26).

Selected Publications

- E. Bates, **C. Hicks**, V. Mavroudis. “**Beyond Rewards in RL for Cyber Defence**”, *ICML*, 2026.
- A. Souly, J. Rando, ... **C. Hicks**, N. Carlini, Y. Gal, R. Kirk. “**Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples**”, *NeurIPS (Spotlight)*, 2026.
- S. McFadden, M. Foley, ..., **C. Hicks**, F. Pierazzi. “**SoK: The Pitfalls of Deep Reinforcement Learning for Cybersecurity**”, *USENIX Security*, 2026.
- S. Vyas, A. Caron, **C. Hicks**, P. Burnap, V. Mavroudis. “**Beyond Training-time Poisoning: Component-level and Post-training Backdoors in Deep RL**”, *AAAI*, 2026.
- S. McFadden, M. Foley, ..., **C. Hicks**, V. Mavroudis, N. Paoletti, F. Pierazzi. “**DRMD: Deep Reinforcement learning for Malware Detection under concept drift**”, *AAAI*, 2026.
- A. Caron, V. Mavroudis, **C. Hicks**. “**On Efficient Bayesian Exploration in Model-Based Reinforcement Learning**”, *TMLR*, 2025.
- E. Shereen, D. Ristea, ..., **C. Hicks**. “**One pic is all it takes: Poisoning visual document retrieval augmented generation with a single image**”, *TMLR*, 2025.

Education

PhD Computer Science, Computer Security

University of Birmingham | 2015 – 2020

- Three first-authored, A-ranked security papers (IEEE EuroS&P ×2, IACR CHES).
- Reverse-engineered a proprietary automotive cipher (AUT64) from patents and firmware, identifying weaknesses and proposing mitigations.
- Designed two cryptographic schemes for unlinkable authenticated broadcast in vehicle-to-vehicle (V2X) using elliptic-curve properties and anonymous hardware attestation.

MEng Electronic and Software Engineering (First-Class Honours)

University of Birmingham | 2010 – 2015

- 80% overall grade, including an industrial placement year at National Instruments (UK).

Skills

AI & ML: Reinforcement Learning (RL, Deep RL, Multi-Agent RL) • Bandits & Online Learning • LLM-based Agents • Generative AI • Deep Learning • Agent-based Modelling

Security of AI: Data Poisoning • Backdoor Attacks & Defences • Model Red-Teaming • Robustness & OOD Detection • LLM Security

AI for Security: Autonomous Cyber Defence • Malware Detection • Software Vulnerability Detection • Network Defence • RL and LLM Agent Environment Design

Computer Security: Cryptography • Provable Security • Systems & Network Security • Identity & Trust • Privacy-Enhancing Technologies

Programming & Tooling: Python (PyTorch, NumPy, Stable-Baselines, RLlib/Ray, W&B, AutoGen, smolagents, OpenAI API) • Bash • C/C++ • Java • JavaCard • JavaScript • Embedded Systems

Research: Experiment Design • Reproducible Research • Peer Review (USENIX Security, PoPETs, NeurIPS, ICML, ICLR, AAAI, AAMAS) • Research Mentorship • Grant Writing