Kai Yao

School of Informatics, University of Edinburgh, UK

PhD Viva Passed (Oct 2026); Formal Degree Conferral Expected Jan 2027

kai.yao@ed.ac.uk | kaiyao17@gmail.com | +86 17302151562 | +44 7867 763608

kaikaiyao.github.io | github.com/kaikaiyao | Google Scholar | linkedin.com/in/kai-yao-ai-algorithm

Research Profile

AI researcher specializing in secure and trustworthy machine learning, generative-model provenance, post-deployment model auditing, and ML systems. My doctoral research integrates theoretical analysis, adaptive threat modelling, and large-scale empirical benchmarking to characterize when provenance and attribution mechanisms remain reliable under realistic editing and deployment changes, alongside work on privacy-induced unfairness and AI governance. Prior full-time industry engineering at Intel and Huawei and earlier scientific-imaging research at Johns Hopkins ground this work in deployable systems and rigorous measurement.

Education

University of Edinburgh — May 2023 – Jan 2027 (expected)
PhD in Cyber Security, Privacy and Trust (Advisor: Dr. Marc Juarez) — Edinburgh, UK

- Viva passed in October 2026; formal degree conferral expected January 2027. Funded by a University of Edinburgh Full PhD Scholarship.

Johns Hopkins University — Sep 2017 – May 2020
MS in Mechanical Engineering (Departmental Research Fellowship) — Baltimore, MD, USA

- Research in computer vision, quantitative microscopy, and deep learning for single-cell measurement.

Fudan University — Sep 2012 – Jun 2017
BS in Theoretical and Applied Mechanics (Outstanding Graduate of the Year) — Shanghai, China

Osaka University — 2014 – 2015
Undergraduate Exchange Student (Japanese Government JASSO Full Scholarship) — Osaka, Japan

Publications and Preprints

Note: * *denotes equal contribution / co-first authorship.*

Peer-Reviewed Conference and Workshop Publications in AI & Machine Learning

[C1] **Yao K**, Juarez M. *FARE: Forensic Acceptance Region Estimation for Catching Bait-and-Switch Image Generators*. Advances in Neural Information Processing Systems (**NeurIPS**), 2026. [arXiv] [Code]

[C2] **Yao K**. *Can Pixels Alone Reveal Image Origin? Minimax Limits and Learnable Interfaces for Passive Provenance*. Advances in Neural Information Processing Systems (**NeurIPS**), 2026. [arXiv] [Code]

[C3] **Yao K***, Szilágyi B*, Kamp S, Poór M, Szilveszter M, Zsoldos MK, Juarez M. *What Breaks Local Watermarks? A Robustness Benchmark for Local Invisible Image Watermarking*. 19th ACM Workshop on Artificial Intelligence and Security (**AISec**), 2026. [arXiv]

[C4] **Yao K**, Juarez M. *Smudged Fingerprints: A Systematic Evaluation of the Robustness of AI Image Fingerprints*. 4th IEEE Conference on Secure and Trustworthy Machine Learning (**SaTML**), 2026 (Oral). [arXiv] [Code]

[C5] **Yao K**. *What Does the Credential Still Certify? Cognitive Stewardship for AI-Mediated Education*. 9th AAAI/ACM Conference on AI, Ethics, and Society (**AIES**), 2026. [arXiv]

[C6] **Yao K**. *When AI Does the Work, What Is Learning For? Post-Instrumental Learning and the Risk of Capacity Dissolution*. 9th AAAI/ACM Conference on AI, Ethics, and Society (**AIES**), 2026. [arXiv]

[C7] **Yao K**, Juarez M. *SoK: What Makes Private Learning Unfair?* 3rd IEEE Conference on Secure and Trustworthy Machine Learning (**SaTML**), 2025 (Oral). [IEEE] [arXiv]

Preprints

[P1] **Yao K**, Juarez M. *SPRINT: Robust Model Attribution of Generated Images via Secret Pixel Reconstruction*. Preprint, arXiv:2508.05691, 2026. [arXiv]

Peer-Reviewed Journal Publications in AI for Science & Biomedical Engineering *(Publications, continued)*

[J1] Rochman ND*, **Yao K***, Perez-Gonzalez NA*, Wirtz D, Sun SX. *Single-Cell Volume Measurement Utilizing the Fluorescence Exclusion Method (FXm). Bio-protocol*, 10(12):e3652, 2020. [DOI]

[J2] **Yao K***, Rochman ND*, Sun SX. *CTRL: A Label-Free Artificial Intelligence Method for Dynamic Measurement of Single-Cell Volume. Journal of Cell Science*, 133(7):jcs245050, 2020. [Journal]

[J3] Perez-Gonzalez NA*, Rochman ND*, **Yao K***, et al. *YAP and TAZ Regulate Cell Volume. Journal of Cell Biology*, 218(10):3472–3488, 2019. [Journal]

[J4] **Yao K***, Rochman ND*, Sun SX. *Cell Type Classification and Unsupervised Morphological Phenotyping from Low-Resolution Images Using Deep Learning. Scientific Reports*, 9(1):1–13, 2019. [Journal]

[J5] Zhang Q, Meng Z, Zhang Y, **Yao K**, et al. *Phantom-Based Experimental Validation of Fast Virtual Deployment of Self-Expandable Stents for Cerebral Aneurysms. BioMedical Engineering OnLine*, 15(2):431–437, 2016. [Paper]

Research Experience

University of Edinburgh 2023 – 2026
Doctoral Research in Generative-Model Provenance, Auditing, and Fairness Edinburgh, UK

- **Post-Deployment Auditing & Passive Provenance (NeurIPS 2026; Preprint):** Developed *FARE*, an output-only audit learning patch-level acceptance regions from certified generator outputs to flag model, checkpoint, and pipeline substitutions (92.6%–99.5% mean TPR at 1% FPR; 92.4%–99.4% under white-box PGD attacks). Established minimax total-variation limits on passive provenance under editing and analyzed verifier-interface leakage enabling surrogate attacks. Designed *SPRINT*, a keyed passive-attribution method using verifier-held pixel-reconstruction targets (98.83% clean accuracy; $<$1% adaptive attack success with private keys).
- **Adversarial Evaluation of Fingerprints & Local Watermarks (IEEE SaTML 2026; ACM AISec 2026):** Formalized white-box and black-box removal and forgery threat models, benchmarking 14 fingerprinting methods across 12 GAN, VAE, and diffusion generators under 5 attack strategies; co-developed a 55-transformation benchmark decoupling spatial localization from payload recovery for local invisible image watermarks.
- **Privacy, Fairness & AI Governance (IEEE SaTML 2025; AAAI/ACM AIES 2026):** Systematized 20 studies on differential-privacy disparate impact into a four-layer taxonomy (population, data, algorithm, mechanism); formulated a capability framework and audited 30 university AI policies across four open-weight LLMs.

Johns Hopkins University 2018 – 2020
Graduate Researcher, Quantitative Biology and Deep Learning Baltimore, MD, USA

- Built label-free deep learning pipelines combining DIC microscopy preprocessing, U-Net/CNN training, and microfluidic fluorescence-exclusion validation for single-cell classification, unsupervised morphological phenotyping, and dynamic volume prediction (5–6% mean error against physical measurements).

Fudan University 2015 – 2017
Undergraduate Researcher, Biomedical Mechanics Shanghai, China

- Contributed to computational modelling and phantom-based experimental validation of fast virtual deployment of self-expandable stents for cerebral aneurysm applications.

Industry Experience

Intel Corp. Jun 2021 – Apr 2023
AI Frameworks Engineer Shanghai, China

- Led development of Neural Coder within open-source Intel Neural Compressor (40 upstream commits), combining Python static program analysis (syntax, type inference, call-graph parsing) with automated benchmarking; Hugging Face integration supported 18 PyTorch tasks ($\sim$1.7$\times$ speedup on 3rd Gen Xeon, typically $<$1% accuracy drop) and was demonstrated in Intel's *Innovation 2022* CEO keynote. [Blogs]
- Built PyTorch adapters and INT8 quantization and inference workflows for Intel Neural Compressor and Intel Extension for PyTorch (including Stable Diffusion on Intel CPU); partnered with Alibaba Cloud (PAI-DSW integration) and AWS on production deployment and cross-platform benchmarking.

Huawei Technologies Co., Ltd. Oct 2020 – Jun 2021
AI Algorithm Engineer Shanghai, China

- Developed and compressed CNN and RNN algorithms for 5G MU-MIMO wireless systems under deployment constraints, working with deployment and validation teams to integrate and verify 5G features.

Invited Talks, Presentations, and Media Coverage

- *Fingerprinting AI Models in a Black-Box Setting.* Invited keynote, Kraus & Lederer Law Firm (Munich), on model attribution and IP protection — 2026
- *Smudged Fingerprints: A Systematic Evaluation of the Robustness of AI Image Fingerprints.* Oral presentation, IEEE SaTML — 2026
- *Image Provenance in the AI Era.* Guest lecture, graduate course *Privacy and Security with Machine Learning*, University of Edinburgh [Slides] — 2026
- *SoK: What Makes Private Learning Unfair?* Oral presentation, IEEE SaTML (Copenhagen) — 2025
- *Algorithmic Unfairness in Private Classifiers.* University of Edinburgh Cyber Security, Privacy, & Trust Research Showcase — 2024
- **Media Coverage:** *Smudged Fingerprints* featured by University of Edinburgh News, DIGIT, and Herald Scotland — 2026

Teaching and Mentoring

- **Guest Lecturer**, *Privacy and Security with Machine Learning* (graduate), University of Edinburgh — 2025 – 2026
- **Teaching Assistant & Lab Demonstrator**, *Privacy and Security with Machine Learning*, University of Edinburgh — 2023 – 2025
- **Teaching Assistant**, *Mathematical and Computational Image Analysis*, Johns Hopkins University — 2019 – 2020
- **Research Mentoring:** Mentored five undergraduate and postgraduate researchers across Johns Hopkins and Edinburgh: Felipe Takaesu, Eliana Crentsil, Lucia Sablich, Shannon Flanary, and Chunhan Fang.

Academic Service

Program Committee	IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2027); ACM Workshop on Artificial Intelligence and Security (AISec 2026)
Reviewer	AAAI Conference on Artificial Intelligence (AAAI-26, AAAI-27); Privacy Enhancing Technologies Symposium (PETS/PoPETs 2027); ACM Digital Threats: Research and Practice (DTRAP)

Honors, Fellowships, and Awards

- PhD Full Scholarship, University of Edinburgh — 2023 – 2026
- IEEE SaTML Conference Attendance Grant — 2024, 2025, 2026
- Intel Division Awards: *Fast Stable Diffusion on Intel CPU* (AIA, 2023); *Neural Coder Partnering Alibaba Cloud* (CESG SW AI, 2023); *Innovation of Neural Coder* (AIA, 2022) — 2022 – 2023
- Departmental Research Fellowship, Johns Hopkins University — 2017
- Outstanding Graduate of the Year, Fudan University — 2017
- JASSO Full Scholarship for Undergraduate Exchange Students, Japanese Government — 2014

Technical Skills and Languages

Programming	Python, C++, MATLAB, Java, SQL
ML & Vision	PyTorch, TensorFlow, Hugging Face Transformers, NumPy, Pandas, CNNs, U-Net, Diffusion, GANs, VAEs, image forensics, quantitative microscopy
Security & Evaluation	Threat modelling, adversarial evaluation, provenance, fingerprinting, watermarking, differential privacy, fairness measurement, fixed-FPR statistical thresholding, ablation design
ML Systems & Tools	Intel Neural Compressor, Intel Extension for PyTorch, Neural Coder, INT8 quantization, model compression, inference optimization, benchmarking, Linux/Unix, Git, Docker, Kubernetes, LaTeX
Languages	Mandarin Chinese (native); English (full professional academic proficiency; US MS & UK PhD); Japanese (JLPT N1; one-year exchange at Osaka University)