Mateo Espinosa Zarlenga

+44-7427-436699 | mateo.espinosazarlenga@trinity.ox.ac.uk | mateoespinosa.github.io | mateoespinosa | mateoespinosa | Google Scholar

Interests: interpretable AI, AI safety, model control, representation learning, and human-in-the-loop AI.

Professional Experience

Trinity College, University of Oxford, *Junior Research Fellow (JRF) in Computer Science*, Oxford, UK — *11/2025 - Today*
Department of Computer Science, University of Oxford, *Associate Junior Research Fellow*, Oxford, UK — *12/2025 - Today*
Department of Computer Science and Technology, University of Cambridge, *Visiting Researcher*, Cambridge, UK — *10/2025 - Today*
Sony AI, *AI Ethics Research Intern*, Cambridge, UK (Remote) — *10/2023 - 01/2024*
Cerebras Systems, (Early Engineering Member) *Machine Learning Software Engineer*, Los Altos, California, USA — *07/2017 - 10/2020*
Stroz Friedberg (now an Aon Company), *Cyber Summer Associate*, Boston, Massachusetts, USA — *Summer 2016*
Dell EMC, *Software Quality Engineer*, Hopkinton, Massachusetts, USA — *Summer 2015*

Education

University of Cambridge — *Cambridge, UK*

Doctor of Philosophy in Computer Science — *Oct. 2021 - Jul. 2025*

- Supervised by Mateja Jamnik and Zohreh Shams. **Gates Cambridge Scholar**. PhD thesis was accepted with **no revisions**.

University of Cambridge — *Cambridge, UK*

Master of Philosophy in Advanced Computer Science — *Oct. 2020 - Jul. 2021*

- Graduated with a **distinction** and awarded the *St Edmund's College Prize* for my performance.

Cornell University — *Ithaca, New York, USA*

Master of Engineering in Computer Science — *Aug. 2016 - May 2017*

- **Dean's list** all semesters. **Highest GPA** of the 2017 MEng cohort (ranked 1/107).

Cornell University — *Ithaca, New York, USA*

Bachelor of Arts in Computer Science (Artificial Intelligence Concentration) — *Aug. 2013 - Dec. 2016*

- **Dean's list** all semesters. **Graduated earlier** with **distinctions in all subjects**.

Honors & Awards

2026 **ICML Travel Award**, Award covering registration and accommodation to attend ICML 2026 — *Seoul, South Korea*
Various **ICML Outstanding Reviewer Award**, Within the top 10% of reviewers for ICML 2022, 2024, 2025, and 2026 — *Various*
2025 **QuantCo Scholarship**, Award covering travel and accommodation to attend ICML 2025 — *Vancouver, Canada*
2025 **ICML Travel Award**, Award covering registration and accommodation to attend ICML 2025 — *Vancouver, Canada*
2024 **Best Paper Candidate**, Paper selected within the top 15 submissions at ECCV 2024 (out of 8,500+ papers) — *Milan, Italy*
2024 **The Alan Turing Institute Enrichment Studentship**, Award to undertake research at the ATI for six months — *London, UK*
2023 **NeurIPS Scholar Award**, Award covering registration and accommodation to attend NeurIPS 2023 — *New Orleans, US*
2022 **Wiseman Prize**, For commendable contributions to Cambridge's Computer Science Department — *Cambridge, UK*
2021 **Gates Cambridge Scholarship**, Full 4-year scholarship (tuition fees and a living stipend) to pursue my PhD — *Cambridge, UK*
2021 **St Edmund's College Prize**, Prize for final MPhil results at the University of Cambridge — *Cambridge, UK*
2017 **Outstanding Performance Award**, Highest GPA of the Computer Science MEng class at Cornell University — *Ithaca, NY, USA*
2013 **University of Excellence Scholarship**, Government-funded full scholarship to study abroad for 4 years — *Ecuador*
2012/3 **International Mathematics Olympiad**, Member of my Ecuador's core national mathematics team — *Multiple Locations*
2013 **Ecuador's National Mathematics Olympiad**, 1st place (perfect score) — *Ecuador*
2013 **Ecuador's National Physics Olympiad - Practical Modality**, 1st place — *Ecuador*

Teaching Experience

Class Instructor

Co-lecturer for "Explainable Artificial Intelligence" (L193, Masters-level Class), University of Cambridge — *2025 and 2026*
Co-lecturer for "Explainable Artificial Intelligence" MPhil Submodule (R255), University of Cambridge — *2023 - 2024*
Recitation Instructor and Head TA for "Functional Programming and Data Structures" (CS 3110), Cornell — *2017*

Teaching Assistant

College Supervisor for "Discrete Mathematics", University of Cambridge — *2021 - 2023*
Teaching Assistant for "Mathematical Foundations for the Information Age" (CS 4850), Cornell — *2016*
Teaching Assistant for "Foundations of Artificial Intelligence" (CS 4700), Cornell — *2015*
Teaching Assistant for "Discrete Structures" (CS 2800), Cornell — *2014 - 2015*

Research Supervisor

Haoyan Luo, *PhD* co-supervisor (main supervisor is M. Jamnik), University of Cambridge — *2025 - Present*
Yael Konforti, *PhD* co-supervisor (main supervisor is M. Jamnik), University of Cambridge — *2024 - Present*
Oscar Hill, *MPhil* and *PhD* co-supervisor (main supervisor is M. Jamnik), University of Cambridge — *2023 - Present*
Muhammad Ayain Fida Rana, *MPhil* thesis co-supervisor, University of Cambridge — *2025 - Present*
Jessy Sodimu, *Master* thesis co-supervisor, University of Bath — *2025 - Present*
Evan Rex, *MPhil* thesis co-supervisor, University of Cambridge (now: McKinsey) — *2023 - 2024*
Thomas Yuan, *Part III* thesis co-supervisor, University of Cambridge (now: Squarepoint Capital) — *2023 - 2024*
Naveen Raman, *MPhil* thesis supervisor, University of Cambridge (now: PhD at CMU) — *2022 - 2023*

Professional Service

LatinX in AI (LXAI) Affinity Workshop, Co-chair and part of organizing committee — *ICML 2023/24/26*
Interpretable Deep Learning Reading Group, Co-organizer — *2026 - Present*
1st Tutorial on Foundations of Interpretable Deep Learning, Co-presenter and Organizer — *AAAI 2026*
Concept-based Explainable AI Special Track, Chair and co-organizer — *XAI 2025*
1st Tutorial on Concept-based Interpretable Deep Learning, Co-presenter and Organizer — *AAAI 2025*
1st Workshop on Human-Interpretable AI, Co-general Chair and Organizer — *KDD 2024*
Conference and Journal Reviewer, Reviewer for NeurIPS, ICML, ICLR, AAAI, TMLR, and JMLR — *2022 - Present*
Workshop Program Committee, SATA @ NeurIPS, HI-AI @ KDD, LXAI @ ICML, Responsible Data @ CVPR — *2022 - 2024*
AAAI Conference on Artificial Intelligence, Volunteer — *AAAI 2023*
LatinX in AI (LXAI) Affinity Workshop, Volunteer — *Various*

Invited Talks

Concept Representations as Shared Vocabularies, ELLIS x UniReps Seminar Series, Invited Talk — *2026*
AI Theory and the Cognitive Sciences, Centre for Human-inspired AI (CHIA), Invited Panellist — *2026*
In Defense of Concept Leakage, Invited Keynote Talk, Unifying Concept Representation Learning Workshop @ ICLR — *2026*
Foundations of Interpretable Deep Learning, Tutorial, 4th Neuro-Symbolic AI Summer School, Online — *2025*
Avoiding Leakage Poisoning, Seminar, LatinX in AI (LXAI) Workshop @ ICML 2025, Vancouver, Canada — *2025*
Challenges in Safe AI Deployment, AI Seminar, Quito, Ecuador — *2025*
Fairness and ML Transparency, Guest Lecture, New York University (NYU), USA — *2025*
Bias Mitigation in the Wild, Guest Talk, Università della Svizzera italiana, Switzerland — *2024*
Learning to Receive Help, Spotlight Talk, NeurIPS @ Cambridge, UK — *2024*
Hands-on Primer on Explainable AI, Tutorial, DEMON workshop: XAI for Neuroimaging in Dementia, Online — *2023*
Invited Talk on Concept-based Interpretability, Seminar, Sidney Sussex College, Cambridge, UK — *2023*
Mind the Explainability Gap, Seminar, AI in Medicine Seminar, Addenbrooke's Hospital, UK — *2023*
Concept Embedding Models, Seminar, University of Cambridge, UK — *2023*

Extracurricular

Ignite Hubs (Educational Outreach Charity), Volunteering Python instructor for kids (weekly) *2021 - 2025*
Researcher Volunteer, Hopes and Fears Lab Popup *2023*
Cambridge Computer Laboratory Postgraduate Forum, Representative of the AI group *2021 - 2023*
St Edmund's Music Officer and Bar Staff Member, Member of the College's CR (Student Body) *2021 - 2022*
Cambridge Computer Laboratory Postgraduate Forum, Representative of MPhil cohort *2020*
St Edmund's College Film and Book clubs, President *2020 - 2022*

Skills

- **Programming:** Python, C++, C, Java, OCaml, LaTeX, Bash, Matlab, R, HTML, CSS, JavaScript, PHP
- **Machine Learning Frameworks:** Tensorflow, PyTorch, XLA/HLO, ONNX, MLIR, JAX

Selected Publications

Impact: 2,280 citations and h-index of 14 (as of September 4, 2026).

For a complete list of publications and preprints, please refer to my Google Scholar profile.

Published Peer-reviewed Works

In Defense of Concept Leakage in Concept-based Models *ICML 2026*
Position paper at the International Conference on Machine Learning (ICML). 2026.
Mateo Espinosa Zarlenga

CB-SLICE: Concept-Based Interpretable Error Slice Discovery *ICML 2026*
The International Conference on Machine Learning (ICML). 2026.
Yael Konforti, **Mateo Espinosa Zarlenga**, Elaf Almahmoud, Mateja Jamnik

Mixture of Concept Bottleneck Experts *ICML 2026 (**Spotlight**)*
The International Conference on Machine Learning (ICML). 2026.
Francesco De Santis, Gabriele Ciravegna, Giovanni De Felice, Arianna Casanova, Francesco Giannini, Michelangelo Diligenti, **Mateo Espinosa Zarlenga**, Pietro Barbiero, Johannes Schneider, Danilo Giordano

Hierarchical Concept-based Interpretable Models *ICLR 2026*
Fourteenth International Conference on Learning Representations (ICLR). 2026.
Oscar Hill, **Mateo Espinosa Zarlenga**, Mateja Jamnik

Deferring Concept Bottleneck Models: Learning to Defer Interventions to Inaccurate Experts *NeurIPS 2025*
In Advances in Neural Information Processing Systems 39. 2025.
Andrea Pugnana, Riccardo Massidda, Francesco Giannini, Pietro Barbiero, **Mateo Espinosa Zarlenga**, Roberto Pellungrini, Gabriele Dominici, Fosca Giannotti, Davide Bacciu

Avoiding Leakage Poisoning: Concept Interventions Under Distribution Shifts *ICML 2025*
The International Conference on Machine Learning (ICML). 2025.
Mateo Espinosa Zarlenga, Gabriele Dominici, Pietro Barbiero, Zohreh Shams, Mateja Jamnik

Addressing Concept Mislabeling in Concept Bottleneck Models Through Preference Optimization *ICML 2025*
The International Conference on Machine Learning (ICML). 2025.
Emiliano Penaloza, Tianyue H. Zhang, Laurent Charlin*, **Mateo Espinosa Zarlenga***

Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning? *ICLR 2025*
The International Conference on Learning Representations (ICLR). 2025.
Gabriele Dominici*, Pietro Barbiero*, **Mateo Espinosa Zarlenga**, Alberto Termine, Martin Gjoreski, Giuseppe Marra, Marc Langheinrich

Efficient Bias Mitigation Without Privileged Information [Best Paper Candidate (Top ∼0.18%)] *ECCV 2024 (**Oral**)*
In European Conference on Computer Vision. Springer, Cham, 2025.
Mateo Espinosa Zarlenga, Swami Sankaranarayanan, Jerone T. A. Andrews, Zohreh Shams, Mateja Jamnik, Alice Xiang

From Must to May: Enabling Test-Time Feature Imputation and Interventions *CEUR 2024*
In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024.
Evan Rex, **Mateo Espinosa Zarlenga**, Andrei Margeloiu, and Mateja Jamnik

Workshop on Human-Interpretable AI *KDD 2024*

In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6708-6709. 2024.

Gabriele Ciravegna, **Mateo Espinosa Zarlenga**, Pietro Barbiero, Francesco Giannini, Zohreh Shams, Damien Garreau, Mateja Jamnik, and Tania Cerquitelli

Understanding Inter-Concept Relationships in Concept-Based Models *ICML 2024*

In Proceedings of the 41st International Conference on Machine Learning, in Proceedings of Machine Learning Research 235:42009-42025.

Naveen Raman, **Mateo Espinosa Zarlenga**, Mateja Jamnik

Learning to Receive Help: Intervention-Aware Concept Embedding Models *NeurIPS 2023 (**Spotlight**)*

In Advances in Neural Information Processing Systems 36. 2023.

Mateo Espinosa Zarlenga, Katherine M Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, Mateja Jamnik

TabCBM: Concept-based Interpretable Neural Networks for Tabular Data *TMLR 2023*

In Transactions on Machine Learning Research. 2023.

Mateo Espinosa Zarlenga, Zohreh Shams, Michael Edward Nelson, Been Kim, Mateja Jamnik

Interpretable Neural-Symbolic Concept Reasoning *ICML 2023*

In International Conference on Machine Learning, pp. 1801-1825. PMLR. 2023.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, **Mateo Espinosa Zarlenga**, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, Giuseppe Marra

Human Uncertainty in Concept-Based AI Systems *AIES 2023*

In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 869-889. 2023.

Katherine M. Collins, Matthew Barker*, **Mateo Espinosa Zarlenga***, Naveen Raman*, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, Krishnamurthy Dvijotham

Towards Robust Metrics For Concept Representation Evaluation *AAAI 2023 (**Oral**)*

In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 10, pp. 11791-11799. 2023.

Mateo Espinosa Zarlenga*, Pietro Barbiero*, Zohreh Shams*, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, Mateja Jamnik

Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off *NeurIPS 2022*

In Advances in Neural Information Processing Systems 35 (2022): 21400-21413.

Mateo Espinosa Zarlenga*, Pietro Barbiero*, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, Mateja Jamnik

Unveiling the Hardware and Software Implications of Microservices in Cloud and Edge Systems *In IEEE Micro 2020*

IEEE Micro 40, no. 3 (2020): 10-19.

Yu Gan, Yanqi Zhang, Dailun Cheng, Ankitha Shetty, Priyal Rathi, Nayan Katarki, Ariana Bruno, Justin Hu, Brian Ritchken, Brendon Jackson, Kelvin Hu,Meghna Pancholi, Yuan He, Brett Clancy, Chris Colen, Fukang Wen, Catherine Leung, Siyuan Wang, Leon Zaruvinsky, **Mateo Espinosa Zarlenga**, Rick Lin, Zhongling Liu, Jake Padilla, Christina Delimitrou

An Open-Source Benchmark Suite for Microservices and Their Hardware-Software Implications for Cloud & Edge Systems *ASPLOS 2019*

In Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, pp. 3-18. 2019.

Yu Gan, Yanqi Zhang, Dailun Cheng, Ankitha Shetty, Priyal Rathi, Nayan Katarki, Ariana Bruno, Justin Hu, Brian Ritchken, Brendon Jackson, Kelvin Hu, Meghna Pancholi, Yuan He, Brett Clancy, Chris Colen, Fukang Wen, Catherine Leung, Siyuan Wang, Leon Zaruvinsky, **Mateo Espinosa Zarlenga**, Rick Lin, Zhongling Liu, Jake Padilla, Christina Delimitrou

Accepted Peer-reviewed Workshop Papers

Don't Lose Focus: Activation Steering via Key-Orthogonal Projections *MechInterp @ ICML 2026*

Workshop on Mechanistic Interpretability at ICML. 2026.

Haoyan Luo, **Mateo Espinosa Zarlenga**, Mateja Jamnik

Actionable Interpretability Must Be Defined in Terms of Symmetries: A Compositional Probabilistic Approach *CompLearning @ ICML 2026*

Workshop on Compositional Learning at ICML. 2026.

Pietro Barbiero, **Mateo Espinosa Zarlenga**, Francesco Giannini, Alberto Termine, Filippo Bonchi, Mateja Jamnik, Giuseppe Marra

Digging Deeper: Learning Multi-Level Concept Hierarchies *Principled Trustworthy AI @ ICLR 2026*

Workshop on Principled Design for Trustworthy AI at ICLR. 2026.

Oscar Hill, **Mateo Espinosa Zarlenga**, Mateja Jamnik

Do Concept Bottleneck Models Obey Locality? *XAI in Action @ NeurIPS 2023*

"XAI in Action: Past, Present, and Future Applications" Workshop at NeurIPS. 2023.

Naveen Raman, **Mateo Espinosa Zarlenga**, Juyeon Heo, Mateja Jamnik

Efficient Decompositional Rule Extraction for Deep Neural Networks *XAI4D @ NeurIPS 2021 (**Spotlight**)*

"Explainable AI Approaches for Debugging and Diagnosis" Workshop (XAI4Debugging) at NeurIPS. 2021.

Mateo Espinosa Zarlenga, Zohreh Shams, Mateja Jamnik