# Lucas Monteiro Paes

lucaspaes@g.harvard.edu | Website | linkedin

## Research Interests

I use theoretical insights to develop safe and trustworthy AI and ML systems. I am particularly interested in efficient prediction explainability, safety alignment, and methods for detecting and mitigating algorithmic bias and arbitrariness in large models.

## Education

**Harvard University** 2021-2025

Ph.D. in Applied Mathematics

2-year Apple Scholar & 1-year Behring Foundation Fellowship

Advisor: Prof. Flavio du Pin Calmon

**Harvard University** 2021-2025

M.S. Engineering Sciences

Advisor: Prof. Flavio du Pin Calmon

**Instituto de Matemática Pura e Aplicada (IMPA)** 2019-2021

M.S. in Computational Mathematics and Modeling

2-year National Council for Scientific and Technological Development Fellowship

Advisor: Prof. Roberto Imbuzeiro Oliveira

**Universidade Federal Fluminense (UFF)** 2017-2021

B.S. in Mathematics

Summa Cum Laude

1-year National Council for Scientific and Technological Development Fellowship

## Professional Experiences

**Google DeepMind** May 2024 - Nov 2024

Student Researcher, Gemini Safety team New York - NY

- Developed sampling strategies for Large Language Models (LLM) to improve generation quality
- Formulated Reinforcement Learning strategies to boost LLM sample efficiency and improve text generation

**IBM T.J. Watson Research Center** May 2023 - Aug 2023

AI Research Scientist Intern Yorktown Heights - NY

- Developed a fast explanation method tailored for Generative Language Models
- Solution patented and being deployed by IBM

**Center for Projects and Innovation at IMPA** Jan 2021 - Jul 2021

Research Assistant Rio de Janeiro - Brazil

- Developed AminioML a model with uncertainty quantification for interpretable amniotic fluid estimates from MRI
- Solution deployed by Dasa, the biggest diagnosis company in Latin America

**Energy Research Office** Jan 2019 - Aug 2019

Energy Research Intern, Mathematical Modeling Group Rio de Janeiro - Brazil

- Optimized oil production unit auctions using machine learning and operations research techniques
- Solution deployed by the Brazilian Ministry of Mines and Energy

**National Council for Scientific and Technological Development** Aug 2018 - Aug 2019

Research Assistant Rio de Janeiro - Brazil

- Improved human voices simulation by creating a novel stochastic mathematical model for producing jitter
- Awarded the Beatriz Neves Award for the best undergraduate work in applied mathematics in Brazil

## Selected Awards

**Best Paper Award** — New England NLP | 2025
- Best paper (1 out of 90 accepted papers) awarded to the work "AI Alignment at your Discretion."

**Apple Scholar in AI/ML Fellowship** — Apple | 2024
- Awarded to 21 senior Ph.D. candidates globally
- Students selected based on innovative research and record as thought leaders and collaborators

**Innovative Applications of Artificial Intelligence Award** — AAAI & IAAI | 2023
- Awarded to "AmnioML: Amniotic fluid segmentation and volume prediction with uncertainty quantification"
- Honors papers that describe deployed applications of AI with measurable benefits

**ISIT Student Travel Grant** — ISIT | 2023
- Awarded to selected authors of the 2023 IEEE International Symposium on Information Theory

**Leadership Fellow** — Fundação Estudar | 2022
- Awarded to 30 Brazilian students selected out of 33k (0.08%)
- Fellowship aims to "develop Brazil's most promising young leaders" (featured at Forbes)

**Behring Foundation Fellowship** — Behring Foundation | 2022
- Awarded to students who "demonstrate potential, resilience, and have high aspirations"

**NeurIPS Scholar Award** — NeurIPS | 2022
- Awarded to selected authors of the 2022 NeurIPS

**Honorable Mention** — Universidade Federal Fluminense | 2021 & 2019
- Awarded to 19 out of 50k students
- For students with the greatest GPAs for two years in a row

**Beatriz Neves Award** — Brazilian Society of Computational and Applied Mathematics | 2019
- Awarded to the paper "A novel source-filter stochastic model for voice production"
- For the best undergraduate research in Brazil in the field of Computational or Applied Mathematics

**Vasconcellos Torres Award** — Universidade Federal Fluminense | 2019
- For the best undergraduate research at UFF in the field of Engineering.

## Leadership

**Artificial Intelligence in Brazil Panel Series** — 2025
- Organized the panel series at Harvard University discussing AI advances and its policy implications in Brazil.
- Speakers included researchers, CTOs/CEOs, and policymakers.

**Science and Technology VP of the Brazil Conference** — 2023
- The Brazil Conference is the most prominent event focused on Brazil with more than 1000 annually attendees
- Led a team with more than 20 members creating science and technology content focusing on impact in Brazil
- Speakers included CTOs, Big Tech directors, Minister of State, and Unicorn Founders

**Symposium on Machine Learning and Information Theory** — 2022
- Organized a symposium at the Congresso Nacional de Matemática Aplicada e Computacional with Flavio Calmon

**Tiny Machine Learning Minicourse** — 2022
- Organized a minicourse on Tiny machine learning with Vijay Janapa, Brian Plancher, Marcelo Rovai, Jose Filho

**EdiMat** — 2019
- Idealized the first event for and by mathematics undergraduate students in Rio de Janeiro
- Led a team of over 10 volunteers for the event and raised funds
- EdiMat had over 100 participants, 16 lectures, and 3 panel discussions

## Selected Presentations

**Apple - Machine Learning Research Team (Invited Talk)** — 2024 & 2025

**FAccT (Talk)** — 2024

**MIT-IBM Watson AI Lab (Invited Talk)** — 2023

**Laboratory of Big Data and Predictive Analytics in Healthcare at USP (Invited Talk)** — 2023

**Fundação Estudar Annual Meeting in New York (Keynote)** — 2022

**XXXIX CNMAC (Special Invited Talk)** 2019

- CNMAC is the biggest applied mathematics congress in Latin America.

## Reviewer Services

**Conferences:** NeurIPS (22, 23, 24, 25), FAccT (24, 25), AAAI (24), AISTATS (24), ICML (23)
**Journals:** IEEE JSAIT, ACM JCSS, TMLR

## In Press

**Jota Newspaper** 2024

- Article about the potential consequences of using AI to score Brazilian university entrance exams

**Harvard News** 2024

- Article about my research was published at Harvard SEAS News

**FAPERJ News** 2022

- The news portal of the Research Support Foundation of Rio de Janeiro created a video about AminioML

**Folha de São Paulo** 2021

- Our ML tool to predict amniotic fluid estimations (AminioML) was featured at the Folha de São Paulo newspaper

## Patents Filed

**Perturb and Explain: In-context Attribution Pipeline for Large Language Models** 2024

Authors: L. Monteiro Paes, D. Wei, R. Luss, A. Dhurandhar, M. Nagireddy, K. Ramamurthy, P. Sattigeri, I. Padhi

U.S. patent application no. 18/595630, March 2024.

## Publications

(* indicates equal contribution)

1. M. Buyl*, H. Khalaf*, C. Mayrink Verdun*, **L. Monteiro Paes***, C. Machado, and F. Calmon. Ai alignment at your discretion. ACM Conference on Fairness, Accountability, and Transparency (**FAccT**) - Best Paper Award at the New England NLP Workshohp, 2025.

2. **L. Monteiro Paes***, D. Wei*, H. Jin Do, H. Strobelt, R. Luss, A. Dhurandhar, M. Nagireddy, K. Natesan Ramamurthy, P. Sattigeri, W. Geyer, and S. Ghosh. Multi-level explanations for generative language models. Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics (**ACL**) - Oral at ACL, 2025.

3. **L. Monteiro Paes**, D. Wei, and F. Calmon. Selective explanations. Conference on Neural Information Processing Systems (**NeurIPS**), 2024.

4. A. Oesterling, C. Mayrink Verdun, C. Xuan Long, A. Glynn, **L. Monteiro Paes**, S. Vithana, M. Cardone, and F. Calmon. Multi-group proportional representation. Conference on Neural Information Processing Systems (**NeurIPS**), 2024.

5. F. Gomez*, C. Machado*, **L. Monteiro Paes***, and F. Calmon. Algorithmic arbitrariness in content moderation. ACM Conference on Fairness, Accountability, and Transparency (**FAccT**), 2024.

6. C. Verdun, C. Tsuzuki, C. Machado, F. Gomez, **L. Monteiro Paes**, and F. Calmon. Ai technologies: Algorithmic monocultures, arbitrariness, and global divides. T20 Inclusive Digital Transformation - **G20 Policy Brief**, 2024.

7. **L. Monteiro Paes**, A. T. Suresh, A. Beutel, F. Calmon, and A. Beirami. Multi-group fairness evaluation via conditional value-at-risk testing. IEEE Journal on Selected Areas in Information Theory (**JSAIT**), 2024.

8. **L. Monteiro Paes**, R. Cruz, F. Calmon, and M. Diaz. On the inevitability of the rashomon effect. IEEE International Symposium on Information Theory (**ISIT**), 2023.

9. D. Csillag, **L. Monteiro Paes**, T. Ramos, J. V. Romano, R. Oliveira, R. Seixas, and P. Orenstein. Amnioml: Amniotic fluid segmentation and volume prediction with uncertainty quantification. Conference on Innovative Applications of Artificial Intelligence (**IAAI | AAAI**) - Innovative Application of AI Award, 2023.

10. A. Lin*, **L. Monteiro Paes***, S. Tanneru*, S. Srinivas, and H. Lakkaraju. Word-level explanations for analyzing bias in text-to-image models. Workshop on Deployment Challenges for Generative AI, **ICML**, 2023.

11. **L. Monteiro Paes***, C. Long*, B. Ustun, and F. Calmon. On the epistemic limits of personalized prediction. Conference on Neural Information Processing Systems (**NeurIPS**), 2022.

12. E. Cataldo, **L. Monteiro Paes**, and C. Soize. A novel source-filter stochastic model for voice production. **Journal of Voice** - Beatriz Neves Award, 2021.