

Lucas MONTEIRO PAES

lucasmp@alumni.harvard.edu | [Website](#) | [linkedin](#)

RESEARCH INTERESTS

I use theoretical insights to develop safe and trustworthy AI and ML systems. I am particularly interested in language model alignment, explainability, and methods for detecting and mitigating algorithmic bias.

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

APPLE

July 2025 - present

AI RESIDENT (PostDoc), APPLE MACHINE LEARNING RESEARCH

Cupertino - CA

- TRUSTWORTHY AI AND SAFETY RESEARCH.

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 RELEASED BY IBM: [GITHUB.COM/IBM/ICX360](https://github.com/IBM/ICX360)

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 USING STOCHASTIC MODELING – AWARDED BEST UNDERGRADUATE WORK IN MATH 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

ACADEMIC ACHIEVEMENT AWARD

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

NAVIGATING MODEL UNCERTAINTY AND THE RASHOMON EFFECT WORKSHOP @ AAAI

2025

- ORGANIZER OF THE FIRST WORKSHOP ON THE RASHOMON EFFECT.

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 New England NLP**, 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 63rd Annual Meeting of the Association for Computational Linguistics (**ACL**) - **Oral at ACL**, 2025.
3. D. Wei, R. Luss, X. Hu, **L. Monteiro Paes**, P. Chen, K. Natesan Ramamurthy, E. Miehling, I. Vejsbjerg, and H. Strobelt. Icx360: In-context explainability 360 toolkit, 2025.
4. **L. Monteiro Paes**, D. Wei, and F. Calmon. Selective explanations. Conference on Neural Information Processing Systems (**NeurIPS**), 2024.
5. 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.
6. F. Gomez*, C. Machado*, **L. Monteiro Paes***, and F. Calmon. Algorithmic arbitrariness in content moderation. ACM Conference on Fairness, Accountability, and Transparency (**FAccT**), 2024.
7. 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.
8. **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.
9. **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.

10. 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.
11. 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.
12. **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.
13. E. Cataldo, **L. Monteiro Paes**, and C. Soize. A novel source-filter stochastic model for voice production. **Journal of Voice - Beatrix Neves Award**, 2021.