Andrew Estornell

Research Scientist · Computer Science

ByteDance Research

andrew.estornell at bytedance.com | andrewestornell.github.io

Research Interests

My research interests fall broadly within the field of multi-agent systems, LLM-agents, and responsible AI. Recently, my work has centered around building and training teams of LLM-agents to collaboratively solve complex problems. In addition to this line of work, I am also interested in responsible AI from both a safety and fairness perspective.

Education

Washington University in Saint Louis *Saint Louis Missouri*

PhD in Computer Science *Aug 2018 - June-2023*

- Advisor: Yevgeniy Vorobeychik
- Advisor: Sanmay Das

Temple University *Philadelphia PA*

BS in Mathematics, Minor in Computer Science *Aug 2015 - May 2018*

- Advisor: Chelsea Walton

Experience

ByteDance Research *San Jose, CA*

Research Scientist *July 2023 - Present*

- Working on training teams of LLM-agents to collaboratively solve complex tasks, as well as improving the trustworthiness of LLMs.

Washington University in St Louis - Dept of Computer Science *St Louis, MO*

PhD Candidate *Aug 2018 - June 2023*

- PhD research: Investigated the safety and fairness of automated decision making systems in the presence of strategic agents.

Temple University - Dept of Computer Science *Philadelphia, PA*

Research Assistant *Dec 2017- Jun 2018*

- Undergraduate research: Collaboration with the Temple University Hospital to develop a GNN-based diagnosis tool for neurological disorders based on patient EEG data.

Temple University - Dept of Mathematics *Philadelphia, PA*

Research Assistant *May 2017 - Jun 2018*

- Undergraduate research: Derived tractable conditions to compute Poincaré-Birkhoff-Witt-Deformations of Quadratic Monomial Algebras and used these conditions to prove that every such Algebra yields at least one non-trivial deformation.

Publications

† equal contribution

[1] ACC-Collab: An Actor Critic Approach to Multi-Agent LLM Collaboration.
Andrew Estornell†, Jean-Francois Ton†, Yuanshun Yao, Yang Liu. International Conference on Learning Representations (ICLR) 2025

[2] Multi-LLM Debate: Framework, Principals, and Interventions.
Andrew Estornell, Yang Liu. Conference on Neural Information Processing Systems (NeurIPS) 2024.

[3] Measuring and Reducing LLM Hallucination Without Gold-Standard Answers via Expertise-Weighting.
Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, Yang Liu. Preprint 2024

[4] User-Creator Feature Polarization in Recommender Systems with Dual Influence.
Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, Yang Liu. Conference on Neural Information Processing Systems (NeurIPS) 2024.

[5] Which Features are the Fairest of them All? The Impact of Features Used by Algorithms on Perceptions of Fairness.
Andrew Estornell†, Tina Zhang†, Sanmay Das, Chien-Ju Ho, Brendan Juba, Yevgeniy Vorobeychik. International Joint Conference on Artificial Intellegence (IJCAI) 2024

[6] Adaptive Recruitment Resource Allocation to Improve Cohort Representativeness in Participatory Biomedical Datasets.
Victor Borza, Andrew Estornell, Ellen Wright Clayton, Chien-Ju Ho, Russell Rothman, Yevgeniy Vorobeychik, Bradley Malin. American Medical Informatics Association (AMIA) 2024 (**awarded best student paper**).

[7] Dataset Representativeness and Downstream Task Fairness. Victor Borza†, Andrew Estornell†, Chien-Ju Ho, Bradley Malin, Yevgeniy Vorobeychik. Preprint 2024

[8] Incentivizing Recourse through Auditing in Strategic Classification.
Andrew Estornell, Sanmay Das, Yang Liu, Yatong Chen, Yevgeniy Vorobeychik. International Joint Conference on Artificial Intellegence (IJCAI) 2023.

[9] Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents.
Andrew Estornell, Sanmay Das, Yang Liu, Yevgeniy Vorobeychik. Coonference on Fairness Accountability and Transparency (FAccT) 2023. Also appeared in Learning with Strategic Agents (LSA) 2022 (**awarded best paper**)

[10] Popularizing Fairness: Group Fairness and Individual Welfare.
Andrew Estornell, Sanmay Das, Brendan Juba, Yevgeniy Vorobeychik. Conference on Artificial Intelligence (AAAI) 2023.

[11] Location Spoofing Attacks on Autonomous Fleets.
Jinghan Yang, Andrew Estornell, Yevgeniy Vorobeychik. Conference on Vehicle Security and Privacy (VehicleSec) 2023

[12] Predicting Customer Goals in Financial Institution Services: A Data-Driven LSTM Approach.
Andrew Estornell†, Stylianos Loukas Vasileiou†, William Yeoh, Daniel Borrajo, Rui Silva. ICAPS Workshop on Financial Planning (FinPlan) 2023.

[13] Manipulating Elections by Changing Voter Perceptions.
Junlin Wu, Andrew Estornell, Lecheng Kong, Yevgeniy Vorobeychik. International Joint Conference on Artificial Intelligence (IJCAI) 2022

[14] Incentivizing Truthfulness Through Audits in Strategic Classification.
Andrew Estornell, Sanmay Das, Yevgeniy Vorobeychik. Conference on Artificial Intelligence (AAAI) 2021.

[15] Election Control by Manipulating Issue Significance.
Andrew Estornell, Sanmay Das, Edith Elkind, Yevgeniy Vorobeychik. Conference on Uncertainty in Artificial Intelligence (UAI) 2020.

[16] Deception Through Half-Truths.
Andrew Estornell, Sanmay Das, Yevgeniy Vorobeychik. Conference on Artificial Intelligence (AAAI) 2020.

[17] PBW Deformations of Quadratic Monomial Algebras.
Andrew Estornell, Zachary Cline, Chelsea Walton, Matthew Wynne. Communications in Algebra 2019.

Awards

Best paper award at (LSA) "Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents" 2022

Francis James Sholomskas Scholarship for Outstanding Students (Mathematics) 2017-2018

Contributed Presentations and Talks

To Present: "ACC-Collab: An Actor-Critic Approach to Multi-Agent LLM Collaboration." at ICLR 2025

Presented "Multi-LLM Debate: Framework, Principals, and Interventions." at NeurIPS 2024

Presented "Which Features are the Fairest of them All? The Impact of Features Used by Algorithms on Perceptions of Fairness." at IJCAI 2024

Presented "Incentivizing Recourse through Auditing in Strategic Classification." at IJCAI 2023

Presented "Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents." at FAccT 2023

Presented "Popularizing Fairness: Group Fairness and Individual Welfare" at AAAI 2023

Presented “Manipulating Elections by Changing Voter Perceptions” at IJCAI 2022

Presented “Unfairness Despite Awareness: Group-Fair Classification with Strategic Agents” at LSA (AAMAS workshop) 2022, and at StratML (NeurIPS workshop) 2021.

Presented “Incentivizing Truthfulness Through Audits in Strategic Classification” at AAAI, 2021.

Presented “Election Control by Manipulating Issue Significance” at UAI 2020.

Presented “Deception Through Half-Truths” at AAAI 2020.

Professional Development and Skills

Teaching Experience

Spring 2022	**Special Topics in Computer Science: Adversarial Machine Learning (CSE.544T)**, Teaching Assistant	*Washington University in Saint Louis*
Spring 2023	**Adversarial AI (CSE.555T)**, Guest Lecturer	*Washington University in Saint Louis*

Program Committee Member and Reviewer for Conferences and Journals

NeurIPS: 2024, 2022, 2021; ICLR: 2025, 2024; AAMAS: 2025, 2023, 2022; AAAI: 2024, 2023, 2022, 2021; AIES: 2023; IJCAI: 2023; JAIR: 2023; ICML: 2023, 2022; FAccT: 2023, 2022; KAIS: 2022; LSA: 2022; AASG: 2020; UAI: 2020

Programming Languages

Languages: Python, Mathematica,