Research Statement

Carlos Mougan

Contact Information

- 96 Euston Rd., London NW1 2DB, Reino Unido
- Email: carmougan@gmail.com
- +34 620 618 275

Research Focus

My research is centred on the field of AI observability and Model Monitoring within the domain of Artificial Intelligence (AI). Model monitoring involves the critical practice of observing and understanding the behavior of AI systems in production. It encompasses data collection and analysis from deployed AI models to ensure their proper functioning and to detect any issues or anomalies during their operation.

I am particularly interested in addressing the challenges of AI Observability, a multifaceted area that involves two primary dimensions: AI Alignment and Model Monitoring. Within the context of AI Alignment, my research aims to ensure that AI systems align with a diverse set of human values, including moral, societal, and justice-oriented principles. To achieve this, I am developing innovative metrics and approaches that go beyond traditional fairness metrics. These novel metrics consider feature attribution explanations and distributions to provide a more comprehensive assessment of AI alignment [1]

In the realm of Model Monitoring, my research explores the intricate relationships between distribution shifts and learned models [2]. I propose the utilization of "explanation shifts" as a pivotal indicator for understanding how changes in data distributions impact model behavior. This research advances conventional methods by incorporating feature attribution explanations, such as Shapley values and LIME, to provide more sensitive and explainable insights into model performance and behavior.

Furthermore, I investigate the concept of explainable uncertainty estimation to identify the sources of model performance deterioration. This research has significant implications for enhancing model accountability, addressing potential biases, and ensuring responsible AI deployment.

Contributions

My contributions extend beyond theoretical advancements to practical tools and methodologies. I have developed open-source tools and packages, such as `explanationspace`, `skshift`, `category_encoders` to make these innovative methods accessible and applicable to the broader research community.

In summary, my research seeks to advance the field of AI Observability by providing innovative solutions to AI Alignment and Model Monitoring challenges. I am dedicated to contributing to the responsible and ethical use of AI systems, aligning with the evolving regulatory landscape, such as the European Commission's AI Act.

My research is a reflection of my commitment to advancing the field of AI while aligning with the department's commitment to excellence in research, innovation, and ethical AI development. I am enthusiastic about the prospect of collaborating with colleagues and students who share a passion for ensuring the responsible and fair application of AI in our increasingly complex and interconnected world.

References

[1] Carlos Mougan, Laura State, Antonio Ferrara, Salvatore Ruggieri, and Steffen Staab. Beyond demographic parity: Redefining equal treatment, 2023.

[2] Carlos Mougan, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[3] Carlos Mougan, Richard Plant, Clare Teng, Marya Bazzi, Alvaro Cabregas Ejea, Ryan Sze-Yin Chan, David Salvador Jasin, Martin Stoffel, Kirstie Jane Whitaker, and Jules Manser. How to data in datathons. *Advances in Neural Information Processing Systems*, 2023.

[4] Carlos Mougan, David Masip, Jordi Nin, and Oriol Pujol. Quantile encoder: Tackling high cardinality categorical features in regression problems. In *Modeling Decisions for Artificial Intelligence*, pages 168–180, Cham, 2021. Springer International Publishing.

[5] Carlos Mougan and Dan Saattrup Nielsen. Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI Conference on Artificial Intelligence*, 2023.

[6] Carlos Mougan, Georgios Kanellos, and Thomas Gottron. Desiderata for explainable AI in statistical production systems of the european central bank. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, volume 1524 of *Communications in Computer and Information Science*, pages 575–590. Springer, 2021.

[7] Carlos Mougan Mougan, George Kanellos, Johannes Micheler, Jose Martinez, and Thomas Gottron. Introducing explainable supervised machine learning into interactive feedback loops for statistical production system. In *Irving Fisher Committee on Central Bank Statistics,2021*, 2021.

[8] Francisco Castillo-Eslava, Carlos Mougan, Alejandro Romero-Reche, and Steffen Staab. The role of large language models in the recognition of territorial sovereignty: An analysis of the construction of legitimacy. In *European Workshop of Algorithmic Fairness*, 2023.

[9] Carlos Mougan, Jose Alvarez, Salvatore Ruggieri, and Steffen Staab. Fairness implications of encoding protected categorical attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 956–966, Montreal, Canada, 2023. Association for Computing Machinery.