

Carlos Mougan

AI Safety | AI Regulation | Machine Learning

cmougan.eu | carmougan@gmail.com | +34 620 618 275 | [LinkedIn](#)

Profile

AI safety and technical policy specialist at the European Commission AI Office, working on the implementation and enforcement preparation of the EU AI Act for general-purpose AI. Current work focuses on systemic risks, particularly harmful manipulation and child safety. Previously a Marie Skłodowska-Curie Research Fellow and Principal Investigator at The Alan Turing Institute, with research spanning model monitoring, uncertainty estimation, tech-policy and data quality. First author publications include work in *Science*, NeurIPS, AAAI, AIES and TMLR.

Experience

European Commission — AI Office

AI Safety

Brussels, Belgium

Oct 2024 – Present

- Contribute to the technical implementation and enforcement preparation of the AI Act for general-purpose AI, including systemic-risk obligations [1].
- Lead work on harmful manipulation and child safety risks, including prohibited practices, CSAM/NCII, addiction and over-reliance, and mental health; develop evaluations and coordinate external research collaborations.
- Coordinate AI Act enforcement work with the Digital Services Act and contribute to EU policy development and legislative discussions on emerging AI risks.

The Alan Turing Institute

Principal Investigator, Applied Skills Team

London, United Kingdom

Apr 2022 – Jan 2025

- Conducted research on AI alignment, AI policy and safety, and data quality, including work published at NeurIPS 2023 Datasets & Benchmarks [2].
- Provided technical and research advisory support to public, nonprofit and industry partners, including the British Antarctic Survey, WWF, Amnesty International, Roche, AstraZeneca, CEFAS, Telenor and Turkcell.

University of Southampton

Marie Skłodowska-Curie Research Fellow & PhD Researcher

Southampton, United Kingdom

Feb 2021 – Feb 2024

- Conducted research on model monitoring, explainability and fairness within the EU-funded NoBias Innovative Training Network.
- Published at NeurIPS, AAAI, AIES and ECML PKDD [2, 3, 4, 5, 6, 7]; completed research visits at SCHUFA, the University of Pisa and BBC DataLab.
- Taught Data Science and Machine Learning to industry professionals, consultants and startups.

European Central Bank

Statistician, Directorate General Statistics

Frankfurt, Germany

Jan 2020 – Jan 2021

- Monitored the quality of balance-of-payments and international-investment-position statistics, including methods and sources across EU Member States.
- Developed a multiclass neural-network model for National Central Banks and the European System of Central Banks.

Deloitte Robotics

Technology Consultant

Madrid, Spain

Dec 2017 – Aug 2018

- Applied OCR and robotic process automation to audit and banking workflows; worked as a UiPath developer and trained Deloitte teams in Belgium and the Netherlands.

Education

PhD, Model Monitoring and AI Ethics — University of Southampton

2021 – 2024

Thesis: Model Monitoring in the Absence of Ground Truth Data via Feature Attribution Explanations; supervisor: Steffen Staab

MSc, Mathematical Modelling — Universitat Autònoma de Barcelona

2018 – 2019

BSc, Physics — Universidad Complutense de Madrid

2013 – 2018

Other & Highlights

- **Open source:** Main contributor to [Category Encoders](#); main developer of [skshift](#).
- **Data Science Stack Exchange:** Ranked among the top two users in 2020.

- **Machine-learning competitions:** 1st — BCG Gamma Datathon; 1st — Novartis 2021 Datathon; Bronze Medal — Kaggle IEEE-CIS Fraud Detection.
- **Sport:** Elite swimming athlete, 2013–2017.

Selected Talks

- **Royal Society for Spanish Mathematics** — *Beyond Demographic Parity: Redefining Equal Treatment* (Jan 2024).
- **NeurIPS 2023** — *How To Data in Datathons* (Dec 2023).
- **Regulatable ML Workshop, NeurIPS 2023** — *Necessity of Processing Sensitive Data* (Dec 2023).
- **AAAI 2023** — *Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-Parametric Bootstrap* (Jan 2023).

Selected Publications

- [1] Carlos Mougan, Lauritz Morlock, Jair Aguirre, James R. M. Black, Jan Brauner, Simeon Campos, Sunishchal Dev, David Fernández Llorca, Alberto Franzin, Mario Fritz, Emilia Gómez, Friederike Grosse-Holz, Eloise Hamilton, Max Hasin, Jose Hernandez-Orallo, Dan Lahav, Luca Massarelli, Vasilios Mavroudis, Malcolm Murray, Patricia Paskov, Jaime Raldua, and Wout Schellaert. The science and practice of proportionality in ai risk evaluations. *Science*, 391(6787):769–771, 2026.
- [2] 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.
- [3] Carlos Mougan and Dan Saattrup Nielsen. Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI Conference on Artificial Intelligence*, 2023.
- [4] 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.
- [5] Carlos Mougan, Laura State, Antonio Ferrara, Salvatore Ruggieri, and Steffen Staab. Beyond demographic parity: Redefining equal treatment. In *First Workshop on AI meets Moral Philosophy and Moral Psychology. Neural Information Processing Systems*, 2023.
- [6] 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.
- [7] 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.