DAN LEY

PhD Student

DETAILS

ADDRESS

96 Winthrop St
Boston, MA 02119
United States

PHONE

+1 857 313 5096

EMAIL

d.w.ley@hotmail.com

LINKS

Personal Website

LinkedIn

Google Scholar

GitHub

Twitter

SKILLS

Explainable AI

Python & PyTorch

LaTeX & Paper Writing

ChatGPT & Copilot

EDUCATION

PhD Computer Science, Harvard University — Cambridge, MA

Sep 2022 — May 2028

Trustworthy machine learning, supervised by Himabindu Lakkaraju

Conference paper *On Minimizing the Impact of Dataset Shifts on Actionable Explanations* [3] accepted to *UAI 2023 (Oral)*

Workshop paper *Consistent Explanations in the Face of Model Indeterminacy via Ensembling* [9] accepted to *ICML 2023*

M.Eng Engineering, University of Cambridge — Cambridge, UK

Oct 2017 — Jul 2021

Explaining uncertainty in deep learning, supervised by Adrian Weller

Research award for outstanding project (top 5% of students)

Workshop papers *d-CLUE: Diverse Sets of Explanations for Uncertainty Estimates* [6] and *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates* [7] accepted to *ICLR/ICML 2021*

1st Year: Class I - 87% (12th of 324); **2nd Year:** Class I - 83% (12th of 310)

3rd Year: Pass (No Classing / COVID); **4th Year:** Distinction

Coursework: Probabilistic ML, Practical Optimization, Computational Statistics, Data Compression, Bayesian Inference

EMPLOYMENT HISTORY

Research Assistant, Harvard University — Cambridge, MA

Jul 2023 — Present

Conducting PhD research at the intersection of explainable AI systems and Large Language Models (LLMs). Investigating the use of LLMs as explainers of other AI systems [10], and the faithfulness of chain-of-thought reasoning in LLMs [11].

Workshop paper *Are Large Language Models Post Hoc Explainers?* [10] accepted to *NeurIPS 2023*

Workshop paper *On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models* [11] accepted to *NeurIPS 2024*

LANGUAGES

English

French

Spanish

AI Researcher, JPMorgan Chase & Co — London, UK

Oct 2021 — Jul 2022

Explainable AI, supervised by Saumitra Mishra and Daniele Magazzeni

Methods to outperform state-of-the-art and cut computational costs by orders of magnitudes for global explanations of AI models

Workshop paper *Global Counterfactual Explanations: Investigations, implementations and improvements* [8] accepted to *ICLR 2022*

Conference paper *GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations* [2] accepted to *ICML 2023*

Research Assistant, University of Cambridge — Cambridge, UK

Jul 2021 — Sep 2021

Continuation of MEng research to explain uncertainty in deep learning; explored the notion of a distribution over counterfactual explanations

Conference Paper *Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates* [1] accepted to *AAAI 2022*

CONFERENCE PUBLICATIONS

[1] Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates — AAAI 2022

Dan Ley*, Umang Bhatt, Adrian Weller

[2] GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations — ICML 2023

Dan Ley*, Saumitra Mishra, Daniele Magazzeni

[3] On Minimizing the Impact of Dataset Shifts on Actionable Explanations — UAI 2023 (Oral)

Anna P. Meyer*, **Dan Ley***, Suraj Srinivas, Himabindu Lakkaraju

[4] Degraded Polygons Raise Fundamental Questions of Neural Network Perception — NeurIPS Datasets & Benchmarks 2023

Leonard Tang, **Dan Ley**

[5] OpenXAI: Towards a Transparent Evaluation of Model Explanations — NeurIPS Datasets & Benchmarks 2022 (Revised '24)

Chirag Agarwal*, Dan Ley, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, Himabindu Lakkaraju

WORKSHOP PUBLICATIONS

[6] d-CLUE: Diverse Sets of Explanations for Uncertainty Estimates — ICLR 2021

Dan Ley*, Umang Bhatt, Adrian Weller

[7] Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates — ICML 2021

Dan Ley*, Umang Bhatt, Adrian Weller

[8] Global Counterfactual Explanations: Investigations, Implementations and Improvements — ICLR 2022

Dan Ley*, Saumitra Mishra, Daniele Magazzeni

[9] Consistent Explanations in the Face of Model Indeterminacy via Ensembling — ICML 2023

Dan Ley, Leonard Tang, Matthew Nazari, Hongjin Lin, Suraj Srinivas, Himabindu Lakkaraju

[10] Are Large Language Models Post Hoc Explainers? — NeurIPS 2023

Nicholas Kroeger*, **Dan Ley***, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju

[11] On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models — NeurIPS 2024

Sree Harsha Tanneru, **Dan Ley***, Chirag Agarwal, Himabindu Lakkaraju

ADDITIONAL

Honours

Scholar of Corpus Christi College, University of Cambridge (2021)

Prize for Outstanding Research Project - Top 5% of Students (2021)

Travel Award for ICLR Workshop Security & Safety in ML Systems (2021)

Dewhurst Scholar for Outstanding Exam Performance (2018-2021)

Mathematics Background

90% average in 1st-3rd Year Mathematics - Highest Modules (2017-20)

Senior Team Mathematics Challenge National Finalists (2016 & 2017)

British Mathematical Olympiad, Top 500 Students in the UK (2016)

50,000 interactions on Brilliant.org mathematics problems/solutions

Ranked 1st of 220,000 users on JobFlare (cognitive speed tests)

Sporting Achievement

Footballer for MIT FC, Bay State Soccer League Div 1 (2022-2024)

Coach for Corpus Christi FC, University of Cambridge (2021-2022)

Marathon and Double Marathon Runner (2020 & 2021)

Footballer for Cambridge University Blues (2017-2021)