# PhD Progress Update

James Hinns

April 2024

## 1 Overview

In this report, I aim to provide a clear view of my objectives for this PhD and the progress I have made towards achieving them.

This report is divided into three principal sections:

1. **Overview**: This section provides a brief introduction to my current progress in both research and the PhD Study Programme, along with the overarching aims of the PhD.
2. **Research**: Discusses both current and projected research activities as part of this PhD programme.
3. **PhD Study Programme**: Details both progress to date and future objectives aligned with the study requirements of the programme.

### Aim

This PhD aims to enhance the understanding and transparency of machine learning predictions through the introduction of novel methods and evaluations of explainable AI (XAI). With a technical focus, I aim to concurrently develop usable, open-source tools that complement my research.

### Current projects

The contributions are categorised as follows: Papers that have been submitted are highlighted in **bold**, and those that are ongoing are in *italics*:

- ***Novel Explanation Methods***:
  - **XAIStories**: Natural language narratives formed from SHAP and counterfactual explanations.
  - **CoF Tables**: A method to aggregate instance-level explanations to identify shortcuts in image classification models.
- ***Evaluation of XAI Methods***:
  - *PCherry*: Evaluates the probablilty that a given set of counterfactuals includes cherry-picked explanations.
- *AXA Collaboration*: Improving mail switching LLM through the analysis of counterfactual explanations.

### Progress towards Study Program

The program requires the accumulation of 30 credits. Below is the summary of credits earned to date:

| Competency | Name | Credits | Date |
|---|---|---|---|
| A | DLCV | 6 | Jun 2023 |
| C | Masters Supervision | 1 | Jan 2023 |
| C | Masters Supervision | 1 | Jan 2024 |
| D | XAIStories arXiv | 2 | Sep 2023 |
| E | ECML Poster | 2 | Sep 2023 |
| | | **Total Credits** | 12 |

# 2 Research

This section overviews the research I am undertaking as part of this PhD. It is split into two subsections:

1. **Current**: This covers the projects I am currently engaged in. Projects that have had papers submitted are highlighted in **bold**, while those still being actively worked on are in *italics*.
2. **Future**: This outlines the research I plan to undertake in the near future.

## Current

### XAIStories

### Abstract

In today’s critical domains, the predominance of black-box machine learning models amplifies the demand for Explainable AI (XAI). The widely used SHAP values, while quantifying feature importance, are often too intricate and lack human-friendly explanations. Furthermore, counterfactual (CF) explanations present ‘what ifs’ but leave users grappling with the ’why’. To bridge this gap, we introduce XAIstories. Leveraging Large Language Models, XAIstories provide narratives that shed light on AI predictions: SHAPstories do so based on SHAP explanations to explain a prediction score, while CFstories do so for CF explanations to explain a decision. Our results are striking: over 90% of the surveyed general audience finds the narrative generated by SHAPstories convincing. Data scientists primarily see the value of SHAPstories in communicating explanations to a general audience, with 92% of data scientists indicating that it will contribute to the ease and confidence of nonspecialists in understanding AI predictions. Additionally, 83% of data scientists indicate they are likely to use SHAPstories for this purpose. In image classification, CFstories are considered more or equally convincing as users own crafted stories by over 75% of lay user participants. CFstories also bring a tenfold speed gain in creating a narrative, and improves accuracy by over 20% compared to manually created narratives. The results thereby suggest that XAIstories may provide the missing link in truly explaining and understanding AI predictions.

**Problem:** Complex machine learning models are often difficult to interpret, many explanations methods are still too intricate for lay-users.

**Solution:** To enhance the understanding behind AI predictions for lay-users by providing natural language explanations.

*Currently revising for Decision Support Systems.*

For further details, see the arXiv paper.

### CoF Tables

### Abstract

The rise of deep learning in image classification has brought unprecedented accuracy but also highlighted a key issue: the use of ‘shortcuts’ by models. Such shortcuts are easy-to-learn patterns from the training data that fail to generalise to new data. Examples include the use of a copyright watermark to recognise horses, snowy background to recognise huskies, or ink markings to detect malignant skin lesions. The explainable AI (XAI) community has suggested using instance-level explanations to detect shortcuts without external data, but this requires the examination of many explanations to confirm the presence of such shortcuts, making it a labour-intensive process. To address these challenges, we introduce Counterfactual Frequency (CoF) tables, a novel approach that aggregates instance-based explanations into global insights, and exposes shortcuts. The aggregation implies the need for some semantic concepts to be used in the explanations, which we solve by labelling the segments of an image. We demonstrate the utility of CoF tables across several datasets, revealing the shortcuts learned from them.

**Problem:** As deep learning models achieve high accuracy in image classification, they often rely on ‘shortcuts’—simple, non-generalisable patterns learned from training data (such as recognising horses by copyright watermarks, or huskies by snowy backgrounds). These shortcuts can lead to failures when models encounter new data. Detecting these shortcuts typically requires examining numerous instance-level explanations, a process that is time-consuming and labour-intensive.

**Solution:** Aggregate instance-level explanations into global insights to expose shortcuts.

*Currently under review at SIGKDD, likely reworking and submitting to a journal*

An initial version of this work was presented at the 2023 ECML PhD Forum.

### Future

Below is a list of some of the projects I am currently working on or plan to work on in the near future. Titles in bold are currently being worked on, those in italics have yet to be started.

- **PCherry**: For a given data point and model, many different counterfactual explanations can be generated due to variations in algorithms, a lack of ground truth to optimise for, and non-deterministic behaviour. This phenomenon is known as the disagreement problem. Given this potential for variability, malicious actors can cherry-pick explanations to support a specific narrative, such as fair-washing. PCherry aims to provide a metric to evaluate the probability that a given set of counterfactuals includes cherry-picked explanations. Collaboration with two members of my research group.
- **AXA Collaboration**: Improving mail switching LLM through the analysis of counterfactual explanations.
- *CoF Tables 2.0*: Improving CoF tables through combinations of segments, new segmentation and edit functions. Ablation studies to measure effectiveness of different components.

## 3 PhD Study Programme

Below is a summary of the credits earned thus far (in bold), and the planned credits (in italics) for the PhD study programme.

| Competency | Name | Credits | Date |
|---|---|---|---|
| **A** | **DLCV** | **6** | **Jun 2023** |
| A | *Optimisation with Meta Heuristics* | 6 | 2024 |
| **C** | **Masters Supervision** | **1** | **Jan 2023** |
| **C** | **Masters Supervision** | **1** | **Jan 2024** |
| C | *Masters Supervision* | 1 | May 2024 |
| **D** | **XAIStories arXiv** | **2** | **Sep 2023** |
| D | *XAIStories WoS Publication* | 6 | 2024 |
| D | *CoF Tables* | 6 | 2024 |
| D | *PCherry* | 6 | 2024 |
| **E** | **ECML Poster** | **2** | **Sep 2023** |
| F | *Doctoral Day Discussant* | 1 | Oct 2024 |
| F | *Doctoral Day Speaker* | 1 | Oct 2024 |
| | | **Total Achieved Credits** | **12** |
| | | **Total Planned Credits** | **39** |

Given that there are maximum credits for each specific competency, here is a revised summation of planned credits:

| Competency | Capped Planned Credits | Maximum Credits |
|---|---|---|
| A | 12 | 15 |
| C | 3 | 4 |
| D | 15 | 15 |
| E | 2 | 8 |
| F | 2 | 4 |
| | **Capped Total Planned Credits** | **34** |

### Courses

The study programme requires at least 12 credits from courses approved by the university.

- **DLCV Summer School**: Deep Learning and Computer Vision summer school, confirmed with department to count as 6 credits for competency A.

- *Optimisation with Meta Heuristics*: I still need to complete the research portion of the course. For this I plan to optimise the combination of segment within the instance level CaFe explanations which form CoF tables. This functionality is required for a number of the papers I plan to do.

## Other Non-Research Credit Contributions

### Masters Supervisions

| Student Name | Date of Completion |
|---|---|
| Camille Dams | Jan 2023 |
| Margot Willemse | Jan 2024 |
| Tibo Vanleke | Expected May 2024 |

### Doctoral Day

The study programme requires that I participate in the Doctoral Day, as a discussant and speaker, as well as attending twice. I have attended before, and in the next Doctoral Day I plan to be both a discussant and a speaker, fufilling the requirements for competency F, and adding 2 credits to my total.

# 4 Other

- **Cambridge Summer School on Probabilistic Machine Learning** I was selected to particpate in the cambridge summer school on Probabilistic machine learning.
- **Reviewer for KDD** I reviewed 6 papers for SIGKDD 2024.
- *Research Stays* I hope that after my next round of papers, to be in a position to apply for some research stays at other institutions, with the aim to foster further collaboration and build my network.