Plan Overview

A Data Management Plan created using DMPonline

Title: Evaluating Explainable Reinforcement Learning with Objective Human Metrics

Creator: Balint Gyevnar

Principal Investigator: Christopher G. Lucas

Data Manager: Balint Gyevnar, Mark Towers

Affiliation: University of Edinburgh

Funder: UKRI Future Leaders Fellowships

Template: UoE Default DMP template for PGRs

Project abstract:

Explanations of machine learning systems are notoriously hard to evaluate because they are intended for humans and thus require running user studies to measure their effectiveness, which are both very expensive and cumbersome to conduct. As a result, explainable reinforcement learning (XRL) is struggling to establish a shared understanding due to compounding issues caused by a lack of comparative evaluations with humans and a lack of standardised benchmarks or rigorous evaluation metrics across the literature. To address these challenges, our competition will provide the first systematic comparison of XRL approaches with thorough human evaluation for debugging agent behaviour. Competitors will develop explanation mechanisms for agents trained in a gridworld-style resource-gathering environment for various publicly visible objectives and tested on held-out goals. To evaluate submissions, we will conduct a large-scale user study where participants will predict an agent's subsequent actions and underlying goals from an explanation of its decision-making. Combining objective metrics (prediction accuracy) with subjective assessments of explanation quality will establish the first large-scale comparative benchmark for XRL algorithms while giving baselines for future research to build upon.

ID: 174742

Start date: 01-03-2025

End date: 31-12-2025

Last modified: 31-03-2025

Grant number / URL: EP/S022481/1

imply that the creator(s) endorse, or have any relationship to, your project or proposal

Evaluating Explainable Reinforcement Learning with Objective Human Metrics

Administrative Information

1) School or Institute

School of Informatics, University of Edinburgh

2) Name and Contact details of supervisor(s)

Christopher G. Lucas, c.lucas@ed.ac.uk

3) Project start date

2025-03-01

4) Project end date

2025-12-31

Data Collection

5) Data Collection

Data will be collected using an online user study via the crowdsourcing platform Prolific and a survey built using the Qualtrics platform. Tabular text data will be collected and stored in a comma-separated file (.csv). We intend to recruit around 300 participants, collecting about 10 data points from each of them. Attention and comprehension checks will be used to guarantee high data quality.

Documentation & Metadata

6) Documentation & Metadata

The data will be accompanied by a README.txt file containing full metadata and the data folder structure. The file will contain a layout of the file structure of the data, including a description of each other file, their extensions, a summary of their contents, and how they can be used. The same file will also contain metadata about the fields of the dataset, including a brief summary of the field, its data type and expected values.

Ethics & Legal Compliance

7) Ethics & Legal Compliance

All researchers with access to the data will have completed the relevant data protection course. Ethics review and approval are sought according to the Informatics Research Ethics Process with reference number 257531. We will ensure that all personally identifying information is removed, keeping only basic demographic fields: age range and education level. No sensitive data will be stored. Data will be processed using password-protected and encrypted cloud storage OneDrive and moved to the DataShare service

of the University of Edinburgh for permanent storage. The data will be released under the MIT license.

Storage and Back-Up

8) Where will your data be stored and backed-up during the project?

Data will be processed using password-protected and encrypted cloud storage OneDrive and moved to the DataShare service of the University of Edinburgh for permanent storage. The data will be released under the MIT license.

Selection and Preservation

9) Where will the data be stored long-term?

The DataShare service of the University of Edinburgh will be used for permanent storage. The data will be released under the MIT license.

10) Which data will be retained long-term?

All cleaned and anonymised data will be retained for the long term to allow future reuse.

Data Sharing

11) Will the data produced from your project be made open?

- Yes: go to 12

12) How will you maximize data discoverability & access?

The data will be released under the MIT license, which allows full open access without the use of restrictions, subject to attribution. No embargo period is relevant for this data. We will store the data on Edinburgh DataShare, which will provide a persistent ID.

Responsibilities & Resources

14) Who will be responsible for the research data management of this project?

Balint Gyevnar

15) Will you require any training or resources to properly manage your research data throughout this project?

We have already done such data management training.