UNIVERSITEIT VAN AMSTERDAM

MSc ARTIFICIAL INTELLIGENCE
MASTER THESIS

Addressing Goal Misgeneralization in Offline RL with Natural Language

by
GIULIO STARACE
13010840

28th February, 2023

48 ECTS
November 2022 - August 2023

Supervisor:
NIKLAS HÖPNER

Examiner:
YOUR EXAMINER

Second Reader:
YOUR SECOND READER

INFORMATICS INSTITUTE (IVI)

Contents

Abstract

Provide a short overview and description of your thesis here.

Chapter 1

Introduction

This document provides an empty template for your thesis. Please keep the following things in mind when writing your thesis:

- The chapter organization of this template is a generic suggestion, please customize to your needs and vision.
- The guideline is to have a 40 page upper limit for the thesis, with additional optional pages for appendices if needed, for example to provide additional experimental results or mathematical details. Most theses do not end up with an appendix.
- For references, you are free to choose your own style. A standard style is included in this template, for example
- The official website of the thesis AI (https://student.uva.nl/ai/content/az/master-thesis-ai/master-thesis-ai-2020.html) contains all information regarding the procedures and details of the thesis process.

Appendix A

Optional appendix

Checking that citation works (Agarwal et al., 2022). Also inline Abbeel and Ng (2004).

Bibliography

Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA. Association for Computing Machinery.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. 2022. Reincarnating Reinforcement Learning: Reusing Prior Computation to Accelerate Progress. In *Advances in Neural Information Processing Systems*.