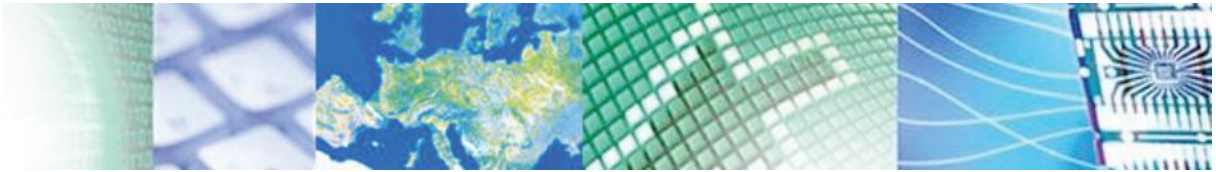


CHIST-ERA Projects Seminar 2022

Topic: XAI, Project: CausalXRL

The University Of Sheffield.

universität wien

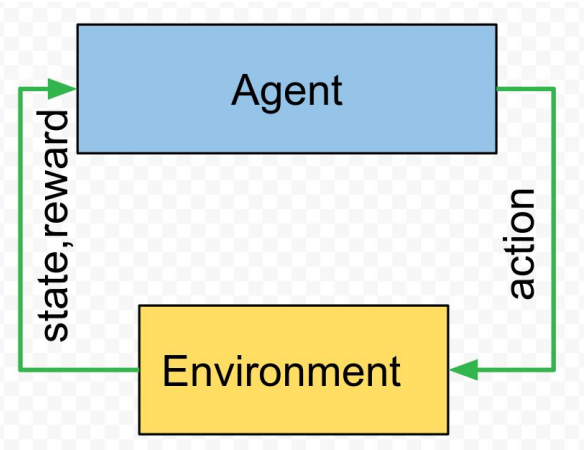
Inria

29 March 2022

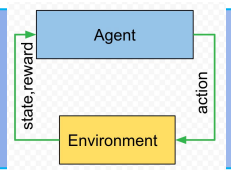
Programme co-funded by the EUROPEAN UNION

Introduction: Causal eXplainations in Reinforcement Learning (CausalXRL)

- **ML typically uses input-output correlations**
- **Reinforcement Learning (RL) uses random exploration**
- **Similar to humans, we will:**
  - Infer a causal model of the environment
  - Use that model to plan & suggest explainable actions on environment in RL loop

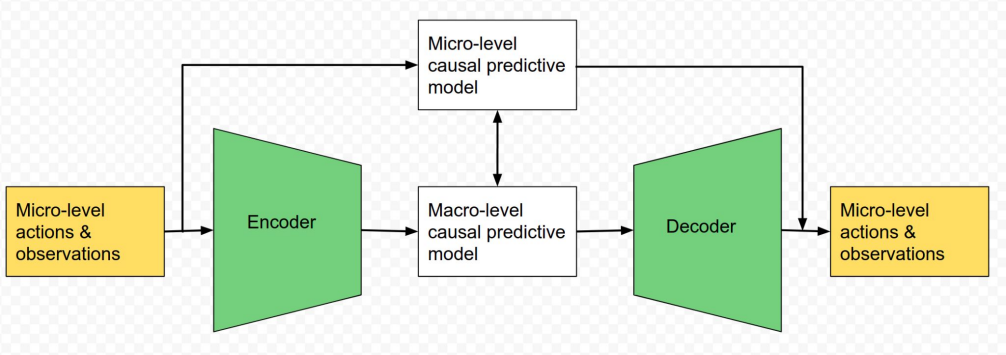


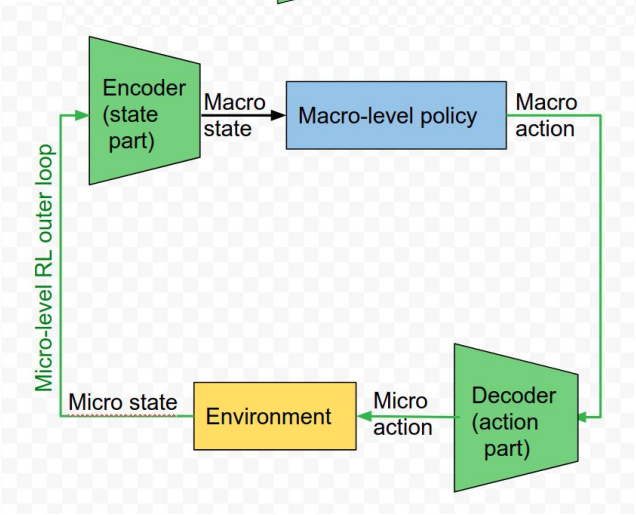
chist-era



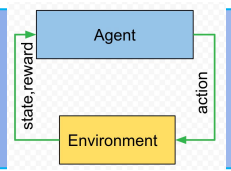
Architecture

- Learn a micro/macro-level causal model
- Explainable macro variables and actions
- Embed in an RL loop:
  - ✔ Macro-control



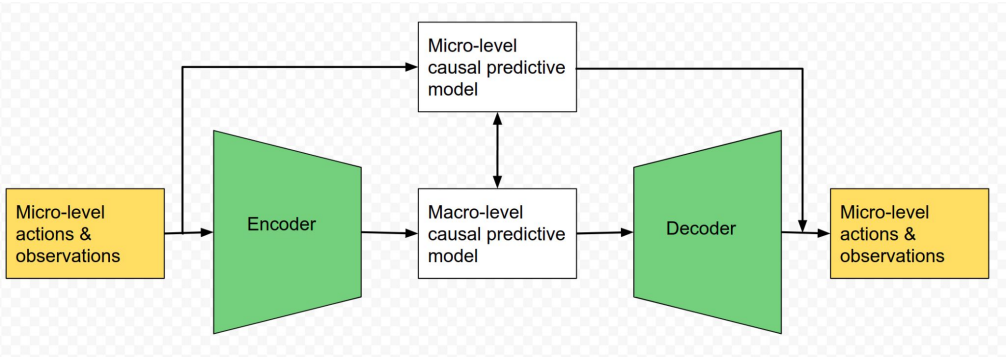


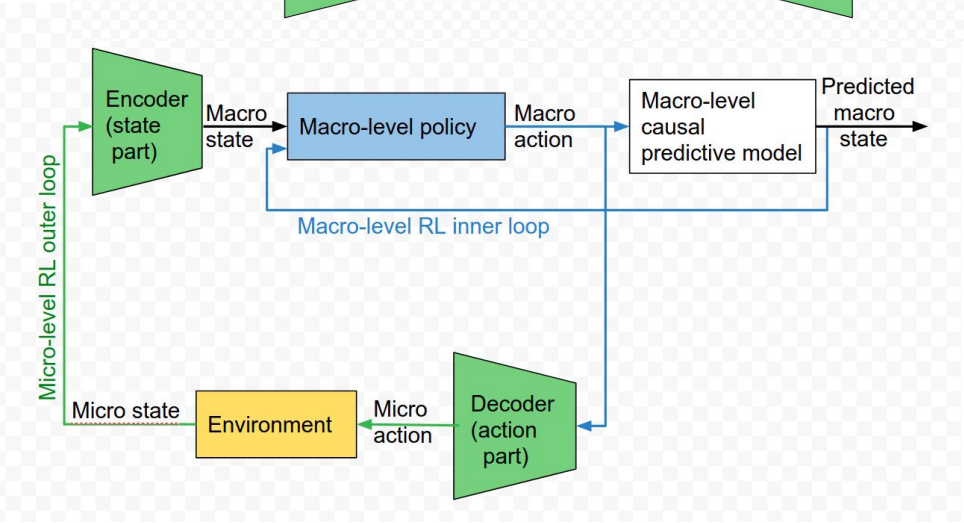
chist-era



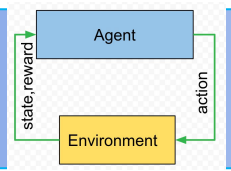
Architecture

- Learn a micro/macro-level causal model
- Explainable macro variables and actions
- Embed in an RL loop:
  - ✔ Macro-control
  - ✔ Macro-planning



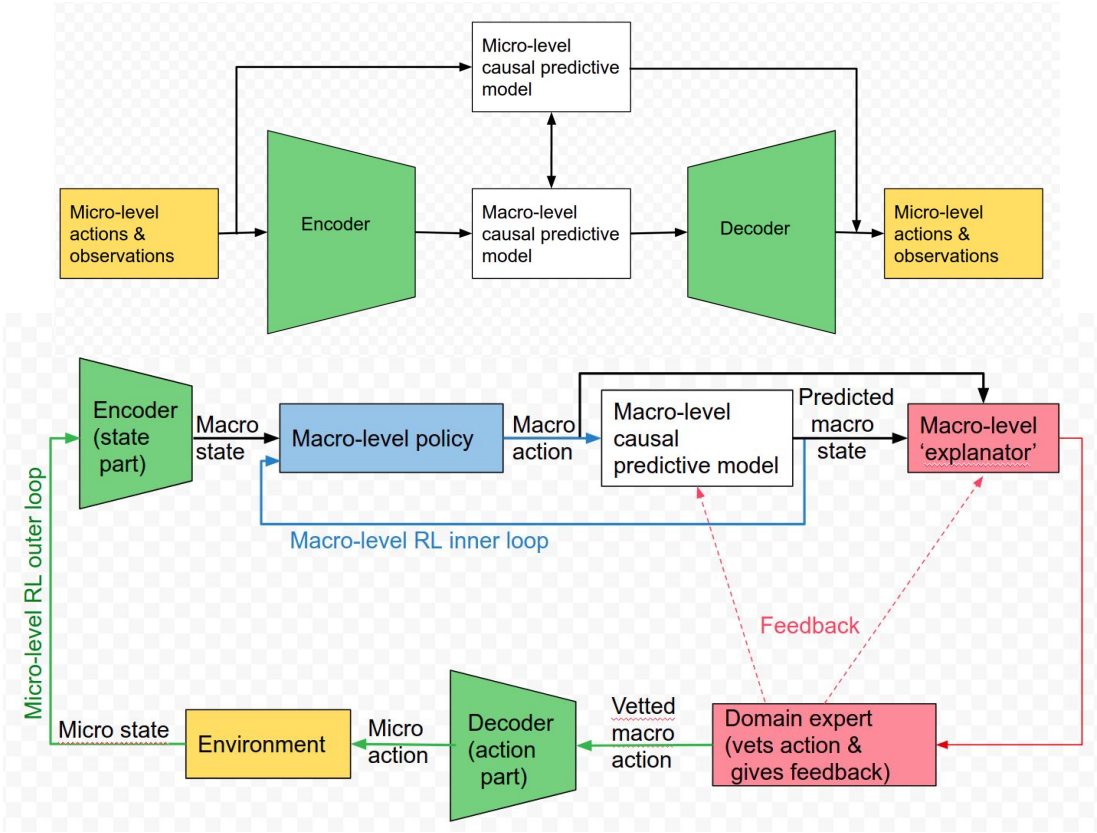


chist-era



Architecture

- Learn a micro/macro-level causal model
- Explainable macro variables and actions
- Embed in an RL loop:
  - ✔ Macro-control
  - ✔ Macro-planning
  - ✔ Explanations & Expert in loop



Partner Expertise & Contributions

The University Of Sheffield.

Chao Han
Aditya Gilra
(ex-coordinator-PI,
@ cwi.nl since Feb'22)
Eleni Vasilaki
(coordinator-PI)

Dynamical models & neural RL
Application: Neuromorphic

universität wien

Mauricio Gonzalez
Moritz Grosse-Wentrup (PI)

Multi-level causal inference
Application: Brain stimulation, E-education

Inria

Riccardo Della Vecchia
Philippe Preux (PI)

RL theory
Application: Intensive care, Farming

Timeline

Delayed start due to covid-compounded admin, visa & travel issues:

Extension till Jan'25 (Vienna), Oct'24 (Sheffield), Aug'24 (INRIA Lille)?

Challenges & progress

- Delve into new literature & code (2y since grant)
- Partially observable environments
- Model learning and RL from offline data, then refine online

The University Of Sheffield.

Auto-encoder architectures for causal model learning and RL

universität wien

Environments & algorithms for causal inference and RL

Inria

Instrumental variables for causal inference & RL

Collaboration: fortnightly meetings, Slack channel, Github

chist-era

Questions

Thank You!
Questions?