Plan Overview

A Data Management Plan created using DMPonline

Title: Safe Artificial Mental Models for Cyber-Physical Systems Learning

Creator: David Broman

Data Manager: David Broman

Affiliation: KTH Royal Institute of Technology

Funder: Swedish Research Council

Template: Swedish Research Council Template

Project abstract:

Recently, machine learning techniques related to large language models and reinforcement learning have witnessed tremendous advances. However, to safely and robustly incorporate these techniques into complex cyber-physical systems—such as humanoid robotics, autonomous vehicles, or industrial automation—is non-trivial. Specifically, large language models may return incorrect results, and training reinforcement learning algorithms on physical systems can lead to unsafe actions. Moreover, even if a combined pipeline of large language models and reinforcement learning of cyber-physical systems would work, it is hard to explain why it acts in a certain way: the explainability problem of using machine learning on cyber-physical systems. This project aims to develop a new theoretical foundation, robust design, and practical implementation for a new concept called artificial mental models, defined as domain-specific language programs. The purpose is to enable (i) safe user interaction using large language models, (ii) safe actuation using efficient reinforcement learning, and (iii) explainable actions validated in a formal setting, resulting in an overall trustworthy learning-based cyber-physical system.

ID: 164485

Start date: 01-01-2025

End date: 31-12-2028

Last modified: 21-12-2024

Grant number / URL: 2024-05043_VR

Safe Artificial Mental Models for Cyber-Physical Systems Learning

General Information

Project Title

Safe Artificial Mental Models for Cyber-Physical Systems Learning

Project Leader

David Broman

Registration number/corresponding

2024-05043

Version

1

Date

2024-12-21

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

Data is collected using physical robots or by using open data sets, such as open-source large language models. No sensitive information will be used. The data is only used to make the robots learn better policies to form certain tasks, i.e. by collecting information from sensors.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

Either publically open data (performance benchmarks or open-source benchmarks) or data collected by the robots that we are constructing. There is no sensitive data collected.
The data will be stored internally within the computer's memory when performing training. If the data from sensors needs to be stored temporally, it will be in raw file format, CSV, or JSON (for simplicity).
Note that the collected data is not any output from the project. It is only used as part of experiments. The output from the project will be new algorithms or artifacts in terms of open-source software code.

Documentation and data quality

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

The source code will be stored publically in the GitHub version source control. There is no secret data.
Data from sensors will not be stored; it will just be used temporarily during training in machine learning policies.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

In several conferences that we submit to, there are artificial validations. When we publish in such conferences, an external person validates that the results correspond to the data provided in the paper.
When we run performance experiments, each experiment is always repeated many times, and uncertainty statistics are presented.

Storage and backup

How is storage and backup of data and metadata safeguarded during the research process?

When we research, we store code and papers in Git repositories. Source code is typically developed in public, i.e., the source code is always open source. Papers are normally written in Latex and are stored in private Git repositories.
If any data is needed for reproducing experiments, we provide it in Github repositories or store it for long-term public availability in Zenodo, https://zenodo.org/.

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

We don't use sensitive or personal data in this research project.
Since we make all source code and other artifacts public via Github, publishers' repositories, or Zenodo, we are not in control of the security of these platforms.

Legal and ethical aspects

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

We don't handle any personal data in this project, and there is no confidential data. For algorithms, software, or other potential IPR innovations, the normal teacher's right applies, as in all Swedish universities.

How is correct data handling according to ethical aspects safeguarded?

There are no ethical aspects to data collection in this project. When it comes to the dual use of innovation of algorithms and software, we intend to make software and hardware design public.

Accessibility and long-term storage

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

There are no conditions, embargoes, or limitations. The software will be open source, and all data needed to reproduce experiments will be provided in Github or other open databases, such as Zenodo.

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

Long-term storage is safeguarded by Zenodo. CERN runs Zenodo and is responsible for its storage. It will be available as long as CERN exists. Publications are stored in the publisher's archive, in particular ACM and IEEE.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

Software is the key contribution of this work, not the data itself. Everything is documented in either GitHub or Zenodo repositories.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

Zenodo provides unique identities.

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

The PI, David Broman, is responsible. Most of the research will be conducted by PhD students and/or postdocs.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

We will use public and free repositories, storage systems, and databases, including GitHub and Zenodo, for the storage. The data will be public.