Neurosymbolic AI for Constraint Learning

A Data Management Plan created using DMPonline.be

Creators: Jessa Bekker, n.n. n.n., Wannes Meert https://orcid.org/0000-0001-9560-3872

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n.

Project Administrator: Jessa Bekker, Wannes Meert https://orcid.org/0000-0001-9560-3872

Grant number / URL: G047124N

ID: 206555

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

Constraints are ubiquitous in artificial intelligence and machine learning.

While everybody uses constraints, only few works exist that learn the constraints needed in constraint programming, combinatorial optimisation and (variations of) SAT-solving.

This project wants to develop the next generation of constraint learning techniques, which will not be based on traditional search and solver technology but rather on neurosymbolic AI. Neurosymbolic AI methods tightly integrate neural networks with symbolic AI methods. Neural methods will be used for learning the structure of constraints, and will also allow to learn constraints from mixed symbolic and subsymbolic data (such as images + logical descriptions). The project will address the learning of both hard and soft constraints as well as its applications.

Last modified: 25-04-2024

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- No

Neurosymbolic AI for Constraint Learning
GDPR

GDPR

Have you registered personal data processing activities for this project?

- No

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

We will reuse existing data from public domain sources and benchmarks and also create some new datasets of an artificial nature. No sensitive data will be used or collected. The datasets that will be used are concerned with neurosymbolic benchmarks, knowledge graphs, relational data etc.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Internal reports as well as papers will be made available using the Lirias system (KU Leuven).

For the short as well as for the long-term, the other relevant research data, such as software and the corresponding user guides and input data, will be stored on a secure NetApp-based storage solution at the Dept. of Computer Science (KU Leuven). Secure backups are automatically stored at a second location at KU Leuven, so loss of data is minimized.

IOF Fellow Wannes Meert and Research Manager Jessa Bekker in the DTAI section will help ensure and advise the PI so that also this project follows the DMP and Ethics requirements.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

We will not deviate from this principle. Rather we will make our data and software available in the public domain, whenever possible.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

In case it would turn out to be useful during the course of the project to use a sensitive dataset, we will first ask for ethical approval to the SMEC Ethics committee of KU Leuven.

Neurosymbolic AI for Constraint Learning
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Synthetic data	Synthetic data for obtaining an understanding of the abilities and limitations of the techniques	Generate new data	Digital	Simulation data		<100MB	
Neurosymbolic database (Application A)	Public domain dataset. Database that contains both symbolic and subsymbolic elements.	Reuse existing data	Digital	Observational		<1GB	
Dynamic systems dataset (Application B)	Camera images of conveyer belt at different points in time. Data available at our lab	Reuse existing data	Digital	Compiled/aggregated data		<1GB	
Constraint satisfaction and optimisation dataset (Application C)	Generated data. e.g. Sudokus	Generate new data	Digital	Simulation data, Compiled/aggregated data		<1GB	
Dataset for constraints as explanations for neural networks (Application D)	Generated data. It will show the use of constraints to help verify the behaviour of neural networks	Generate new data	Digital	Simulation data, Compiled/aggregated data		<1GB	
Algorithm implementations	Implementations of the developed algorithms	Generate new data	Digital	Software	.py	<100MB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

TBD

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Each newly introduced dataset will be accompanied with a README file that contains an explanation of the file structure and the content of each of the fields in the different files.
With each paper published within the context of this project, a runnable notebook or script will be provided that allows reproducing the experiments.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

The research is algorithm-oriented. The code will be shared according to the standards in our field. Typically, our algorithms are implemented in python and will then be shared in the the form of an easily installable python package.

3. Data storage & back-up during the research project

Where will the data be stored?

Software will be developed in KU Leuven GitLab or private GitHub repositories
Datasets will be stored either on the gitlab repositories or the NetApp-based storage solution at the Dept. of Computer Science (KU Leuven).

How will the data be backed up?

GitLab and the NetApp-based storage system have secure backups that are automatically stored at a second location at KU Leuven, so loss of data is minimized.

**Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.**

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Only authorized users can access the NetApp-based storage, GitLab repositories and private GitHub repositories.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

All storage and backup costs are covered by the project budget. After the project has ended, the costs will be covered by the DTAI research group.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data listed above will be retained for at least 10 years after the end of the project, according to KU Leuven policy.

Where will these data be archived (stored and curated for the long-term)?

All data will be archived on the NetApp storage service offered by the Department of Computer Science in the form of snapshots of all text, source code, data and presentations.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

All storage and backup costs are covered by the project budget. After the project has ended, the costs will be covered by the DTAI research group.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

The software implementation and datasets will be made available upon publication of the corresponding papers.

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

The software implementations will be made publicly available through GitHub repositories, as is common practice in our field.
The generated datasets will be made available alongside the code to generate them in GitHub repositories.

When will the data be made available?

Upon publication of the corresponding papers.

Which data usage licenses are you going to provide? If none, please explain why.

Permissive licenses. For software, this will be MIT, BSD, or Apache 2.0. For other data, this will be CC-BY-4.0.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

For software, we use the DOI of the relevant publication. The publication details where the code can be downloaded.

What are the expected costs for data sharing? How will these costs be covered?

Negligible. All storage and backup costs are covered by the DTAI research group.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The PhD and Postdoctoral researchers working on this project

Who will manage data storage and backup during the research project?

The PhD and Postdoctoral researchers working on this project

Who will manage data preservation and sharing?

Luc De Raedt

Who will update and implement this DMP?

Luc De Raedt and Stefano Teso