Plan Overview

A Data Management Plan created using DMPonline

Title: Generalization in Mind and Machine

Creator: Jeff Bowers

Principal Investigator: Jeff Bowers

Data Manager: Jeff Bowers

Affiliation: University of Bristol

Funder: European Research Council (ERC)

Template: ERC DMP

ORCID iD: 0000-0001-9558-5010

Project abstract:

Is the human mind a symbolic computational device? This issue was at the core Chomsky's critique of Skinner in the 1960s, and motivated the debates regarding Parallel Distributed Processing models developed in the 1980s. The recent successes of "deep" networks make this issue topical for psychology and neuroscience, and it raises the question of whether symbols are needed for artificial intelligence more generally. One of the innovations of the current project is to identify simple empirical phenomena that will serve a critical test-bed for both symbolic and non-symbolic neural networks. In order to make substantial progress on this issue a series of empirical and computational investigations are organised as follows. First, studies focus on tasks that, according to proponents of symbolic systems, require symbols for the sake of generalisation. Accordingly, if non-symbolic networks succeed, it would undermine one of the main motivations for symbolic systems. Second, studies focus on generalisation in tasks in which human performance is well characterised. Accordingly, the research will provide important constraints for theories of cognition across a range of domains, including vision, memory, and reasoning. Third, studies develop new learning algorithms designed to make symbolic systems biologically plausible. One of the reasons why symbolic networks are often dismissed is the claim that they are not as biologically plausible as non-symbolic models. This last ambition is the most high-risk but also potentially the most important: Introducing new computational principles may fundamentally advance our understanding of how the brain learns and computes, and furthermore, these principles may increase the computational powers of networks in ways that are important for engineering and artificial intelligence.

ID: 60487

Start date: 01-09-2017

End date: 31-08-2022

Last modified: 22-06-2021

Grant number / URL: 741134

Copyright information:

Generalization in Mind and Machine

Summary

Project Acronym

M and M

Project Number

741134

Provide a dataset summary

The data from the behavioral experiments and simulation studies will allow us to assess whether neural networks used in computer science provide a reasonable description of how the mind works. We are making the data available so that researches can replicate our analyses and confirm that the data we collected support the conclusions we draw. The data have been collected in various ways, from on-line experiments to eye-tracking data collected in laboratory. We have also made the code and the results from our simulation studies available. This will ensure that researchers can fully understand the simulation studies we have carried out and replicate our findings.

The main objective is to compare generalization in neural networks with human generalization across a range of domains (vision, language, memory, problem solving) to assess how similar neural networks are to human cognition. In addition, we try to modify networks in order to make them more human like, and this may involve adding “symbolic” mechanisms to networks.

The behavioral experiments are designed to assess human generalization in a range of domains so that we can compare human and model performance. Similarly, the modelling work assesses the generalization across a range of domain so we can compare human and model performance. In many cases the modelling work does not involve any parallel behavioral work as we are comparing the performance of models to data previously collected by others.

Types and formats of data generated/collected

- The Code folder for behavioral experiments consist of all the files necessary to run the experiment. Typically, this will consist of a .psyexp file which runs the experiment on Psychopy, and corresponding files (e.g., .png images used in the experiment, or a .yaml file used to interact with the eyetracker).
- The data folders for behavioral experiments typically consist of an analysis file (e.g., R script) which reads a specific datasheet (e.g., .csv file)
- The datasets for simulations consist of data files in .png / .jpg format
- The code is generally written in Python using Pytorch or Tensorflow & Keras

Origin of the data: Behavioral data was collected either in the laboratory (with participants recruited via University of Bristol’s participant mailing list or by their course credit scheme for Psychology undergraduates) or online using an online experimental platform like gorilla.ac or pavlovia.org (with participants recruited and paid via prolific.ac).

Datasets have generally been created by us. Where they have been borrowed from other sources, we have identified the source of the data in the paper and / or the Readme file associated with each project.

Expected size of the data

The behavioral data takes up little data storage, with all experimental findings expected to take up less than 1GB of storage. The simulation studies consist of large datasets varying between 5GB-50GB per project.

Data utility: to whom will it be useful

The behavioral data will be of interest to cognitive psychologists interested in a range of issues, from vision, language, and memory. It will allow researchers to replicate our analyses if there are any questions regarding our conclusions. The modelling data will be of interest to psychologists, neuroscientist, and computer scientists interested in our claims regarding the similarities of artificial networks and brain. Again, the code and data will allow researchers to fully understand our simulations as well as replicate our findings.

FAIR data and resources

1. Making data findable

All data will be uploaded together with the relating metadata, including project context to the university RDSF server. These collections will be linked to scientific articles, conference proceedings, reports, and other sources to be published. For this, we will make use of persistent and unique Digital Object Identifiers (DOI) via the data storage facility. A description of available data collections will also be added to the PIs website.

Metadata consists of a README.md file that instructs on how to run the code and scripts is included, as well as metadata that defines the software requirements. This file also links each data folder to the related publication identified using it's DOI.

In addition, most projects are linked to a version-controlled Github repository which tracks changes made to the project and provides a platform for sharing code as well as data. See: https://github.com/mmrl

All data used in any publication (journal article, conference proceeding.) will be made openly available and linked to via DOI from the original publication.

Outline naming conventions used

Data is organised into folders, with each folder consisting of two sub-folders:

- exp_mind: This contain data, code and scripts for human experiments
- exp_machine: This contains data, code and scripts for simulations

Each of these folders consist of sub-folders for: "code" and "data". The directories are, in turn, organised according to the experiments / simulations / figures in the paper linked to the folder and identified in the README.md file.

Approach towards search keyword

1. Data organized by paper or conference proceeding.

Approach for clear versioning

Data and code are linked to Github repositories that provides version control.

Metadata creation

No standards are known to the PI at the point of the DMP creation. As metadata, we will thus provide a README.md file associated with each folder that identifies:

Published results,

Title,

Authors including contact information,

Description,

Date of creation, and

References to all publications referring to the dataset.

2. Making data openly accessible

All data will be made available, along with open access to all publications.

Data will be made available through: (a) published results, (b) Github / OSF repository and (c) the university's Research Data Storage Facility (RDSF). All software, data collected during experiments and datasets will be openly available for any researchers to use. Where subject data has been collected, it will be anonymised.

All data will be made available on University of Bristol's Research Data Storage Facility (RDSF). When data is ready to be published a DOI for this data will be generated and the data will be made openly available through (i) the University of Bristol Research Portal (https://research-information.bris.ac.uk), and (ii) DataCite (https://datacite.org). Where possible, data and code will also be made available over Github and through the Open Science Framework (https://osf.io).

There will be no restrictions.

3. Making data interoperable

For the simulations, most of the code is written is Python using either Pytorch or Tensorflow+Keras libraries, which means code written can be easily ported from one machine to another. The code for conducting human experiments is generally written in PsychoPy and analysis is performed in either Python, R or SPSS. We will identify the key requirements for running the code as well as scripts in the README.md file associated with each project.

We will use standard vocabulary for all data types present in our datasets.

4. Increase data reuse

How the data will be licenced to permit the widest reuse possible

We will use a GPL-compatible license

All data will be put up at the same time as open access publications are uploaded on public repositories as well as the ERC website at: https://mindandmachine.blogs.bristol.ac.uk In no case is an embargo period foreseen.

Anyone will be welcome to use the data we make available. There will be no restrictions.

Data quality assurance processes

Loss of data is avoided by storing a copy of data on the University of Bristol's Research Data Storage Facility (RDSF), which is automatically backed up. In addition, most projects use Github and OSF for version control.

Length of time for which the data will remain re-usable

The data will remain re-usable for at least 20 years, which is the policy of University of Bristol's Research Data Storage Facility.

5. Allocation of resources and data security

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

No costs.

Responsibilities for data management in your project

The PI takes primary responsibility to ensure data management is fully compliant with ERC requirements.

Costs and potential value of long term preservation

There are no costs to the funder, and the datasets and code will allow others and members of the team to replicate any simulations or analyses if the need arises.