Plan Overview

A Data Management Plan created using DMPTuuli

Title: AI Personality and Cognition (AiPerCog): AI learning from Humans in Games

Creator: Benjamin Cowley

Principal Investigator: Benjamin Cowley

Data Manager: Benjamin Cowley, Lauri V Ahonen

Project Administrator: Natalia Postnova

Affiliation: University of Helsinki

Funder: The Research Council of Finland (former The Academy of Finland)

Template: Academy of Finland data management plan guidelines

ORCID iD: 0000-0001-8828-2994

Project abstract:

Digital games are played by an estimated 3 billion people today. AI technology embedded in widely-played games could make them a fine instrument to study psychological states of large populations. AiPerCog aims to find behavioural markers of psychologically vulnerabilty, such as propensity to depressive or anxiety disorders, based on computer gaming behaviour. To do this we will: (1) data-mine very large datasets of game replays to identify human patterns of play, (2) train AI agent models to replicate these patterns, & (3) build classifiers to link psychological profiles to play patterns. We apply this methodology firstly to open databases of game-play records from popular games (Poker, CounterStrike 2, StarCraft II), and secondly to play data from corporate partners, combined with survey data from open recruitment among their customer base. This approach will allow us to understand how different types of reinforcement from game situations drive human player behaviour. Our work will produce AI that can profile and produce insights into its users, and advance the understanding of how human cognition in digital environments.

ID: 11790

Start date: 01-09-2023

End date: 31-08-2027

Last modified: 30-10-2023

Grant number / URL: 355200

Copyright information:

AI Personality and Cognition (AiPerCog): AI learning from Humans in Games

1. General description of data

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Also give a rough estimate of the size of the data produced or collected?

What kinds of data

We use mainly questionnaires and behavioural data from game players.

Questionnaires consist of surveys on player preferences and motivations to play, plus background data, for validation datasets.

Behavioural data is the log of individual actions within a game, defined as activity-related variables of sufficient detail that complete games can be replayed, given a suitable game engine software.

Much of this data is already available either from online repositories, or from project partners including several game development companies. A smaller portion of validation data will be collected from recruited subjects, via a downloadable software platform.

Some portion of the replays available online, produced by eSports tournaments, can be linked to commentaries on the games given by domain experts (as in, sports commentaries). Tournament games can also be linked limited background data on the players themselves, such as name, age, gender, and competition history.

All data will be controlled by HiPerCog group, University of Helsinki, using servers provided by CSC (epouta) and IT services (datacloud).

Table 1: non-sensitive data: *no possibility of privacy violations*

Gb	Datatype	Source	Format	#
100	StarCraft II game replays	Blizzard Games	*.SC2replay	1
1-2	StarCraft II / Hearthstone tournament game replays	Various websites, e.g. lotv.spawningtool.com	*.SC2replay	2
<1	StarCraft II / Hearthstone tournament game player background	Various websites, e.g. lotv.spawningtool.com	*.csv	3
<5	StarCraft II / Hearthstone tournament game expert commentaries	Various websites, e.g. twitch.tv	*.mp3	4
<5	Hearthstone public repo	hearthscry.com	*.json	5
<1	Existing survey responses on play styles & preferences	Kinrate Analytics Oy	*.csv	6

Table 2: sensitive data: *identifying, revealing of personal information, or legally protected*

Gb	Datatype	Source	Format	#
<5	Replays for testbed games, Starcraft II, Hearthstone, poker, chess	Live participants recruited online	*.SC2replay & *.json	7
<1	Player personality questionnaires	Live participants recruited online	*.csv	8
>10	Corporate partner game replays	Next Games Oy, a Netflix company	to be determined	9
>10	Corporate partner game replays	Futureplay Games Oy	to be determined	10
<1	Newly gathered survey responses on: 1.play styles/preferences 2.psychological state 3.lifestyle/wellbeing state	Open recruitment from customer base of corporate game development partners	*.csv	11

What data

The project collects data primarily from large open datasets of game-play recordings published online by games companies to promote research, and by eSports tournaments to promote their events. These are not always available as single repositories, so (where needed) we will gather them by web-scraping and curating from the sources. In addition, the project will collect game replays from state of the art AI agents released online (AlphaStar Team, 2019), and generate its own very large datasets of game replays produced by AI agents playing the testbed games.

To facilitate study of player personality, the project will have access to ~14000 respondents to surveys on player preferences and motivations to play completed by respondents from Finland, Denmark, Canada, Japan, USA, UK, Korea and Brazil (one of the largest datasets of its kind, gathered by Dr Jukka Vahlo and his startup company Kinrate Analytics in 2014-2019). Further, in order to validate the modelling work, the project will gather game records for each testbed game, plus background data and psychological self-reports including playing personality tests, from at least 100 live subjects.

The project will work with corporate partners to analyse game replay data from their customer base. We will then conduct open recruitment (advertising to those customers via our own channels) to obtain a sample of survey data on player preferences and personality which we can match to the play data by customers consent.

File formats and size estimate

Game replay and self-report / background data are in common open plaintext data formats, such as json, csv, or similar. Large datasets obtained online scale to hundreds of gigabytes of such data. Validation data (100+ subjects playing three games) will generate much less data, on the order of gigabytes.

Audio files of tournament game commentaries are converted from online videos, typically saved in mp3 format. These files will take tens of gigabytes of storage. This data is already public-domain and speakers are identified along with the recordings, so privacy violation is not possible.

1.2 How will the consistency and quality of data be controlled?

We will use a relational database software to manage data organisation, version control, and access.

Online data sources are of consistent quality: game replays are created by the game engines themselves, programmed by the original game developer and tested by millions of users. Video/audio commentaries are produced at a professional standard equivalent to traditional sports reporting, with high consistency.

Protecting the identity of players who produce the data, by minimisation, pseudonymisation, and anonymisation will not affect data quality.

For the validation data gathered from live subjects, the project researchers have extensive experience from similar past recordings, and will use an automated platform to encapsulate data gathering. This means (a) we have extensive ability to plan failsafes, (b) we can run the data gathering for as long as needed to get the quantity, quality, and consistency of data we need.

Corporate-sourced data is also of high quality, as corporations themselves also study their data for their own purposes.

2. Ethical and legal compliance

2.1 What ethical issues are related to your data management, for example, in handling sensitive data, protecting the identity of participants, or gaining consent for data sharing?

For data gathered from live subjects (datasets 7 & 8 from table 2), typical ethical procedures will be followed, including approval by an ethical review board (at one of the participating institutions), following guidelines of the Declaration of Helsinki, full briefing and informed signed consent. Such data will also be anonymised before publishing, and will only consist of forms of data which can and have been safely anonymised.

The HiPerCog group and Faculty of Educational Sciences, who are responsible for collecting data from live subjects, have extensive experience performing similar research to collect similar data, and are thus very well equipped to manage the data safely and ethically. The project will obtain the ethical permission needed from UH's relevant institutional review board, promptly after being granted funding.

For online data gathered from open sources, the project will not use or share the data in any way not covered by the licence terms that already exist. Additionally, all such data will be analysed in anonymised form, such that identification of original individuals is not possible (either because the data has been anonymised at source, or by removing all identifiers prior to analysis). For inherently identifying forms of data, such as voice recording, preliminary analysis will be conducted in order to extract features which are not identifying, such as sentiment labels, and only those anonymous features will be fused with other data sources for further analysis.

2.2 How will data ownership, copyright and IPR issues be managed? Are there any copyrights, licences or other restrictions that prevent you from using or sharing the data?

Ownership, copyright and IPR issues

All sources of online data (datasets 1 -- 6, table 1) have been confirmed to provide explicit rights to use the data for research. For sharing such data as a project output, the original source will be identified and linked (rather than re-published).

Data recorded directly from live subjects (datasets 7 -- 11, table 2), or produced by AIs, will be owned by the collaborating institutions of the AiPerCog project (University of Helsinki, datasets 7, 8, 11; corporate partners, datasets 9 & 10).

Agreements with all partners / researchers will be made in the beginning of the project to define issues of ownership and sharing of the data.

For sharing data, we will use MIT licence.

Restrictions

There are no restrictions that prevent the datasets 1 -- 5 being used and shared in the manner planned in the project. Datasets 7, 8 & 11 require ethical approval and anonymisation. Datasets 6, 9 & 10 require legal contracts of data sharing with the corporate partners.

Agreements

Contracts will be drawn up to govern how the data will be shared by corporate partners: the partners will provide data for analysis but not for redistribution, nor can all details of the data be disseminated.

3. Documentation and metadata

3.1 How will you document your data to make them findable, accessible, interoperable and reusable for you and others? What kinds of metadata standards, README files or other documentation will you use to help others understand and use your data?

Data documentation

During the project, data will be managed by standard methods:

Metadata standards
Data management software, i.e. database
File naming conventions
Fixed directory structure
Readme file(s)
Version control

Filenaming conventions: All data will be identified with short, human-readable but anonymous keys, which contain following information.

- ID number unique to each human subject or AI agent
- Short text key for data type, e.g. SC2 for StarCraft2, or MOT1 for play motivations questionnaire 1
- Identifier for session number, e.g. match3 for 3rd match played by that individual

This system essentially allows each piece of data to be uniquely identified.

Datasets will be published with documentation describing the structure, licencing, and associated software resources (scripts for processing or analysis).

Metadata

Meta-data for the overall study will be formulated as a tabular database, organised congruently to the data itself. Thus the main key for data will be subject ID, linking to tables for each session (including trait data recording).

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will they be backed up?

Sensitive data will be stored on large-capacity server solutions provided by **CSC – IT CENTER FOR SCIENCE LTD.**, which provides high performance computing solutions that can be shared across project institutions. Non-sensitive data will be stored on the datacloud.helsinki.fi server provided by the UH IT services.

Both servers automatically handle data backups.

HiPerCog group already has an own node provided by CSC, and own datacloud solution.

The directory structure for data storage will then be sorted by subject ID. Into each subject ID folder, all subject-related data will be arranged in terms of meta-data, trait-wise background data, and session-wise data recordings.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

Project data will be hosted on secure servers with access controlled by project PI, and available only to project researchers. Identification of subjects will be even more tightly controlled, with the key to link individuals to their unique ID codes encrypted and accessible only by project PI.

5. Opening, publishing and archiving the data after the research project

5.1 What part of the data can be made openly available or published? Where and when will the data, or their metadata, be made available?

All data is envisioned to be publishable in different senses.
The large datasets already available online need only be identified/linked in our research reports. This will happen in the first few months of the project.
The project will be given permission by subjects whose data are recorded to share it in anonymised form. Their data will be available midway through the project.
Towards the end of the project, software scripts to process or analyse the data will be published in online repositories, on Github.
Once the associated data is fixed (all recordings complete, all processing software finalised), meta-data will be published to the Etsin service.
The datasets from live subjects will be made available through online data respository Zenodo, receiving a permanent digital identifier.
For large third-party datasets, as a contingent measure, we can mirror those data on Finnish servers (without breaching any licence terms), to ensure they will not be removed in future.

5.2 Where will data with long-term value be archived, and for how long?

n/a

6. Data management responsibilities and resources

6.1. Who will be responsible for specific tasks of data management during the research project life cycle? Estimate also the resources (e.g. financial, time and effort) required for data management.

Project PI Cowley will be responsible for initial planning of the data management procedures. They will then delegate responsibility for detailed implementation to their respective post-doctoral researchers (Lauri Ahonen and Natalia Postnova), who handle design of the web-scraping software, the psychometric questionnaire software, and the software to encapsulate games for live data collection.
Finally, day to day data management tasks will be performed by post-graduate students. This is expected to take no more than 3h / week during 2-3 weeks before opening the data in a repository.