EMERGE

WP9 Project Management and Coordination

D9.4 Updated Data Management Plan

Version: 1.0

Date: 31/03/2025

Document control

Project title	Emergent awareness from minimal collectives
Project acronym	EMERGE
Call identifier	HORIZON-EIC-2021-PATHFINDERCHALLENGES-01-01
Grant agreement	101070918
Starting date	01/10/2022
Duration	48 months
Project URL	http://eic-emerge.eu
Work Package	WP9 Project Management and Coordination
Deliverable	D9.4 Updated Data Management Plan
Contractual Delivery Date	M30
Actual Delivery Date	M30
Nature[1]	DMP
Dissemination level[2]	PU
Lead Beneficiary	UNIPI
Editor(s)	Davide Bacciu (UNIPI)
Contributor(s)	Monica Marrucci (UNIPI), Nadine Meertens (LMU), Simon Jones (UOB)
Reviewer(s)	Renan Picoreti Nakahara (DVL)
Document description	This deliverable provides an updated version of the Data Management Plan for the EMERGE project, building on the deliverable "D9.2 – Data Management Plan" provided at M6. It reports an update on the data collected within the project and outlines how these data, along with other research outcomes, are made available and exploitable by the community

---

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Version control

Version[3]	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.1	Monica Marrucci	12/03/2025	Revision of the first version of the document and updating of the basic information to align with the updated version
0.2	Davide Bacciu	14/03/2025	Integration of changes to the data and code management platform. Updates on benchmark datasets.
0.3	Nadine Marteen	19/03/2025	Updates on datasets collected by LMU.
0.4	Simon Jones	20/03/2025	Updates on datasets collected by UOB.
0.5	Davide Bacciu	20/03/2025	Complete draft ready for review
0.9	Renan Picoreti Nakahara	25/032025	Document approved by reviewer
1.0	Davide Bacciu	31/03/2025	Document released by Project Coordinator

---

[3] 0.1 – TOC proposed by editor; 0.2 – TOC approved by reviewer; 0.4 – Intermediate document proposed by editor; 0.5 – Intermediate document approved by reviewer; 0.8 – Document finished by editor; 0.85 – Document reviewed by reviewer; 0.9 – Document revised by editor; 0.98 – Document approved by reviewer; 1.0 – Document released by Project Coordinator.

Abstract

The report is the updated version of the Data Management Plan for the EMERGE project. It provides an update on the data collected within the project as well as the details on how those data, as well as other research outcomes, are made available and exploitable by the community.

Consortium

The EMERGE consortium members are listed below.

Organization	Short name	Country
Università di Pisa	UNIPI	IT
TU Delft	TUD	NL
University of Bristol	UOB	UK
Ludwig Maximilian University of Munich	LMU	DE
Da Vinci Labs	DVL	FR

Disclaimer

This document does not represent the opinion of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA), and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

This document may contain material, which is the copyright of certain EMERGE consortium parties, and may not be reproduced or copied without permission. All EMERGE consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the EMERGE consortium as a whole, nor a certain party of the EMERGE consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk and does not accept any liability for loss or damage suffered by any person using this information.

Acknowledgement

Table of contents

List of abbreviations

CSV	Comma Separated Value (file)
DM	Data Manager
DMO	Data Management Officer
DMP	Data Management Plan
DoW	Description of Work
FACT	Fair, Accurate, Confidential, and Transparent
FAIR	Findable, Accessible, Interoperable, and Reusable
HLEG	High Level Expert Group
RI	Research Infrastructure
WP	Work package

List of figures

1. Executive summary

This document is a deliverable of the EMERGE project, funded under HORIZON-EIC-2021-PATH FINDER CHALLENGES-01-01 under grant agreement number 101070918.

The document updates and integrates the former deliverable "D9.2 Data Management Plan" submitted at M6 and which structure is adapted from the Horizon Europe DMP template available here: https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf

The DMP describes the data management life cycle for the data to be collected, processed, and/or generated by EMERGE. As part of the effort of making research data findable, accessible, interoperable, and reusable (FAIR), this DMP includes information on:

- what data will be collected, processed and/or generated, and whether such data will be shared/made open access (Section 2);
- the handling of research data during & after the end of the project, which methodology & standards will be applied (Section 3);
- what further research outputs are produced and how they are made FAIR-ly available (Section 4);
- how data will be curated & preserved (Section 6 and 7).

The partners who handle data in the EMERGE project are: UNIPI, TUD, LMU and the associated beneficiary UOB. DVL does not handle research data.

The deliverable "D9.4 Data Management Plan" describes the updated version of the Data Management Plan (DMP) at M30 of the EMERGE project. The current deliverable is still conceived as a living document as it was its former version, i.e. as needs arise and data or data management tools or standards change, we will include them in the report and we will submit the updated and final version of the document, as foreseen by the Grant Agreement, by 30 September 2026 (D9.6 Final Data Management Plan).

To ease versioning and tracking of the changes, the table below reports the significant updates contained in the different versions of the DMP (and relevant documentation and resources) released to date. We do not track minor linguistic changes.

The key change to the DMP is the shift in FAIR policies, designating the AI-on-Demand (AIoD) platform as the primary point for EMERGE data sharing instead of SoBigData. This decision aligns with the EC's guidance, identifying AIoD as the reference platform for the European AI ecosystem of tools, data and knowledge.

Title	Description	Link
D9.2 Data Management Plan (*Date: 31 March 2023*)	First version of the DMP	Publicly available on the EMERGE project website at this link.

D9.4 Updated Data Management Plan (*Date: 31 March 2025*)	Updated version of the DMP containing the following significant updates: - Revised description of the data collected and used within the project, with statistics and details about the data collected until M30 (Section 2.1). - Revised list of public benchmark used in the project, based on the scientific works published until M30 (Section 2.2). - Update on FAIR data policies, describing a change in the service being used for data sharing from the SoBigData platform, to the AI-on-Demand (AIoD) platform (Section 3). - Updates on project software released publicly (Section 4). - Removal of Appendix A (previously reporting information on SoBigData).	Will be made available on the link above upon approval by the EC.

2. Data summary

2.1 Type, format and size of collected/generated data

In the following we describe the original datasets that are collected within the project, providing as much details as possible at the current stage of the project. Further details will be added in the final version of this report. For each dataset we identify the partner responsible for its collection (in bold) and we list all partners which are planning to handle the data therein.

Moral Psychology & Experimental Ethics Data (**LMU**; UNIPI, UOB)

TYPE

This data are collected within the scope of the activities of WP2 (in particular T2.2 and T2.3) aimed at testing how the new concept of collaborative awareness stands regarding two key

ethical risks for humans when relating to AI agents, i.e. the lack of understanding and the tendency to anthropomorphise their awareness. The leading partner managing the experiments and data collection is LMU.

For the purposes above, we collect questionnaires, demographic data, and behavioural data in psychological experiments. This is done both via online experiments as well as physical lab-based experiments.

We conduct empirical studies to investigate human evaluations of and interactions with artificial agents using established and fairly well-known behavioural game theory and vignette-based study methods. The collected data involve choice in various scenarios, judgements of situations, and ratings, e.g., of preferences and emotions. This entails recording human participants' decisions and other evaluative and performance metrics in the given tasks, and collecting demographic data such as the participant's age, gender, nationality.

FORMAT

At the start of the emerge project we planned for all data to be collected in the form of CSV files. All data would be and now is stored in offline digital storage. For online experiments, we planned to collect data via online dedicated platforms such as Amazon Mechanical Turn (Mturk) and Prolific: these platforms maintain the identification of participants and only an anonymized ID would be entered into the offline storage at the partner site. For in-laboratory based experiments, the data is collected on hardcopies, and again only an anonymized ID is stored on the digital storage.

To date, all our experiments within the project were conducted online. We used both MTurk and Prolific to recruit our experiment participants. Our conducted studies were set up using LIONESS Lab (for behavioural game theory) and Qualtrics (for vignette-based studies) software tools.

SIZE

We estimated that the data size would be less than 100 GB. The sample size of each experiment would generally be around 100 participants. As part of the EMERGE activities, we expected to perform about 10 experiments, with a total sample size ranging in about 1,000-1,500 people.

Thus far, we conducted 8 experiments and collected and analysed data from 1,975 (134 + 161 + 199 + 200 + 200 + 201 + 440 + 440) participants' responses. All raw data collected to date is stored in CSV files and, together with data analysis code and research reports that were prepared predominantly using the statistical analysis software R.

The Ethics Committee of the Faculty of Philosophy, Philosophy of Science and Religious Studies at LMU Munich (ref. 238497) and/or The University of London School of Advanced Study Research Ethics Committee (ref. SASREC_1819_313A) approved our studies after they were reviewed for compliance with ethical research standards and GDPR provisions. We obtained informed consent from all participants who took part in our studies and all studies adhered to the ethical principles of the Declaration of Helsinki.

Swarm Experimental Data (**UOB**; UNIPI, TUD)

TYPE

These data are collected within the scope of WP5 and WP6 activities, consisting of logs from physical experiments with DOTS robots. The data support the demonstration of the Awareness in Swarm use case (T6.3) and the validation of emergent awareness measurement and control methodologies from T5.1 to T5.3. UOB is the leading partner responsible for the experiments and data collection.

Initial work on the Distributed Spatial Awareness experiment was conducted in simulation, and physical robot experiments are now beginning. For the Framework for the Examination of Awareness in Artificial Systems, all work was completed in simulation. Results from both experiments include raw CSV files and are processed using Python scripts. The experimental code is developed in C++ for both experiments, with additional Python code used specifically for the Framework for the Examination of Awareness in Artificial Systems.

FORMAT

Data includes logs from onboard sensors and state information from the DOTS robots, as well as video recordings of the experiments. Complementary resources, such as scripts and code for experimental analysis, are also included. The data and associated resources will be made openly available in accordance with Section 3 provisions.

For the Distributed Spatial Awareness project, results so far include 400 MB of raw CSV files, with analysis performed using Python scripts. For the Framework for the Examination of Awareness in Artificial Systems, the results amount to 300 MB of raw CSV files, with similar analysis methods using Python.

SIZE

The expected data size is in the range of gigabytes per experiment. The Distributed Spatial Awareness project has currently produced 400 MB of raw data, while the Framework for the Examination of Awareness in Artificial Systems has generated 300 MB of raw data.

Swarm user studies (**UOB**; UNIPI, TUD, LMU)

TYPE

These data are collected within the scope of the activities of WP5, and in particular to assess the ability of users to monitor and control emergent awareness of our robot swarm (activities T5.1, T5.3 and T5.4). The leading partner managing the experiments and data collection is UOB.

For this purpose, we collect anonymised questionnaires and interviews, associated with logs of experiments. The data will also be linked to scripts for its analysis.

These data and associated resources will be made openly available in anonymised form, according to the provisions in Section 3.

FORMAT

To date, collected data include video files of trials and eye-tracking data stored in CSV files, supplemented with scripts for analysis. Specifically:

- Participant Questionnaires: 21 individual responses in PDF format (approx. 5.3 MB each).
- Trial Videos: 21 recordings of participant interactions in MP4 format (approx. 482 MB each).
- Eye-Tracking Data: Each participant has 7 CSV files and one video recording stored in a dedicated folder (total size per participant: ~82.8 MB).

SIZE

The data collected so far (including questionnaires, trial videos and eye-tracking data) total about 12 gigabytes of storage space. Further experiments will expand the data size accordingly.

2.2 Reuse of existing data and its origin

In the first phases of development of the Archetype Computing System, we made use of publicly available datasets to benchmark in learning architectures and algorithms, also for publication purposes. Specifically:

ADIAC - Publicly available dataset containing sinusoidal-like patterns extracted from images with the objective of identifying 37 different diatoms classes. (SOURCE: timeseriesclassification.com)

CIFAR10 - A public dataset representing a classic benchmark for neural networks. This is a dataset of 50,000 32x32 colour training images and 10,000 test images, labelled over 10 categories. As for the MNIST dataset, it is used to test different architectures and learning strategies, in a variety of settings. (SOURCE: https://www.cs.toronto.edu/~kriz/cifar.html)

ECG200 – Publicly available datasets where each time series traces the electrical activity of a subject recorded during one heartbeat. The task is a binary classification prediction between a normal heartbeat and one highlighting a Myocardial Infarction. (SOURCE: timeseriesclassification.com)

LIBRAS - Publicly available dataset focusing on the recognition of 15 different hand movements from the Brazilian sign language. (SOURCE: timeseriesclassification.com)

MALLAT - Publicly available signal processing dataset containing input signal belonging to 8 different classes and characterized by an high imbalance between training and test sequences (in favor of the latter). (SOURCE: timeseriesclassification.com)

MNIST – A public dataset representing a classic benchmark for neural networks. The dataset contains 70000 28x28 images representing handwritten digits. This dataset is used to test different architectures and learning strategies, in different settings. (SOURCE: http://yann.lecun.com/exdb/mnist/)

TRACE - Publicly available dataset is a dataset focusing on the recognition of 4 different classes of instrument failures in a nuclear power plant, from timeseries information. (SOURCE: timeseriesclassification.com)

WESAD – A publicly available dataset for recognition of human emotional state recognition. The dataset features multimodal physiological data gathered through wearable devices from 15 subjects. (SOURCE: https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/)

3. FAIR data

3.1 Making data findable, including provisions for metadata

The data collected/generated within the project are made available to all the project partners for its duration. The datasets collected as part of the project (referenced in Section 2.1) are also made openly available to the community, in anonymised form whenever personal data is involved.

To maximise dissemination and uptake of project results, we have opted to leverage primarily the AI-for-Europe/AI-on-Demand (AI4EU/AIoD) platform for the purpose of dataset, software and other research outcomes, opting-out from the previous choice of using SoBigData (see D9.2, Section 3).

AIoD is a European community-driven platform, initially developed as part of the AI4EU project (ICT26-2019), and designed to empower European research and innovation in AI by providing and open and easily accessible tool to facilitate knowledge sharing, research experimentation and development of state-of-the art solutions and technologies related with Artificial Intelligence.

AIoD provides a broad set of services, data, methods, training materials and technologies for Artificial Intelligence. It allows GDPR-compliant data sharing and facilitates interoperability by providing standardized data formats and secure access controls, integrated with EU-Login Single Sign On (SSO). AIoD integrates data, code and experiment management tools in the Virtual Lab tool, which facilitates (re)use of datasets and models shared on the platform and simplifying AI model development while maintaining full compliance with European data governance regulations. Virtual lab users have access to datasets as well as other types of resources through a searchable web-based catalogue: resources can then be saved and organized in a personal virtual lab space accessible from anywhere through the SSO. Figure 1 shows a snapshot the my Assets functionality in Virtual Lab.

EMERGE data will be complemented with metadata needed for AIoD indexing, following the documentation in https://api.aiod.eu/redoc. Whenever possible/meaningful the data will be complemented with performance results from baseline methods (e.g. from AI models developed in the EMERGE). To foster reproducibility and simplify data use, we will also release and associate source codes for replicating the baseline results within the AIoD platform.

The motivations that have led us to favour AIoD over SoBigData include the fact that AIoD makes available all the features that we intended to leverage in SoBigData, but in addition to that it offers a (prospectively) better integration with the AI community, as simplified access to knowledge and data, and the possibility to have a single archival point for the EMERGE project after its conclusion. A concrete example of this feature is the Project catalogue (see Figure 2) that will allow us to mirror the content of the current EMERGE website into a long-term storage, associated with all the assets (data, code, workflows) produced by the project (see an example in Figure 3).

In order to ensure the most effective diffusion of the data, datasets may also be released in other sharing platforms, for instance taking into consideration aspects such as:

1. the interests of the community revolving around the platform, its popularity, adequacy of the access interface to the specificity of the data.
2. availability of a scientific publication associated with the dataset, if the publisher offers an associated data sharing platform compliant with FAIR and FACT policies.
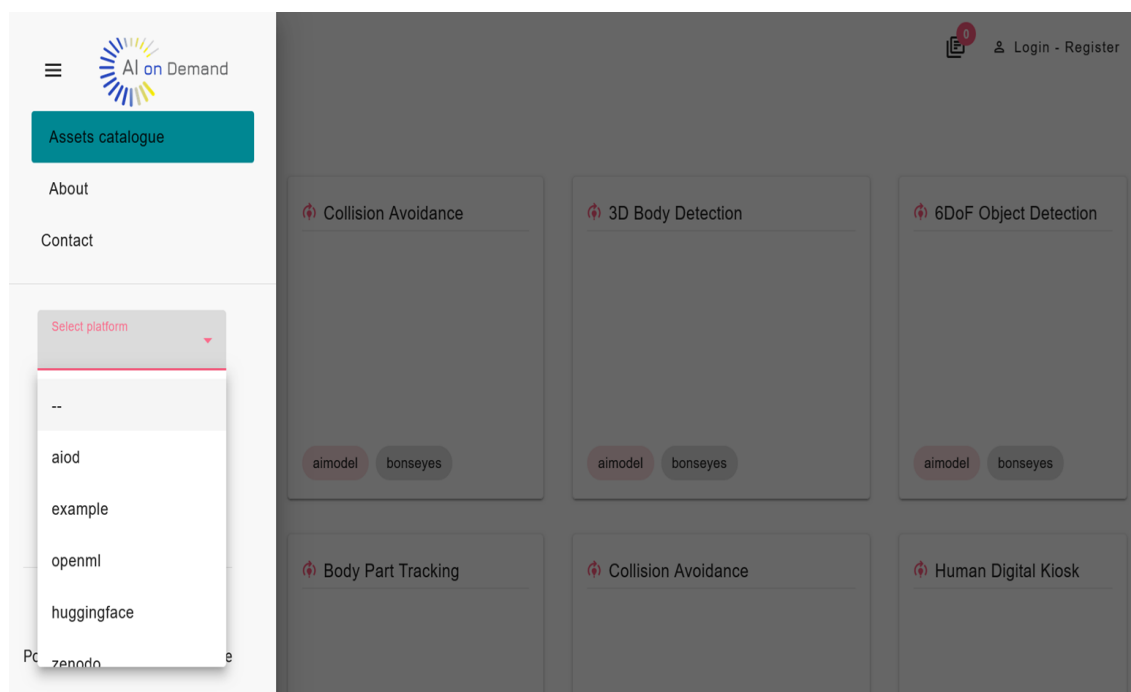
Figure 1 – Assets catalogue on the AIoD platform: assets can be organized and saved into personalized libraries for experimentation.

14

Figure 2 – Searchable catalogue of EU project on the AIoD platform.

Figure 3 – Example of an entry of the Project catalogue for the AI-Sprint project. The figure shows both the general project information as well as the associated AI assets produced.

3.2 Making data interoperable

Released data will be in standard formats widely adopted by the Artificial Intelligence community, such as formatted text readable by standard Python Pandas API methods, allowing easy code-based access, use and combination with other data sources. As a general principle data will be openly shared as soon as possible whenever associated with scientific publications making use of it. In a first phase this may entail releasing the datasets outside of AIoD. However, we will make sure that all data and assets produced by EMERGE will be on the AIoD platform within the end of the project.

The Virtual Lab functionality in AIoD makes data interoperable and employable by other services. It offers an online virtual research environment, supporting secure and controlled data storage and sharing. Each virtual research environment has an associated workspace, partitioned in private and public areas, where users can store, access, and share documents and results related to their activities. The AIoD platform allows to select and execute available methods on the datasets: this is performed within the workspace of the user and can be made publicly available.

3.3 Increase data re-use

We expect most of EMERGE project data to be re-usable for a long time, given the simulation-based nature of several datasets. The choice of a consolidated sharing platform such as AIoD is also intended to ensure persistence of shared data across time. Datasets in AIoD include a unique reference and an assessment of their nature, scale, and availability (such as related scientific publications, privacy issues, data governance policies, licensing, or similar resources). The preservation and re-using procedures bundled in the platform describe how data is stored, which technology is used, and for how long the data is available.

EMERGE data will be made available together with accompanying code that will allow, for instance, to run pre-processing and data cleaning, to replicate experiments with benchmarking models and EMERGE methodologies, and to generate further data (in case of simulated datasets). The AIoD infrastructure will be a key enabler for such an increased data reuse, as it provides data, methods and computational infrastructures made accessible to the user via an integrated interface through the Workspace service.

4. Other research outputs

The research outcomes that EMERGE will make available to the community are not limited to data alone.

One of the key outputs of WP2 is the release of an ethical toolkit including both templates and codes to generate (a) text and graphic vignettes, (b) game theoretic scenarios, (c) surveys, (d) visualisation of dimensions of awareness for different systems, and (d) visualisation of

distribution and simple statistics regarding ethical responses. The ethical toolkit will therefore be a mixture of workflows, code and training material whose nature matches very well breadth of content allowed by the AIoD under the AI Asset category. Additionally, the focus of the AIoD platform and its community is heavily oriented towards social, ethical and regulatory implications of AI. We expect that such matching interests will promote the impact and favour the uptake of the EMERGE ethical toolkit.

A main research output for the technical WPs in EMERGE is the release of open software, including all software and codes related to methodologies presented in academic papers, robotic simulators, and all the code needed to replicate EMERGE published experiments. Also, UOB follows an open-hardware approach for robots and interfaces developed as part of EMERGE.

The released software and code will be made available in the AIoD Virtual Lab as EMERGE assets once consolidated. Currently, the software being developed by the project is already publicly available (also for the purposes of reproducibility and open science) through standard platforms, namely Bitbucket for UOB and GitHub for UNIPI/TUD. An example of this is the Archetype Computing System, whose incremental (stable) version are maintained on a dedicated and public project Github (https://github.com/EU-EMERGE/archetype-computing-adaptive-system) along with the accompanying documentation (see Figure 4).
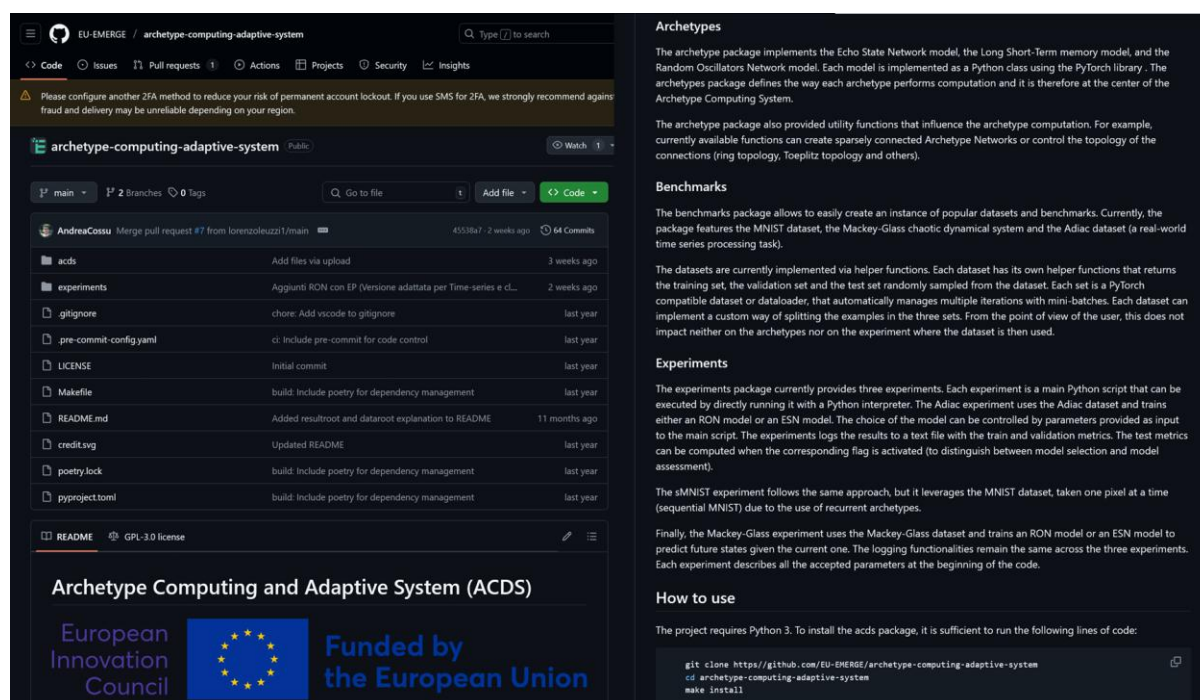

Figure 4 - Github page for the Archetype Computing System

5. Allocation of resources

The costs for data acquisition and generation are covered by the project budget. The data sharing platforms identified in the previous sections are free of charge, hence there is no need to plan additional costs for long term preservation (e.g. following the end of the project). Even when considering data repositories associated with scientific publishers, these are typically free of charge (costs for open publication fees are already foreseen in the project budget).

The Data Management Officer (DMO) for EMERGE is the Coordinator, Davide Bacciu (davide.bacciu@unipi.it). The DMO is responsible for managing application of the data sharing policies and for supporting the partner in charge for the data repository in their implementations. Leading partners for each dataset have been identified in Section 2.1. These partners are primarily responsible for data collection/generation, its secure storage and preservation, for providing access to it to each project partner and, ultimately, for preparing data for sharing on public repositories. Each partner has identified a Data Manager (DM) who will implement the data sharing procedures locally. The list of the DMs along with their contact is provided in the following:

- UNIPI: Davide Bacciu (davide.bacciu@unipi.it)
- TUD: Yasemin Türkyilmaz-van der Velden (datasteward-3mE@tudelft.nl)
- LMU: Nadine Meertens (nadine.meertens@lrz.uni-muenchen.de)
- DVL: Renan Picoreti Nakahara (renan.picoreti@davincilabs.eu)
- UOB: Simon Jones (simon2.jones@bristol.ac.uk)

6. Data security

Project private information (not meant for sharing) collected as part of the moral psychology and swarm user studies will be maintained either on hard-drives (disconnected from the network), or on a server in the local area network of the responsible partner, in encrypted folders or within a Database Management System with an access control management. Access to these data will be granted upon agreement with the DPO and the data maintainer partner. Further details on personal data protection measures are provided in the dedicated "D2.1 Ethics Monitoring" deliverable.

7. Ethics

Project data are used to train and validate machine learning models. As such both data and the trained machine learning models adhere to the ethical guidelines of the HLEG on Artificial Intelligence. These aspects were discussed and articulated in a dedicated deliverable D2.1, as indicated in the Description of the Action document. All datasets in the project are either synthetically generated or collected in experiments involving only volunteers. In the latter case, a detailed statement of consent is collected, and data is duly anonymised.

EMERGE

8. Other issues

We primarily follow the EU General Data Protection Regulation (GDPR). Additional procedures and provisions for the specific partners are detailed below.

LMU: The behavioural and experimental ethics studies performed by LMU are pre-evaluated and approved by the ethics committee of either the Faculty of Philosophy, Philosophy of Science and Religious Studies, or the Psychology faculty at Ludwig-Maximilians-Universität München (LMU Munich).

TUD: Activities are carried in compliance with the TU Delft Research Data Framework Policy stating that research data, code and any other materials needed to reproduce research findings are appropriately documented and shared in a research data repository in accordance with the FAIR principles (Findable, Accessible, Interoperable and Reusable) for at least 10 years from the end of the research project, unless there are valid reasons not to do so.

UOB: The University of Bristol provides a Research Data Service as part of the Library and supports pan-University research activities to help realise the impact of research on society. They help individual researchers or major interdisciplinary research initiatives to plan, manage and share research data. More information can be found here: http://www.bristol.ac.uk/staff/researchers/data/.

DVL: The project maintains a mailing list of email addresses (and, in some cases, names) of people participating in the events organised by the Consortium. This mailing lists are collected upon registration to the event, provided the registrant authorises the inclusion in the list. This mailing list is held on the web-based email marketing service GetResponse. The mailing list is administered by and under the responsibility of DVL. Project members cannot view any details about subscribers. The subscriber list contains personally identifiable information (e-mail addresses and potential names) and is subject to GDPR. A privacy statement will be accessible, included (or linked) in each message posted to the list. The list will be used only to send communications about upcoming events organised by the consortium.