HumanE AI Net:

The HumanE AI Network

Grant Agreement Number: 952026
Project Acronym: HumanE AI Net

Project Dates: 2020-09-01 to 2024-08-31
Project Duration: 48 months

D5.2 Second period microproject results on societal, ethical and responsible AI deposited for general use on the AI4EU platform

Author(s): Nina Khairova, Andrea Aler Tubella, Virginia Dignum
Contributing partners: UMU
Date: May 31, 2023
Approved by: John Shawe-Taylor
Type: Report (R)
Status: Draft
Contact: virginia@cs.umu.se

Dissemination Level		
PU	Public	x

DISCLAIMER

DOCUMENT INFO

0.1. Authors

Authors	Institution	E-mail
Nina Khairova	UMU	nina.khairova@umu.se
Andrea Aler Tubella	UMU	andrea.aler@umu.se
Virginia Dignum	UMU	virginia@cs.umu.se

0.2. Document History

Revision		
Date	**Lead Author(s)**	**Comments**
02.04.2023	AAT	Section Outline
17.04.2023	NK	Microproject collection, writing of descriptions
19.04.2023	AAT	Overall summaries of contributions to WPs 4, 5.

20.04.2023	NK	Update microproject descriptions and results
21.04.2023	NK	Update the list of microprojects, their descriptions and results
24.04.2023	AAT, NK	Update the list of microprojects, overall summary WP6, and adding executive summary section 9.
14.05.2023	VD	Included section on New Microproject Procedures from January 2023
29.05.25	NK	Update the list of microprojects

Table of Contents

1. Introduction 5
1.1. Purpose of this document 5
1.2. Societal, legal and ethical AI objectives 5
2. Microprojects included in the analysis 6
2.1. Contributions of the microprojects 8
3. WP4-related results: Social AI 11
3.1. Overall summary 11
3.2. Microproject descriptions 13
4. WP5-related results: AI Ethics and Responsible AI 19
4.1. Overall summary: 19
4.2. Microproject descriptions 20
5. WP6-related results: Applied research with industrial and societal use cases 26
5.1. Overall summary 26
5.2. Microproject descriptions 28
6. WP7-related results 31
6.1. Overall summary 31
6.2. Microproject descriptions 32
7. WP8-related results 33
7.1. Overall summary 33
7.2. Microproject descriptions 33
8. New Microproject Procedures from January 2023 33
Measuring, modelling, predicting the individual and collective effects of different forms of AI influence in socio-technical systems at scale. (WP4 motivated) 34
ELS evaluation projects (WP5 motivated) 35
Creation/Augmentation of realistic Datasets (WP 6 motivated) 35
Innovation projects (WP6&7 motivated) 36
Education & training projects (WP8 motivated) 36
9. The Legal Protection Debt of Training-Datasets – Executive summary 37

1. Introduction

1.1. Purpose of this document

This deliverable reports on the second period results of microprojects (MPs) related to the societal, ethical and responsible AI objectives of HumanE AI Net. It includes all microprojects with reported results between September 2021 and April 2023 with results relevant to WPs 4,5,6,7 and 8. Earlier reported results can be found in Deliverable 4.1[1].

First, the document presents an overview of the microprojects whose results are reported on, as well as concrete outputs in terms of publications, datasets, outreach, etc.

Following that, there are separate sections for each of the main work packages (WPs) focused on societal, ethical and responsible AI. These include work packages 4,5,6,7 and 8. These sections cover the MPs that contribute to each WP. Following an overview of the MP activity in that WP, a description of each individual MP is included.

Additionally, this deliverable includes the results of VUB's engagement with a MP of WP2, about the curation and augmentation of user-activity training data. The report concerns a legal analysis of collection, curation and augmentation of training data, based on the GDPR and the upcoming EU legal framework. It conducts the analysis under the heading of 'legal protection debt', tracing the debt in terms of legal protection that builds while constructing such data bases. This analysis includes a detailed assessment of the research exemption under the GDPR.

1.2. Societal, legal and ethical AI objectives

The societal, legal and ethical (ELS) objectives can be made concrete along two lines:

1. ELS by design: tools for development, assessment and evaluation of systems along ELS requirements. This includes elements such as robustness, explainability, fairness assessments or evaluations of social impact.
2. ELS for design: tools and methods to guide the responsible design process of intelligent systems. This includes work such as guidelines for developers, surveys or guidance in the form comparisons of different explainability/fairness tools and their applicability for different needs.

The MPs reported in this deliverable address these areas in different ways, which we showcase in a few examples.

In terms of ELS by design, diverse MPs involve assessment of intelligent systems along ELS factors. For example, the MP "Trustworthy Voting Advice Applications" consists of evaluating the implementation and adherence of voting advice applications to the Ethics Guidelines for Trustworthy Artificial Intelligence, developing concrete questions for the assessment.

The MP "Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems" focuses on voice dialog systems. By providing tools to improve context exchange among component and dialog sides (AI agent and human), this MP

[1] https://www.humane-ai.eu/wp-content/uploads/2021/10/Deliverable-4.1.pdf

includes transparency and robustness aspects. Focusing on XAI techniques, the MP “XAI model for human readable data aimed at connected car crash detection” aims at the development of an XAI system to address connected car crash detection.

For ELS for design, several MPs consider modeling and simulation tools for designers to assess the societal impact of different design decisions. By addressing prediction problems and focusing on their application in different human processes in social media sharing, healthcare, finance, disaster management, public security, and daily life, the project “Human Behavior is a matter of Time! – Modeling Events Interactions through Temporal Processes” provides tools for understanding the societal impact of AI systems when integrated in complex human networks. Likewise, the MP “Polarization with the Friedkin-Johnsen model over a dynamic social network” can have applications in the design of social networks, as it studies the effect different configurations can have on polarization.

Taking a high-level perspective of design, the MP “What idea of AI? Social and public perception of AI” addresses the social narratives surrounding AI, proposing the involvement of sociologists and tools from sociology at the design stage.

Concerning legal aspects, the work reported in Section 8 provides and essential analysis of the GDPR to be considered by designers and developers.

Overall, the MPs reported in this deliverable showcase work across a wide spectrum of challenges, areas of AI and applications, keeping a focus on ELS aspects.

2. Microprojects included in the analysis

ID, WP(s)	Project title	Partners	Date start	Date finish
1, WP3 WP5	Human-machine collaboration for content analysis in context of Ukrainian war	UMU, TILDE, CNR, UNIBO	2022-06-01	2023-02-27
2, WP4	Human Behavior is a matter of Time! – Modeling Events Interactions through Temporal Processes	CNR, INESC TEC, Università di Pisa	2022-09-01	2022-12-31
3, WP4 WP2	Decentralized AI on simple social networks	CNR, CEU	2022-12-01	2023-03-31
4, WP7 WP8	Matching the right people! Creating a functional demonstrator for online matching of people and expertise for innovation	ETH Zurich AI Center, European Laboratory for Learning and Intelligent Systems (ELLIS), LMU Munich	2022-06-01	2022-12-31

5, WP2 WP3 WP5	Multilingual and Multimodal conversational agent combined with search engine models. Humane.AI a framework building	THALES SIX, LISN, CNRS, UNIBO	2022-09-01	2023-04-28
6, WP2 WP6	Neural Mechanism in Human Brain Activity During Weight Lifting	TUBITAK BILGEM, DFKI Kaiserslautern	2021-05-03	2022-01-31
7, WP2 WP4 WP6	Polarization with the Friedkin-Johnsen model over a dynamic social network	CNR, Central European University (CEU)	2022-04-01	2022-12-31
8, WP5 WP6	Trustworthy Voting Advice Applications	ETHZ, UMU, TU Delft, University of Amsterdam	2023-01-01	2023-06-30
9, WP2 WP3 WP5	Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems	Brno University of Technology, Charles University	2022-11-01	2023-02-28
10, WP6	The temporal and biological factors of our vulnerability to disinformation	ETHZ, FBK	2022-04-01	2022-04-01
11, WP6	Telefonica 2– validating the air quality prototype in a real city	TID City Council of Valladolid	2022-01-03	
12, WP6	Telefonica 3– assessing the ethical and societal impact of the air quality system	TID City of Madrid City of Valladolid National statistics office of Spain	2022-01-06	
13, WP4	What idea of AI? Social and public perception of AI	UNIBO, UMU, Consiglio Nazionale delle Ricerche	2021-02-01	2021-10-01
14, WP4	Social AI gossiping	Consiglio Nazionale delle ricerche, Central European University	2021-05-17	2021-11-17
15, WP4	Agent based modeling of the Human-AI ecosystem	UNIBO, Central European University	2021-07-01	2022-04-01
16, WP4	Pluralistic Recommendation in News	UNIPI, Consiglio Nazionale delle Ricerche	2021-04-01	2021-12-01
17, WP4	Algorithmic bias and media effects	Consiglio Nazionale delle Ricerche, UNIPI, Central European University	2021-07-01	2021-11-01

18, WP6	Multi-layer evaluation sets for speech translation of web-based meetings	Tilde Charles University	2022-06-01	2022-12-31
19, WP6, WP5	XAI model for human readable data aimed at connected car crash detection	Generali Italia (Industrial Partner), CNR Pisa, Università di Pisa	2021-10-14	2022-02-23
20, WP6, WP8	X5LEARN: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site interface for access to openly licensed educational materials	Knowledge 4 All Foundation, University College London Institut "Jožef Stefan"	2020-12-14	2023-04-20

2.1. Contributions of the microprojects

ID	Publications	Other contributions
1	Working on a conference and journal papers.	Github repository of datasets and software: https://github.com/fablos/ruwa-dataset
2	publication on arXiv "Modeling Events and Interactions through Temporal Processes - A Survey" https://arxiv.org/abs/2303.06067	https://github.com/Angielica/temporal_processes: A list of Point Processes resources. https://github.com/Angielica/datasets_point_processes: A list of relevant datasets.
3	None yet (the micro-project is still ongoing)	All the strategies are implemented in the SAI Simulator (SAISim). The repository is hosted on GitHub. However, it has not been publicly released since the micro project is still ongoing.
4	None yet, we plan on publishing our results later.	Functional prototype with real-world data from Github: http://143.42.16.26:3000/
5	-	Jira & Confluence : https://humane-dialog.atlassian.net/jira/software/projects/HAD/pages Github [PRIVATE]: - project : https://gitlab.com/humane-ai-chatbot/chatbot-fmk - submodules: - T-KEIR: https://github.com/ThalesGroup/t-keir.git - erc-unibo-module: https://github.com/helemanc/erc-unibo-module

6	Conference proceeding: "Prediction of Lifted Weight Category Using EEG Equipped Headgear", published in 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics.	Paper: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9926744 Dataset: https://www.ai4europe.eu/research/research-bundles/neural-mechanism-human-brain-activity-during-weight-lifting?category=ai_assets
7	No publications yet. The collaboration is still ongoing.	Github link of the code of the simulator for the new dynamic model: https://github.com/elisabettabiondi/FJ_rewiring_basic.git
8	-	Video: https://www.youtube.com/watch?v=5riTfDuRTlk
9	-	(WIP) Code for audio data collection: https://github.com/oplatek/speechwoz (WIP) Code for end-to-end response generation: https://github.com/knalin55/augpt (WIP) Report for end-to-end response generation: https://docs.google.com/document/d/1iQB1YWr3wMO8aEB08BUYBqiLh0KreYjyO4EHnb395Bo/edit Workshop links: https://www.clsp.jhu.edu/2023-jelinek-summer-workshop, https://jsalt2023.univ-lemans.fr/en/index.html Workshop proposal: https://docs.google.com/document/d/19PAOkquQY6wnPx_wUXIx2EaInYchoCRn/edit?usp=sharing&ouid=105764332572733066001&rtpof=true&sd=true Workshop team (not finalized yet): https://docs.google.com/spreadsheets/d/1EsHZ-_OREkvf8ODiN7759OYHqSb6MBAX/edit?usp=sharing&ouid=105764332572733066001&rtpof=true&sd=true
10	-	Accepted poster presentation at the International Conference on Computational Social Science (https://www.ic2s2.org/, July 17-20 2023)
11	-	A report with the results of the evaluation of the prototype. Adaptations to the system based on feedback from the local city. Dissemination activities jointly by Telefonica and the local city. Potential policy measures to improve the air quality in Valladolid.
13	- "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control, in Ethics and Information technology". L. Sartori, Laura, A. Theodorou. Published in Ethics and Information Technology 24.1 (2022) https://link.springer.com/article/10.1007/s10676-022-09624-3	-

	- "Minding the gap(s): public perceptions of AI and socio-technical imaginaries". L. Sartori, G. Bocca. Published in AI & SOCIETY (2022) https://link.springer.com/article/10.1007/s00146-022-01422-1	
14	In preparation	Code: SAIsim, C. Boldrini, L. Valerio, A. Passarella, https://zenodo.org/record/5780042#.Ybi2sX3MLPw
15	Human-AI ecosystem with abrupt changes as a function of the composition. Contucci P, Kertész J, Osabutey G (2022) PLOS ONE 17(5): e0267310 https://doi.org/10.1371/journal.pone.0267310	-
16	A data-paper in preparation	Dataset: European News with political bias with metadata (around 16 million articles) and European News with political bias with metadata plus topic annotation for each article (around 4 million articles)
17	"Mass Media Impact on Opinion Evolution in Biased Digital Environments: a Bounded Confidence Model". V. Pansanella, A. Sirbu, J. Kertez, G.Rossetti. Submitted to Scientific Report https://doi.org/10.21203/rs.3.rs-2662381/v1	Code: https://github.com/GiulioRossetti/AlgorithmicBias
18	-	Speech translation evaluation and development data sets for English->Latvian (4 hours and 40 minutes), Latvian->English (4 hours and 52 minutes), and Lithuanian->English (3 hours and 31 minutes) Test sets for AutoMin 2023: for Task A: 10 English + 10 Czech meeting transcripts with manually created reference minutes; for Task D: manually created alignments between AutoMin 2021 system outputs and transcripts. WMT translation evaluations (en->cs) including the relevant domain
19	Explaining Crash-Predictions on Multivariate Time Series Data https://link.springer.com/chapter/10.1007/978-3-031-18840-4_39 . The	-

	Lecture Notes in Computer Science book series (LNAI,volume 13601)	
20	Watch Less and Uncover More: Could Navigation Tools Help Users Search and Explore Videos? In Proceedings of the 2022 Conference on Human Information Interaction and Retrieval (CHIIR '22). https://doi.org/10.1145/3498366.3505814 X5Learn: A Personalised Learning Companion at the Intersection of AI and HCI. In 26th International Conference on Intelligent User Interfaces - Companion. https://doi.org/10.1145/3397482.3450721 Power to the Learner: Towards Human-Intuitive and Integrative Recommendations with Open Educational Resources. Sustainability 14, 18: 11682. https://doi.org/10.3390/su141811682 Bulathwela, S., Pérez-Ortiz, M., Holloway, C., & Shawe-Taylor, J. (2021). Could ai democratise education? socio-technical imaginaries of an edtech revolution. arXiv preprint arXiv:2112.02034.	Videos: X5Learn Demo Video: https://youtu.be/aXGL05kbzyg Workshop presentation (AAAI'21): https://www.youtube.com/watch?v=gYtmL2XdxHg Video: https://youtu.be/E11YUWad7Lw Workshop Presentation (AAAI'21): https://youtu.be/4v-fizLvHwA X5Learn Platform: https://x5learn.org/ Code: TrueLearn Codebase: https://github.com/sahanbull/TrueLearn TrueLearn Python library: https://truelearn.readthedocs.io

Additionally, this deliverable includes the results of VUB's engagement with a MP of WP2, about the curation and augmentation of user-activity training data. This engagement has resulted in datasets creation, curation and dissemination in the scope of the MP in WP2, as well as a report concerning a legal analysis of collection, curation and augmentation of training data, based on the GDPR and the upcoming EU legal framework.

3. WP4-related results: Social AI

3.1. Overall summary

The work package "Social AI" addresses the societal dimension of AI, with a focus on the complex interactions between humans and AI agents. Examples range from urban mobility, with travelers helped by smart assistants to fulfill their agendas, to the public discourse and the markets, where diffusion of opinions as well as economic and financial decisions are

shaped by personalized recommendation systems. The areas comprised in this WP consist of:

1. graybox models of society-scale, networked hybrid human-AI systems;
2. individual vs. collective goals of social AI systems;
3. societal impact of AI systems;
4. self-organized, socially distributed information processing in AI-based techno-social systems.

Over the second period, the MPs related to WP4 are:

- “Decentralized AI on simple social networks”
- “Human Behavior is a matter of Time! – Modeling Events Interactions through Temporal Processes”,
- “Polarization with the Friedkin-Johnsen model over a dynamic social network”
- “What idea of AI? Social and public perception of AI”
- “Social AI gossiping”
- “Agent based modeling of the Human-AI ecosystem”
- “Pluralistic Recommendation in News”
- “Algorithmic bias and media effects”

Many projects **engage with social networks**, either through modeling the networks themselves or through modeling processes with direct effects on such networks, thus addressing the first task.

For example, the MP “Decentralized AI on simple social networks” studied the effect of simple social structures on the decentralized learning process in a human-AI ecosystem and how the lack of coordination impacts the resulting learned model. This goal addresses areas 2 and 4, investigating how decentralized learning can function in different configurations and taking into account human factors.

Similarly, the MP “Social AI gossiping” aims to model “gossiping” for accomplishing a decentralized learning task and to study what models emerge from the combination of local models, where such a combination takes into account the social relationships between the humans associated with the AI. This idea engages with areas 1 and 4

By addressing prediction problems and focusing on their application in different human processes in social media sharing, healthcare, finance, disaster management, public security, and daily life, the project “Human Behavior is a matter of Time! – Modeling Events Interactions through Temporal Processes” addresses the third and fourth areas. Indeed, process modeling can be key in understanding and assessing the societal impact of AI systems when integrated in complex human networks.

The MP “Agent based modeling of the Human-AI ecosystem” aims at investigating systems composed of a large number of agents belonging to either human or artificial types, clearly addressing area 1.

Both within models of social networks and empirically, the **idea of studying polarization** is also frequent.

The MP “Polarization with the Friedkin-Johnsen model over a dynamic social network” studies a model of polarization and opinion dynamics over different social network

typologies, thus addressing areas 1 and 3 in terms of modeling hybrid networks and studying social impact.

The MP “Algorithmic bias and media effects” contributes a model of opinion dynamics where algorithmic bias is introduced, with interaction more frequent between similar individuals, simulating the online social network environment. The objective of this microproject is to enhance this model by adding the biased interaction with media in an effort to understand whether this facilitates polarisation. This idea follows areas 1 and 3, modeling society-scale netwoerks and studying societal impacts of media.

The MP “Pluralistic Recommendation in News” develops two datasets for European News with political leaning labeling, addressing goal 3 of providing tools for analysis of societal impact.

The MP “What idea of AI? Social and public perception of AI” addresses the social narratives surrounding AI, performing political-bias classification by filtering out all the articles which do not carry political-bias, such as those dealing with sports or gossip. This MP directly addresses area 3.

3.2. Microproject descriptions

Human Behavior is a matter of Time! – Modeling Events Interactions through Temporal Processes

Date Start: 2022-09-01
Date finish: 2022-12-31
Duration: 4 months
Partners:

1. Consiglio Nazionale delle Ricerche (CNR), Giuseppe Manco (2 PM)
2. INESC TEC (2 PM)
3. Università di Pisa (UNIPI) (1 PM)

Description:

This Micro project will investigate methods for learning deep probabilistic models based on latent representations that can explain and predict event evolution within social media. Latent variables are particularly promising in situations where the level of uncertainty is high, due to their capabilities in modeling the hidden causal relationships that characterize data and ultimately guarantee robustness and trustworthiness in decisions. In addition, probabilistic models can efficiently support simulation, data generation and different forms of collaborative human-machine reasoning.

Main results of micro project:
Our microproject aims at investigating methods for modeling event interactions through temporal processes. We revisited the notion of event modeling and provided the mathematical foundations that characterize the literature on the topic. We defined an ontology to categorize the existing approaches in terms of three families: simple, marked, and spatio-temporal point processes. For each family, we systematically reviewed the existing approaches providing a deep discussion. Specifically, we investigated recent

machine and deep learning-based methods for modeling temporal processes. We focused on studying prediction problems from event sequences to understand their structural and temporal dynamics. In fact, understanding these dynamics can provide insights into the complex patterns that govern the process and can be used to forecast future events. Among existing approaches, we investigated probabilistic models based on latent representations that represent an appropriate choice to model event sequences. Event sequences are pervasive in several application contexts, such as business processes, smart industry as well as scenarios involving human activities, including especially information diffusion in social media. Indeed, our study has been focused on works whose aim is the prediction of events within social media. Social media focus on the interactions among individuals within context-sharing platforms such as Twitter, Instagram, etc. Interactions can be modeled as event sequences since events can be user actions over time. In addition, we also provided an overview of other application scenarios such as healthcare, finance, disaster management, public security, and daily life. The analyzed literature provides several datasets that we categorized according to the application scenarios they can be used for. For each dataset, we reported its description, the papers containing experiments over it, and, when available, a source web link.

Decentralized AI on simple social networks

Date Start: 2022-12-01
Date finish: 2023-03-31
Duration: 4 months
Partners:

1. Consiglio Nazionale delle ricerche (CNR), Lorenzo Valerio, lorenzo.valerio@iit.cnr.it (3 PM)
2. Central European University, János Kertész, kerteszj@ceu.edu (3 PM)

Main results of micro project:

In this microproject, we set out to study the effect of simple social structures on the decentralized learning process in a human-AI ecosystem and how the lack of coordination impacts the resulting learned model. We have considered the following learning policies: federated learning (FedAvg), average-based decentralized learning (called DecAvg, i.e., an adaptation of FedAvg to the decentralised settings), difference-based decentralized learning (we designed a novel strategy called DecDiff), and KD-based decentralized learning (with a virtual teacher). For decentralized strategies, we considered both homogeneous and heterogeneous initial conditions (e.g., common initialization of models, IID and non-IID data distribution among nodes). The common benchmark is centralized learning (i.e., we assume that all users upload their data to a central server). From the social network standpoint, we initially focused on dyadic and triadic social networks, then moved on to richer topologies like erdős-rényi graphs and SBM graphs. As a learning task, we considered a standard classification problem on the MNIST dataset. Other and more challenging datasets are currently under investigation. We have so far observed the following.

1. In small networks where data availability is not an issue, DecAvg in model-homogeneous settings (i.e., all the users' AI models are commonly initialised) is as good as federated learning using FedAvg (despite the lack of a central controller). Without the common initialization (i.e., all the AI models are independently and randomly initialised), the accuracy strongly depends on the initial conditions. DecDiff definitely mitigates the problem, yielding

a higher accuracy despite being slower in the transient phase. The virtual teacher clearly outperforms a basic centralized approach. Instead, when data is a bottleneck, the learning strategy plays a limited role in the observed performance.
2. In larger networks, DecDiff doesn't suffer from the initial disruption caused by the averaging process that DecAvg suffers from. However, at steady state, it is not better than the simpler DecAvg. When the data distribution is extremely uneven, DecDiff seems to provide more reliable performance. Interestingly, KD-based decentralized learning always performs well, surpassing standard federated learning.

Polarization with the Friedkin-Johnsen model over a dynamic social network

Date Start: 2022-04-01
Date finish: 2022-12-31
Duration: 4 months
Partners:

1. Consiglio Nazionale delle Ricerche (CNR), Elisabetta Biondi, elisabetta.biondi@iit.cnr.it
2. Central European University (CEU), Janos Kertesz, kerteszj@ceu.edu, Gerardo Iniguez, IniguezG@ceu.edu

Description:

The Friedkin-Johnsen model is a very popular model in opinion dynamics, validated on real groups, and well-investigated from the opinion polarization standpoint. Previous research has focused almost exclusively on static networks, where links between nodes do not evolve over time. In this micro-project, we want to fill this gap by designing a variant of the Friedkin-Johnsen model that embeds the dynamicity of social networks. Furthermore, we will design a novel definition of global polarization that combines network features and opinion distribution, to capture the existence of clustered opinions. We have analyzed the polarization effect of the new dynamic model, and identify the impact of the network structure.

Results:

Human social networks are very complex systems and their structure has an essential impact on opinion dynamics. However, since my main goal is to study the impact of the opinion dynamics model per se, we decided to deal with two different social network typologies: an Erdős–Rényi (ER) and a stochastic block model (SBM).

Design of the Friedkin-Johnsen (FJ) dynamic model.

We have implemented a rewiring policy that has been extensively studied in discrete opinion diffusion models. This involves substituting edges that connect nodes with different opinions with other edges. We have adapted this scheme to work with the FJ model's opinions, which are within the range of [-1,1], in both the asynchronous and synchronous versions. According to two parameters θ (the disagreement threshold) and p_rew (the rewiring probability):

- With probability 1-p_rew the FJ is applied
- With probability p_rew, if i and j disagree, i.e. $|x_i - x_j| > \theta$, the edge (i,j) is replaced with an edge (i,k) where k agrees with i, i.e. $|x_i - x_j| <= \theta$.

The above algorithm was specifically designed and implemented for the ER graph. However, in the case of the SBM, I have limited the potential candidates for rewiring to nodes within a maximum of two hops distance. This decision was made to prevent the block structure from becoming entirely irrelevant. The rationale behind this choice is based on the triadic closure mechanism, which suggests that individuals are more inclined to choose new acquaintances among the friends of their friends.

Design of the polarization metric.

The design of the polarization metric involved developing a definition for identifying highly polarized networks. We defined a highly polarized network as one in which there are two distinct opinions that are clustered into two tightly connected communities. To achieve this, we needed to consider both the network structure and the distribution of opinions. Therefore, we decided to use two different metrics to measure these aspects: bimodality for the opinion distribution and homogeneity for its correspondence with the network structure.

Bimodality.

The bimodality coefficient was used to measure the extent to which a distribution is bimodal. It is calculated using the skewness and kurtosis values and represents how similar the distribution is to one with two modes.

Homogeneity

To measure the homogeneity of the opinion distribution with the network structure, we examined the local distribution of nodes' opinions. We looked at whether each node's opinion was similar to those of its neighbors, which would suggest that it was in line with the overall opinion distribution over the network. The final homogeneity value was close to zero if the distribution of opinions was close to linear.

Experimental evaluation.

We have developed a Python simulator that can compute the dynamic FJ (rewiring included), and polarization metrics over time based on the given network and initial opinions. To test the model, we ran simulations on a small network comprising 20 nodes and compared the outcomes of the FJ with rewiring to those without rewiring. For the ER network, we used a vector of uniformly distributed opinions over [-1,1] as the initial opinions. However, for the SBM networks, we employed a different configuration, where the initial opinions were uniformly extracted over the intervals [-0.5,0-0.1] and [0.1,0.5], depending on whether the nodes belonged to one or the other block.

In conclusion, this microproject involves the design of a dynamic version of the FJ model for synchronous and asynchronous cases. Additionally, we have developed a new definition of polarization that considers both the distribution of opinions and the network topology. To assess the model's effectiveness, we conducted simulations on two different network types: an ER network and an SBM network. Our findings indicate that the rewiring process has significant effects on polarization, but these effects are dependent on the initial network.

What idea of AI? Social and public perception of AI

Date Start: 2021-02-01
Date finish: 2021-10-01
Duration: 7 months

Partners:

1. Università degli studi di Bologna (UNIBO), Laura Sartori, l.sartori@unibo.it
2. Umeå University (UMU)
3. Consiglio Nazionale delle Ricerche (CNR)

This microproject wants to conduct empirical research that explores the social and public attitudes of individuals towards AI and robots. AI and robots will enter many more aspects of our daily life than the average citizen is aware of while they are already organizing specific domains such as work, health, security, politics and manufacturing. Along with technological research it is fundamental to grasp and gauge the social implications of these processes and their acceptance into a wider audience. Some of the research questions are: Do citizens have a positive or negative attitude about the impact of Ai? Will they really trust a driverless car, or will they passively accept a loan or insurance's denial based on an algorithmic decision? Do states alone have the right and expertise to regulate the emerging technology and digital infrastructures? What about technology governance? What are the dominant AI's narratives in the general public?

The immediate plan is to perform political-bias classification exploiting the new dataset by filtering out all the articles which do not carry political-bias, such as those dealing with sports or gossip.

Social AI gossiping

Date Start: 2021-05-17
Date finish: 2021-11-17
Duration: 6 months
Partners:

1. Consiglio Nazionale delle ricerche (CNR), Andrea Passarella, andrea.passarella@iit.cnr.it
2. Central European University, János Kertész, kerteszj@ceu.edu

We envision a human-AI ecosystem in which AI-enabled devices act as proxies of humans and try to learn a model in a decentralized way collectively. Each device will learn a local model that needs to be combined with the models learned by the other nodes, in order to improve both the local and global knowledge. The challenge of doing so in a fully decentralized AI system entails understanding how to compose models coming from heterogeneous sources and, in case of potentially untrustworthy nodes, decide who can be trusted and why. In this micro-project, we focus on the specific scenario of model "gossiping" for accomplishing a decentralized learning task and we study what models emerge from the combination of local models, where combination takes into account the social relationships between the humans associated with the AI. We will use synthetic graphs to represent social relationships, and large-scale simulation for performance evaluation.

Agent based modeling of the Human-AI ecosystem

Date Start: 2021-07-01
Date finish: 2022-04-01

Duration: 9 months
Partners:

1. Università di Bologna UNIBO, Pierluigi Contucci, pierluigi.contucci@unibo.it
2. Central European University (CEU), Janos Kertesz, kerteszj@ceu.edu

Description:

The project aims at investigating systems composed of a large number of agents belonging to either human or artificial types. The plan is to study, both from the static and the dynamical point of view, how such a two-populated system reacts to changes in the parameters, especially in view of possible abrupt transitions. We are planning to pay special attention to higher order interactions like three body effects (H-H-H, H-H-AI, H-AI-AI and AI-AI-AI). We hypothesize that such interactions are crucial for the understanding of complex Human-AI systems. We will analyze the static properties both from the direct and inverse problem perspectives. This study will pave the way for further investigation of the system in its dynamic evolution by means of correlations and temporal motifs.

Pluralistic Recommendation in News

Date Start: 2021-04-01
Date finish: 2021-12-01
Duration: 8 months
Partners:

1. Università di Pisa (UNIPI), Paolo Ferragina, paolo.ferragina@unipi.it
2. Consiglio Nazionale delle Ricerche (CNR), Giulio Rossetti, giulio.rossetti@isti.cnr.it

Description:

The first tangible objective of this micro-project is to develop two datasets for European News with political leaning labeling. The second step is the building a bias-minimizing recommender system for European news.

The first dataset comprehends millions of European news and has been enriched with metadata from Eurotopics.net. Each entry in the dataset contains the main text, title, publishment date, language, news source, and news source metadata. This metadata comprehends the political leaning of the news source and its country. An article bias classifier is built to predict the political label of single articles using the labels obtained through distant supervision. Then explainableAI methods were applied to the classifier and concluded that the classifier is predicting the news source rather than the political leaning. In order to overcome this issue, a second dataset was built, which has the same features as the first one described above, but with the addition of topics chosen between 7 macro-topics.

Algorithmic bias and media effects

Date Start: 2021-07-01
Date finish: 2021-11-01
Duration: 4 months
Partners:

1. Consiglio Nazionale delle Ricerche (CNR), Giulio Rossetti, giulio.rossetti@isti.cnr.it
2. Università di Pisa UNIPI, Dino Pedreschi, dino.pedreschi@unipi.it
3. Central European University (CEU), Janos Kertesz, kerteszj@ceu.edu

Description:

Recent polarisation of opinions in society has triggered a lot of research into the mechanisms involved. Personalized recommender systems embedded into social networks and online media have been hypothesized to contribute to polarisation, through a mechanism known as algorithmic bias. In a recent work, a model of opinion dynamics with algorithmic bias is introduced, where interaction is more frequent between similar individuals, simulating the online social network environment. The objective of this microproject is to enhance this model by adding the biased interaction with media in an effort to understand whether this facilitates polarisation. Media interaction will be modeled as external fields that affect the population of individuals. Furthermore, the case of whether moderate media can be effective in counteracting polarisation was studied.

4. WP5-related results: AI Ethics and Responsible AI

4.1. Overall summary

WP5 is dedicated to ensuring that AI systems operate within an ethical, legal and social frameworks, in verifiable and justified ways. This WP addresses theory and methods for Responsible Design of AI systems. The focus is on the integration of engineering, policy, law and ethics approaches concerning legal, ethical and trustworthy aspects. In addition, technical aspects such as robustness, interaction design, fairness, transparency and accountability and their wider socio-legal interpretations are in the scope of WP5.

The micro-projects contributing to this WP are:

- “Human-machine collaboration for content analysis in context of Ukrainian war”
- “Multilingual and Multimodal conversational agent combined with search engine models.”
- “Trustworthy Voting Advice Applications”
- “Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems.”
- “XAI model for human readable data aimed at connected car crash detection”

The MP “Human-machine collaboration for content analysis in context of Ukrainian war” addresses issues of disinformation, reaching WP5 goals by focusing on methods ensuring responsible design of AI Systems and compliance to ethical, trust, fairness, public perception and societal principles.

The MP “Multilingual and Multimodal conversational agent combined with search engine models” provides a proof of concept of the Framework for collaboration described in the Humane-AI Net revised strategic work plan. The proposed architecture looks to adhere to Responsible AI principles, thus connecting to WP5.

The MP “Trustworthy Voting Advice Applications” relates to the call topic “ELS evaluation projects”. It consists of evaluating the implementation and adherence of voting advice applications to the Ethics Guidelines for Trustworthy Artificial Intelligence, thus participating to the WP5 goal of investigating technical and social factors of AI as related to ethical guidelines.

The MP “Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems” focuses on voice dialog systems. By providing tools to improve context exchange among component and dialog sides (AI agent and human), this MP includes transparency and robustness aspects related to WP5.

Finally, the MP “XAI model for human readable data aimed at connected car crash detection” aims at the development of an XAI system to address connected car crash detection. Improving efficiency in connected car crash detection (reduction of false positives) can reduce the number of car crashes with fatal or severe injury outcome and also improve road safety, thus addressing the goal of utilizing responsible AI methods for socially relevant applications.

In addition to these microprojects, VUB’s engagement with a MP of WP2, about the curation and augmentation of user-activity training data contributes to the tasks of WP5 in terms of the study and development of legal methodologies. This activity is reported on in Section 8 of this deliverable, in the form of an executive summary.

4.2. Microproject descriptions

Human-machine collaboration for content analysis in context of Ukrainian war
Date Start: 2022-06-01
Date finish: 2023-02-27
Duration: 6 months
Partners:

1. Umeå University (UMU), Nina Khairova,nina.khairova@umu.se (6 PM)
2. Consiglio Nazionale delle Ricerche (CNR), Carmela Comito, carmela.comito@icar.cnr.it
3. Università di Bologna (UNIBO), Andrea Galassi, p.torroni@unibo.it

Description:

In this project, which we work with a Ukrainian academic refugee, to combine methods for semantic text similarity with expert human knowledge in a participatory way to develop a training corpus that includes news articles containing information on extremism and terrorism.

Main results of micro project:
1) Collection and curation of two event-based datasets of news about Russian-Ukrainian war.
The datasets support analysis of information alteration among news outlets (agency and media) with a particular focus on Russian, Ukrainian, Western (EU and USA), and international news sources, over the period from February to September 2022.

We manually selected some critical events of the Russian-Ukrainian war. Then, for each event, we created a short list of language-specific keywords. The chosen languages for the keywords are Ukrainian, Russian, and English.
Finally, besides the scraping operation over the selected sources, we also gather articles using an external news intelligence platform, named Event Registry which keeps track of world events and analyzes media in real-time. Using this platform we were able to collect more articles from a larger number of news outlets and expand the dataset with two distinct article sets. The final version of the RUWA Dataset is thus composed of two distinct partitions.

2) Development of an unsupervised methodology to establish whether news from the various parties are similar enough to say they reflect each other or, instead, they are completely divergent and therefore one is likely not trustworthy.
We focused on textual and semantic similarity (sentence embeddings methods such as Sentence-BERT), comparing the news and assess if they have a similar meaning.
Another contribution of the proposed methodology is a comparative analysis of the different media sources in terms of sentiments and emotions, extracting subjective points of view as they are reported in texts, combining a variety of NLP-based AI techniques and sentence embeddings techniques.
Finally, we applied NLP techniques to detect propaganda in news article, relying on self-supervised NLP systems such as RoBERTa and existing adequate propaganda datasets.

Preliminary Qualitative results:
When the events concern civilians all sources are very dissimilar. But Ukraine and Western are more similar. When the event is military targets, Russian and Ukraine sources are very dissimilar from other sources, there is more propaganda in Ukraine and Russian ones.

Multilingual and Multimodal conversational agent combined with search engine models. Humane.AI a framework building

Date Start: 2022-09-01
Date finish: 2023-04-28
Duration: 6
Partners:

1. Thales Eric Blaudez, eric.blaudez@thalesgroup.com
2. Unibo Paolo Torrini, p.torroni@unibo.it
3. LISN Christophe Servan c.servan@qwant.com

Description:
The micro-project provides a demonstration of the hierarchical Framework for collaboration described in the Humane-AI Net revised strategic work plan, by constructing multimodal and multilingual conversational agents focused on search.
The framework is based on hierarchical levels of abilities:

- Reactive (sensori-motor) Interaction: Interaction is tightly-coupled perception-action where actions of one agent are immediately sensed and interpreted as actions of the other. Examples include greetings, polite conversation and emotional mirroring

- Situated (Spatio-temporal) Interaction Interactions are mediated by a shared model of objects and relations (states) and shared models for roles and interaction protocols.

On this microproject we focused on the 2 first levels (Reactive and Situational) and designed the global framework architecture. The results are to be demonstrated in a Proof of Concept (PoC).

Results:

Components development

The development of the components is ongoing. The mini project deals with three modules: the HMI to propose an interaction interface, the reactive module to recognize emotions and the situational module to manage the dialog situation. The framework is developed with the strategy proposed by [1]; in this project, we focus on reactive and situational part.

HMI [Status: DONE]

The HMI is a simple Flask application connected to Framework REST Service.

Reactive module [Status: ON GOING, framework integration almost finished]

The objective of the reactive module was to recognize emotions in conversation, so as to enable the other modules to drive the conversation accordingly. We thus addressed an Emotion Recognition in Conversation (ERC) task. The first step was a literature review and a survey of the available datasets and architectures. We chose the MELD [3] dataset as it is considered a benchmark for ERC and one of the few datasets available for building our ERC module. We considered different options for capturing the context, including the current utterance and its corresponding dialogue history were analyzed. As an architecture, we chose EmoBERTa [1], due to its strong performance and availability of implementation. We trained EmoBERTa using MELD splits for training, validation and test dataset. We carried out several experiments to establish baselines and examine the impact of different context representations. Moreover, since MELD conversations differ from those that may arise in the application domain envisaged for our chatbot, since we do not have any validation datasets available, we explored the generalization capability of the model by transfer learning and few-shot learning. In particular, we considered SetFit [4] as a few-shot learning technique and DailyDialog [2] for transfer learning validation. The experiments conducted on EmoBERTa have yielded promising results, indicating that this architecture is a suitable starting point for constructing the reactive module. However, the task of generalization and adaptability of the model is challenging, and ongoing experiments are being carried out to address these issues. References

Situational Module [Status: ONGOING]

This module is responsible of the “Interactions mediated by shared models of objects (entities) and relations and shared models of roles and interaction protocols”. Knowledge graph are used to represent the situation, the relations and keep the context of the dialog.

- The document Analysis module has been developed in AI4EU project and is adapted and used as submodule for the microproject. The module provides a data processing and a knowledge extraction functions by using Natural Language Processing. It builds a structured representation of the data (knowledge graph, named entities, keywords, summaries …) and provide semantic search. This module is in constant

evolution; it will provide at the end of year a capability to manage more than 100 languages.

- The LISN part aims to create a spoken language understanding (SLU) part and the dialogue manager part. LISN proposed a first proof of concept based Speech recognition on RASA [1] in order to propose a full dialogue framework. For this part, at LISN, we focus on the SLU part of the module. The SLU module is based on the DIET-Classifier model (DIET stands for Dual Intent and Entity Transformer) [2] for both performing intent detection and concept detection. For the next part, we will move to another framework since we observed some limitation in the use of the RASA framework. We will also use another model to perform both intent and concept detection, by using the JointBERT model [3]. Preliminary results are encouraging.
- Thales & LISN works on Situational context representation based on (Temporal)-Knowledge graph and is based on Endsley model [4].

References:

Components Development

[1] Crowley, James & Coutaz, Joëlle & Grosinger, Jasmin & Vázquez-Salceda, Javier & Angulo, Cecilio & Sanfeliu, Alberto & Iocchi, Luca & Cohn, Anthony. (2022). A Hierarchical Framework for Collaborative Artificial Intelligence. 10.48550/arXiv.2212.08659.

Reactive Module (UniBO)

[1] Kim, T., Vossen, P.: Emoberta: Speaker-aware emotion recognition in conversation with Roberta. CoRR abs/2108.12009 (2021)

[2] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. In: IJCNLP(1). pp. 986–995. Asian Federation of Natural Language Processing (2017)

[3] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: ACL (1). pp. 527–536. Association for Computational Linguistics (2019)

[4] Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M., Pereg, O.: Efficient few-shot learning without prompts. CoRR abs/2209.11055 (2022)

Situational Module (Thales & LISN)

[1] Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open source language understanding and dialogue man-agement. In the proceedings of the first NIPS workshop on Conversational AI.

[2] Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. (2020). Diet: Lightweight language understanding for dialogue systems. ArXiv preprint arXiv:2004.09936.

[3] Chen, Q., Zhuo, Z., and Wang, W. (2019). Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909.

[4] M. R. Endsley and D. J. Garland (Eds.). Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum, 2000.

Trustworthy Voting Advice Applications

Date Start: 2023-01-01
Date finish: 2023-06-30
Duration: 6 months
Partners:

1. ETHZ, Elisabeth Stockinger, elisabeth.stockinger@gess.ethz.ch (6 PM)
2. UMU, Virginia Dignum, virginia@cs.umu.se (2.1 PM)
3. TU Delft, Jonne Maas, J.J.C.Maas@tudelft.nl (1 PM)
4. University of Amsterdam (external partner) Christopher Talvitie, christalvitie@gmail.com

Results:

Voting Advice Applications (VAAs) are increasingly popular throughout Europe. While commonly portrayed as impartial tools to measure issue agreement, their developers must take several design decisions at each step of the design process. Such decisions may include the selection of issues to incorporate into a questionnaire, the placement of candidates or parties on a political spectrum, or the algorithm measuring the distance between user and candidate. These decisions have to be made with great care, as they can cause substantial differences in the resulting list of recommendations.

As there is no known ground truth, by which to measure different VAA designs, it is imperative that their design follows guidelines and best practices of pro-ethical design. Similarly, as VAAs aim to directly inform voter decisions in a democratic election, users must be able to trust the fulfillment of these guidelines based on the information available to them.

First, we adapted the Ethics Guidelines for Trustworthy Artificial Intelligence designed by the High-Level Expert Group on Artificial Intelligence appointed by the European Commission to the context of VAAs. The result is a comprehensive set of questions covering the key requirements and ethical principles at the core of these guidelines. These questions focus in particular on transparency, which is one of the key elements of democracies, as well as trustworthy AI. Our findings will be made available in a scientific publication, when the work related to this MP has been finished.

Second, we created a video explaining the project to the general public in the spirit of science dissemination, see https://www.youtube.com/watch?v=5riTfDuRTlk .

Third, we presented the project at the HumanE AI conference.

We are working on an evaluation of VAAs within Europe (Stemwijzer, Wahl-O-Mat, Smartvote, Sanoma, and possibly SVT Nyheter Valkompass) to conclude in a scientific publication.

Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems

Date Start: 2022-11-01
Date finish: 2023-02-28
Duration: 4 months

Partners:

1. Brno University of Technology, Petr Schwarz, schwarzp@fit.vutbr.cz (4 PM)
2. Charles University, Ondrej Dusek, odusek@ufal.mff.cuni.cz (3 PM)

Results:

This project brings us data, tools, and baselines that enable us to study and improve context exchange among component and dialog sides (AI agent and human) in voice dialog systems. A better context exchange allows us to build more accurate automatic speech transcription, better dialog flow modeling, more fluent speech synthesis, and more powerful AI agents. The context exchange can be seen as an interactive grounding in two senses - among dialog sides (for example, technologies like example automatic speech transcription rarely use the other dialog side information to adapt itself) and among dialog system components (the speech synthesis rarely uses dialog context to produce more fluent or expressive speech). The individual project outputs are summarized below:

1) Audio data collection software based on the Twilio platform and WebRTC desktop/mobile device clients. The purpose is to collect audio data of communication between agents (company, service provider, for example, travel info provider) and users. This software enables us to collect very realistic voice dialogs that have high-quality audio (>= 16 kHz sampling frequency) on the agent side and low telephone-quality audio on the user side. The code is available here: https://github.com/oplatek/speechwoz

2) We have established a relationship with Paweł Budzianowski (Poly.AI) and Izhak Shafran (Google). Paweł created the MultiWoz database – an excellent dialog corpus (https://arxiv.org/abs/1810.00278) that we use for the text-based experiment. We decided to collect our audio data similarly. Izhak organized DSTC11 Speech Aware Dialog System Technology Challenge (https://arxiv.org/abs/2212.08704) and created artificial audio data for MultiWOZ through speech synthesis, reading, and paraphrasing. Both provided us with the necessary advice for our data collection.

3) Speech dialog data – the data collection platform preparation and data collection are very time-consuming. The data collection is in progress and will be released before June 26th, 2023.

4) Initial experiments with context exchange between dialog sides (user and agent) were performed. These experiments show a nice improvement in the component of automatic speech recognition side. The results will be re-run with the collected data and published when the collection is finished.

5) Initial experiments with training instance weighting for response generation – which brings context to dialog system response generation, were performed. Experiments were based on the AuGPT system, previously developed at CUNI. The code is available here: https://github.com/knalin55/augpt. Instance weighting increases the re-use of context, compared to normal training, and can go even beyond natural occurrences in data. Simple weighting (threshold) seems better than designing a complex instance weight (in terms of automated metrics, limited manual evaluation is not conclusive). Cross entropy loss works better than unlikelihood loss, where dialogue success may be reduced.

6) There is ongoing work on building a team for JSALT research summer workshop[2]. This is a prestigious workshop organized by John Hopkins University every year. This year it is supported and co-organized by the University of Le Mans. Our topic is the Automatic design of conversational models from observation of human-to-human conversation (https://jsalt2023.univ-lemans.fr/en/automatic-design-of-conversational-models-from-observation-of-human-to-human-conversation.html). The topic passed a scientific review by more than 40 world-class researchers in AI in Baltimore, USA, in December 2022, and was selected for this workshop out of 15 proposals together with three others. The workshop topic builds on the outcome of this microproject and will reuse the collected data.

XAI model for human readable data aimed at connected car crash detection

Date Start: 2021-14-10
Date finish: 2022-23-02
Duration:
Partners:

1. Generali Italia (Industrial Partner),
2. CNR Pisa, Università di Pisa

For insurance business a connected car is a vehicle where an embedded telematics device streams acceleration data, GPS position and other physical parameter of the moving car. This live streaming is used for automatic real time detection of car crash. The project is focused on the development of an XAI layer which translates the logical outcome of an underneath LSTM used for crash detection into a human readable format.
Industrial outcome: the LSTM automatic labeling of a signal event from a car telematics box as a 'crash' triggers an emergency live call from a contact center to the driver's phone for health assessment and further help. If the driver is not responding or is out of reach, more actions could follow (e.g. call to emergency service). In order to improve the efficiency of this emergency procedure, is vital for the contact center operator to reduce the number of false positive events (e.g. being able to read the outcome of the box and discriminate a false positive event)
Societal outcome: an improved efficiency in connected car crash detection (reduction of false positives) can reduce the number of car crashes with fatal or severe injury outcome and also improve road safety.

5. WP6-related results: Applied research with industrial and societal use cases

5.1. Overall summary

The work package is focused on translating core research into applications. The three main objectives are (1) ensuring that the needs of important European industry are adequately then into account within the research agenda, (2) making sure that key results are evaluated

[2] https://www.clsp.jhu.edu/2023-jelinek-summer-workshop , https://jsalt2023.univ-lemans.fr/en/index.html

in industrially (and socially) relevant use cases and (3) making sure that the knowledge created by the microprojects of WP1-5 reaches key European industrial players.

The MPs related to the goals of WP6 in the second period are:

- “Neural Mechanism in Human Brain Activity During Weight Lifting”
- “Polarization with the Friedkin-Johnsen model over a dynamic social network”
- “Trustworthy Voting Advice Applications”
- “The temporal and biological factors of our vulnerability to disinformation”
- “Telefonica 2– validating the air quality prototype in a real city”
- “Telefonica 3– assessing the ethical and societal impact of the air quality system”
- “Multi-layer evaluation sets for speech translation of web-based meetings”
- “XAI model for human readable data aimed at connected car crash detection”

The goal of **fulfilling the research needs of industry** is addressed by several MPs. The MP “Neural Mechanism in Human Brain Activity During Weight Lifting” is an industrial collaboration investigating the use of EEG (electroencephalography) signals for detecting the motion as well as the variable weights a person is lifting. These result could result in a variety of applications, fulfilling the goal of WP6 of fulfilling the research needs of industry. For example, this system could be used to develop rehabilitation systems robust to dynamic changes in weight. Moreover, information regarding weight change could contribute to a better estimation of fatigue condition to be used in sports and training applications. Finally, it has been evaluated that the approach to predict different categories of lifted weight could be used in further optimizations in industrial applications for which usage of exoskeleton can be given as an example. The MP “Multi-layer evaluation sets for speech translation of web-based meetings” produced datasets for translation (for English->Latvian, Latvian->English, and Lithuanian->English). The microproject produced data that will be beneficial for future developments of speech translation and automatic minuting systems.
Several MPs contribute to the WP6 goal of making sure that **key results are evaluated in socially relevant use cases**. The MP “Polarization with the Friedkin-Johnsen model over a dynamic social network” can have applications in the design of social networks, as it studies the effect different configurations can have on polarization. Likewise, the MP “Trustworthy Voting Advice Applications” studies voting advice applications already deployed on the market, therefore clearly evaluating socially relevant use cases. The MP “The temporal and biological factors of our vulnerability to disinformation” analysed a comprehensive dataset, examining the reliability of information relating to the COVID-19 pandemic shared on Twitter. The MP “XAI model for human readable data aimed at connected car crash detection” aims at the development of an XAI system to address connected car crash detection. Improving efficiency in connected car crash detection (reduction of false positives) can reduce the number of car crashes with fatal or severe injury outcome and also improve road safety, thus clearly addressing a socially relevant use case.

The objective of making sure that the knowledge created by the microprojects of WP1-5 **reaches key European industrial players** is addressed by the MPs “Telefonica 2– validating the air quality prototype in a real city” and “Telefonica 3– assessing the ethical and societal impact of the air quality system”. These MPs build on each other to evaluate and assess the ethical and social impact of the air quality system, which supports city governments to take data-driven decisions based to better manage challenges related to air

quality. These MPs have been conducted in collaboration with the administrations of the cities of Madrid and Valladolid.

5.2. Microproject descriptions

Neural Mechanism in Human Brain Activity During Weight Lifting

Date Start: 2021-05-03
Date finish: 2022-01-31
Duration: 8 months
Partners:

1. TUBITAK BILGEM
2. DFKI Kaiserslautern

Description:
In this project, it was investigated whether EEG (electroencephalography) signal can be used for detecting the motion as well as the variable weights a person is lifting. To do this, an experimental paradigm has been designed and EEG data have been acquired during performing biceps flexion-extension motions for different weight categories: lifting with no weight (empty), medium, and heavy lifting.
Features in EEG data generating difference for each lifted weight of category have been investigated. EEG data via different two EEG headsets have been collected from various participants while they lift different categories of load, namely empty, medium and heavy, in this project. Then, EEG data have been analyzed to realize if different category of weights result in difference in EEG data by applying different deep learning methods together with different machine learning methods. According to the obtained results, it can be said that that EEG signals can be successfully used as a method to predict different loads during dynamic bicep curl motion. Therefore, this result could result more researches to develop rehabilitation systems robust to dynamic changes in weight. Moreover, information regarding weight change could contribute to a better estimation of fatigue condition to be used in sports and training applications. Finally, it has been evaluated that the approach to predict different categories of lifted weight could be used in further optimizations in industrial applications for which usage of exoskeleton can be given as an example.

Results:
Presentation at the IEEE-EMBS International Conference on Biomedical and Health Informatics jointly organised with the IEEE-EMBS International conference on Wearable and Implantable Body Sensor Networks organized in Ioannina, Greece between 27-30 September 2022.
Publication with the title “Prediction of Lifted Weight Category Using EEG Equipped Headgear" in 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics Conference Proceedings.

The temporal and biological factors of our vulnerability to disinformation

Date Start: 2022-04-01
Date finish: 2022-12-31
Duration: 9 months

Partners:

1. ETHZ Dirk Helbing secretary-coss@ethz.ch, Elisabeth Stockinger (ETHZ) Elisabeth.stockinger@gess.ethz.ch
2. FBK, Riccardo Gallotti rgallotti@fbk.eu

Description:
Despite all efforts to mitigate mis- and disinformation, they continue to be a substantial problem. This project contributed to the literature base on mis- and disinformation about social media with an analysis of the interaction effects between temporal rhythms of disinformation and social media usage in the context of the COVID-19 pandemic.
Specifically, consider how mis- and disinformation spread on Twitter varies throughout the day and whether there are individual differences in users' propensity to spread mis- and disinformation on Twitter based on the activity patterns.
We analysed a comprehensive dataset, examining the reliability of information relating the COVID-19 pandemic shared on Twitter. We clustered users into pseudo-chronotypes based on their activity patterns on Twitter throughout the day, identified times of waking and prolonged waking states per cluster as well as times of increased susceptibility.
We aggregated out results into a paper and submitted an extended abstract to the International Conference on Computational Social Science (https://www.ic2s2.org/, accepted) and are in the process of preparing a paper for submission for a reputable journal. Elisabeth Stockinger from ETHZ spent a 3-week mobility period at FBK in Trento, Italy, to work with the partner directly. She has continued the collaboration as a virtual visiting student.

Multi-layer evaluation sets for speech translation of web-based meetings

Date Start: 2022-06-01
Date finish: 2022-12-31
Duration: 7 months
Partners:

1. Tilde
2. Charles University

Description:

Within the scope of the project, we have created evaluation and development data sets for speech translation for meetings (for English->Latvian, Latvian->English, and Lithuanian->English) (http://hdl.handle.net/20.500.12574/74), an automatic minuting test set for the AutoMin 2023 shared task on automatic creation of meeting summaries (“minutes”) for English and Czech (https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4692). We also contributed to the human evaluation of machine translation systems for the Seventh Conference on Machine Translation (WMT) for the English->Czech translation direction.

The microproject produced data that will be beneficial for future developments of speech translation and automatic minuting systems. The data will allow us to better understand capabilities of these systems and also identify potential areas of improvement. The results

will contribute towards developing robust and trustworthy AI systems. Furthermore, the project contributed towards mobilization of the research landscape by involving a European research institution (Charles University) and an industry partner (Tilde).

Telefonica – validating the air quality prototype in a real city

Date Start: 2022-01-03
Date finish:
Duration: 6 months
Partners:

1. TID
2. City of Madrid
3. City of Valladolid
4. National statistics office of Spain

Description:

This second micro project of WP6.10 will validate the air quality prototype developed in the first micro project with a second, real city, Valladolid in Spain. In principle, there will be no new developments except for feedback about the system from the city. This project is also in line with the objective of Humane AI: to shape the AI revolution in a direction that is *beneficial to humans both individually and societally*, and that adheres to European ethical values and social, cultural, legal, and political norms. The focus of the project will be on insights generated from data collected with mobile air quality measurement stations to be placed on top of vehicles, and to test how does insides help local governments in better managing the challenges around air quality.

Data from mobile air quality measurement stations that are placed on top of vehicles that and drive through all the streets of the city. Open data that is published by the local city of Valladolid. Aggregated and anonymized mobility data generated from a telecommunications network.

The city of Valladolid has agreed to run a pilot for 6 months to evaluate the air quality platform developed in the first micro project with the City of Madrid. The interest of the city council is to monitor the air quality in the low-emission area and beyond to understand whether they can take additional measures for improving the air quality to what they currently are planning.

Telefonica – assessing the ethical and societal impact of the air quality system

Date Start: 2022-01-06
Date finish:
Duration: 4 months
Partners:

1. TID
2. City of Madrid
3. City of Valladolid
4. National statistics office of Spain

This third micro project of WP6.10 focuses on assessing the ethical and social impact of the air quality system, which supports city governments to take data-driven decisions based to better manage challenges related to air quality. We want to, however, make sure that those decisions are fair and do not have undesired negative consequences such as boosting inequality and negatively affecting vulnerable groups. That is the objective of this last micro project in a series of three. The first micro project developed the prototype and the second validated it in a real city. We will do the assessment for the city of Madrid because this city has more relevant data available. This will be a collaboration with WP5. For assessing the ethical in fact, we will use open data from the city as well as from the National Spanish statistics office. Demographic data from census information such as gender, foreigners, age range, socioeconomic level, et cetera.

The descriptions of MPs “Polarization with the Friedkin-Johnsen model over a dynamic social network” and “Trustworthy Voting Advice Applications” can be found in Section 4.2.

X5LEARN: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site interface for access to openly licensed educational materials

Date Start: 2020-1214
Date finish: 2023-04-20

Partners:

1. Knowledge 4 All Foundation, Davor Orlic, davor.orlic@gmail.com
2. University College London, John Shawe-Taylor, j.shawe-taylor@ucl.ac.uk
3. Institut "Jožef Stefan", Davor Orlic, davor.orlic@gmail.com

Results:

Under this microproject, a series of extensions to the X5Learn platform was added. A new user-friendly user interface was developed and deployed. X5Learn, being an intelligent learning platform, a series of human-centric AI technologies that enable educational recommendation, intelligent previewing of information and scalable question generation that can help different stakeholders such as teachers and learners were developed backed by scientific research. The results have been published in peer reviewed conferences such as AAAI, AIED and CHIIR and also published in the Journal of Sustainability. The new earning platform is now available to the public including a python library that implements the recommendation algorithms developed.

6. WP7-related results

6.1. Overall summary

WP7 aims at providing research and mechanisms to support the creation of start-ups, the transformation of traditional (non-digital) SMEs into high-tech companies, and to push agile innovation in major industries. The main objectives are:

1. providing means and mechanisms to transform basic and applied research results into ventures and businesses that are provide value to European citizens,
2. to ensure that applied research is guided by real world challenges and steered toward domains that are beneficial for society.

The MP “Matching the right people! Creating a functional demonstrator for online matching of people and expertise for innovation” addresses the goals of this WP by considering a use case that is valuable for European businesses by developing a recommender system that suggests potential collaborators based on their demonstrated expertise. Social factors were considered by a user-centered design approach in the development of the system.

6.2. Microproject descriptions

Matching the right people! Creating a functional demonstrator for online matching of people and expertise for innovation

Date Start: 2022-06-01
Date finish: 2022-12-31
Duration: 6 months

Results:
Connecting the right people with the right expertise is essential for innovation, but finding suitable collaborators can be challenging. In this micro project, we explored a novel approach that leverages publicly available performance data (e.g., Github profiles) to automatically match people and expertise for innovation. Building on such data, we developed a recommender system that suggests potential collaborators based on their demonstrated expertise.
The project builds on the previous two micro projects, which identified a set of functional questions to match people for innovation and implemented a prototype for creating user profiles based on those questions. In this third micro project, we developed a demonstrator that creates recommendations for collaboration based on the automated collection and analysis of GitHub data. In contrast to the previous micro projects, the approach focuses on demonstrated expertise, rather than relying on self-reported knowledge, which can be inaccurate or incomplete.
To develop the demonstrator, we used the GitHub API to retrieve information about users' activities and built a recommender system that suggests potential collaborators based on their GitHub statistics. We also created a user interface for founders to find potential co-founders based on the expertise shown. This system offers a reliable and objective way of matching people with the right expertise, which can be especially valuable in high-tech environments where expertise is critical for success.
We evaluated the effectiveness and usability of the developed recommender system and user interface through a series of interviews and online studies. First, we interviewed domain experts to establish the requirements and important features of the model. Second, we interviewed potential founders to evaluate the general approach and gather feedback on the recommender system. Finally, we conducted an online study to evaluate the usability of the user interface. The study involved a group of participants using the interface to find potential collaborators and provide feedback on their experience. Overall, the evaluations provided valuable insights and feedback for improving the system's effectiveness and usability. We plan to incorporate the feedback from the evaluations to

further improve the system and enhance its ability to match people and expertise for innovation in the future.
The user-centered design process benefited from close collaboration and interaction with experts working at external partners ELLIS and the ETH Zurich AI Center. The long-term goal is to help connect experts and entrepreneurs across physical boundaries, creating a vibrant and agile high-tech environment on a European scale.

7. WP8-related results

7.1. Overall summary

WP8 aims at boosting the efficiency of collaboration, disseminating the latest and most advanced knowledge to all the academic and industrial AI laboratories in Europe.

The MP “Matching the right people! Creating a functional demonstrator for online matching of people and expertise for innovation” addresses the goals of this WP by considering a use case that aims at the distribution of expertise across industry and academia: the development of a recommender system that suggests potential collaborators based on their demonstrated expertise.

7.2. Microproject descriptions

The description of the MP “Matching the right people! Creating a functional demonstrator for online matching of people and expertise for innovation” can be found in Section 6.2.

8. New Microproject Procedures from January 2023

Following publication of the modified Humane AI Net Strategic Research agenda, documented in HAI Net Deliverable D6.1, during the consortium general assembly in October 2022 it was decided to implement new procedures concerning proposal and funding of micro-project. The procedure have come into force in January 2023.

Starting January 2023, the consortium will publish maintain a current list of research challenges that are to be updated 3 or 4 times year. Consortium partners may propose micro-projects responding to the current call at any time. Changes to the call are discussed in regular meetings of each workpackage, prior to publication of the new challenges. The administrative rules of the micro projects are as before (at least two partners, ideally 2-6 months duration, with 2-4 PMs per partner and obligation to produce a tangible result to be made available through the AI4Europe platform. Deviations from these guidelines may be approved by the project management committee with suitable justification.

The HumaneAI Net website must now be used for submission proposals for micro-projects. Each microproject should include visits between the sites. Partners who have run out for funds can now apply for micro-projects from our “reserve fund”. Preference for use of reserve funds will be given to projects that involve external partners those that involve industry

For the external group travel funds (including subsistence for longer stays) will be paid by the consortium coordinator. No person months can be paid for the external partners.

All proposals that meet the following evaluation Criteria will be automatically approved:

- Alignment with call objectives and the specific research direction
- Feasibility, innovation and societal relevance of the proposed approach
- Measurable impact potential of the solution on theories, methods, and societal or economic/business impact
- Quality of the proposed collaboration and partnership

These calls for proposals have been published in January 2023 for workpackages 4-8:

Measuring, modelling, predicting the individual and collective effects of different forms of AI influence in socio-technical systems at scale. (WP4 motivated)

The rise of large-scale socio-technical systems (STS) in which humans interact with AI systems, including assistants and recommenders, multiplies the opportunity for the emergence of collective phenomena and tipping points, with unexpected, possibly unintended, consequences. A better understanding is needed of the impact of AI systems on complex STS and the unique feedback loop they generate: the past evolution of a complex system influences the training of the AIs, which in turn influences the complex system's future evolution. For example, navigation systems' suggestions may create chaos if too many drivers are directed on the same route, and personalised recommendations on social media may amplify polarisation, filter bubbles, and radicalisation. On the other hand, we may learn how to foster "wisdom of crowds" and collective action effects to face social and environmental challenges.

This topic focuses on methods for measuring, modeling, predicting the individual and collective effects of different forms of AI influence in socio-technical systems at scale. In order to understand the impact of AI on socio-technical systems and design next-generation AIs that team with humans to help overcome societal problems rather than exacerbate them, we need to lay the foundations of Social AI, a new discipline at the intersection of Complex Systems, Network Science, AI and the (Computational) Social Sciences.

Activities that will be funded, include but are not limited to case studies, experiments, simulations and novel models and methods exploring the frontier of Social AI along the following dimensions:

- How can we describe STS rigorously in mathematical terms?
- What is the impact of AIs on individual and collective goals?
- What are the new network effects and collective phenomena due to the interacting human-AI system?
- How to design next-generation human-centered AI architectures that balance individual and collective goals with platform sustainability?
- How to make people aware of the impact of AIs on collectivity?

ELS evaluation projects (WP5 motivated)

This topic focuses on micro-projects that aim to assess/evaluate/monitor the implementation and adherence to ELS principles and guidelines (ethical, legal, societal). The micro-projects should involve collaboration between at least one partner from the network and at least one external organization or industrial player. Micro projects are expected to take 2-4 months and deliver a tangible output (e.g., demo, dataset, publication…)

Activities that will be funded, include but are not limited to:

- Research on methods and tools for assessment and monitoring ELS, with particular relevance those that address the European Trustworthy AI guidelines or AI act/
- Implementation and testing of ELS principles and guidelines in real-world scenarios
- Development and validation of metrics to evaluate ELS principles
- Dissemination and communication of the results and impact to relevant stakeholders
- Methods for the explanation and justification of the output of machine learning

Creation/Augmentation of realistic Datasets (WP6 motivated)

Having access to data and ensuring data privacy are difficult to realize at the same time. While data protection laws, such as GDPR, provide a form of safety for users, they can also create challenges for data engineers and AI practitioners. Most often, data is even available within a company, but cannot be accessed by other departments due to legal restrictions

This leads to a chicken-and-egg problem. Researchers cannot provide sufficient reasoning for data access as the merit of analyzing the data is not known a priori. Yet this merit can only be assessed if access to the data is granted.

Consequently, there is a need for datasets that do not fall under these restrictions. One option is anonymized data. However, completely anonymizing data is often complex and very costly. In this call for microprojects, we propose an alternate option: generating artificial data that has the same characteristics as restricted personal data. This artificial data could be used for preliminary data analysis, possibly warranting access to the real data. Apart from technical challenges, this call encompasses selected ethical and societal challenges:

- Creating a latent representation of the original data, that can be used to generate artificial data
- Evaluating the quality of artificial data in terms of its usefulness but also its degree of anonymization
- User modeling/personalization based on a latent representation instead of personal data
- Ethical aspects and legal boundaries of modeling users via digital twins (a latent representation of their personal data)

We invite micro projects covering one or more of the aforementioned challenges. Additionally, micro projects should focus on conducting applied research within industrial applications and with societal use cases in mind. While relevant for our overall research agenda (responsible usage of data), this call particularly addresses pillars 2 (providing

usable data for multimodal perception and modeling) and 5 (ethical aspect and legal boundaries of artificially created data).

Innovation projects (WP6&7 motivated)

We recognize that many research results remain in laboratory and do not reach market or end-users. This is why we aim to run a call on innovation projects to transfer research results and generate outcomes that benefit society across various domains including healthcare, finance, transportation, and more.

This topic aims to support the development and implementation of innovative AI solutions that not only have significant technological advancements but also have a measurable impact on society and the economy. Activities aim to address real-world challenges and opportunities in various domains such as healthcare, transportation, energy, and agriculture among others.

We invite proposals for microprojects that aim to develop and implement innovative AI solutions with significant socio-economic impact. The microprojects should involve collaboration between at least one partner from the network and at least one external organization or industrial player.

Activities that will be funded, include but are not limited to:

- Research and development of innovative AI solutions
- Implementation and testing of the solutions in real-world scenarios
- Measuring and evaluating the socio-economic impact of the solutions
- Dissemination and communication of the results and impact to relevant stakeholders

Evaluation Criteria:

- Feasibility and innovation of the proposed solution
- Relevance and alignment with the specific research direction
- Impact potential of the solution on society and the economy
- Quality of the proposed collaboration and partnership

Education & training projects (WP8 motivated)

Human-Centered AI mobilizes several disciplines such as AI, human-machine interaction, philosophy, ethics, law and social sciences.

The ambition of HumanE AI Net is to establish a training agenda to improve the education of a new generation of creative researchers and innovators, knowledgeable and skilled in Human-Centered AI. This call for micro-projects aims to create and distribute relevant dissemination and knowledge spreading materials such as Human-Centered Curricula, lectures, practicals, tutorials, MOOCs, which could take the forms of online materials as well as training events.

We encourage micro-projects engaging external partners, in particular micro-projects conducting in the context of the International Artificial Intelligence Doctoral Academy (AIDA), which gather the four ICT-48 networks (AI4Media, ELISE, HumanE AI NET, TAILOR) and the VISION project.

9. The Legal Protection Debt of Training-Datasets – Executive summary

This section is contributed by LSTS, as an executive summary of the Report which details the results of the involvement of LSTS in the microproject "Collection of datasets tailored for HumanE-AI multimodal perception and modelling".

LSTS (VUB) is a partner of the HumanE-AI Project, participating to WP5 on the legal and ethical bases for responsible AI. Under T5.2, LSTS visits microprojects and interacts with AI researchers to engage in a constructive technology assessment, teasing out potential risks for the rights and freedoms of natural persons who may suffer the consequences of implementation of AI.

The present Report is the result of the involvement of LSTS in the microproject "Collection of datasets tailored for HumanE-AI multimodal perception and modelling". Such microproject has been carried out in the context of WP2, Multimodal Perception and Modeling. The goal of WP2 is to provide integrated multi-modal perception and modelling to develop systems that can understand complex human actions, motivations and social settings. The microproject aimed at contributing to the research field by putting at disposal of the scientific community ready-to-use, curated datasets for multimodal perception and modelling of human activities and gestures. The partners involved have created, curated and released datasets for Human Activity Recognition (HAR) tasks, in particular, the extended dataset OPPORTUNITY++ and Wearlab BeachVolleyball dataset.

The participation to the microproject has offered the chance to get a closer look at the practices, doubts and difficulties emerging within the scientific community involved in the creation, curation and dissemination of training datasets. Considering that one of the goals of the HumanE-AI Net is to connect research with relevant use cases in European society and industry, the participation to the microproject has offered the occasion to situate dataset collection, curation, and release within the broader context of AI pipeline.

Under T.5.2., the task of LSTS is to provide researchers with recommendations on how to integrate legal protection by design into the architectures developed in the microprojects. To this end, the Report examines the potential issues that arise within current ML-practices and provide an analysis of the relevant normative frameworks that govern such practices. By bridging the gap between practices and legal norms, the Report provides researchers with the tools to assess the risks to fundamental rights and freedoms that may occur due to the implementation of AI research in real world situations and recommends a set of mitigating measures to reduce infringements and to prevent violations.

The Report acknowledges that datasets constitute the backbone infrastructure underpinning the development of Machine Learning. The datasets that are created, curated and disseminated by ML practitioners provide the data to train ML models and the benchmarks to test the improvement of such models in performing the tasks for which they are intended.

However, until recently, the practices, processes and interactions that take place further upstream the ML-pipeline, between the collection of data and the use of dataset for training ML-models, have tended to fade into the background.

The report argues that the practices of dataset creation, curation and dissemination play a crucial role in the setting of the level of legal protection that is afforded to all the legal subjects that are located downstream ML-pipelines. Where such practices lack appropriate legal safeguards, a “Legal Protection Debt” can mount up incrementally along the stages of ML-pipelines. In section 1.1., the Report provides a brief overview of how current data science practices depend on and perpetuate an ecosystem characterised by a lack of structural safeguards for the risks posed by data processing. This can lead to the accumulation of “technical debt”. Such debt, in turn, can assume relevance in the perspective of the compliance with legal requirements. Taking inspiration from the literature on technical and ethical debt, the Report introduces the concept of Legal Protection Debt. Because of this legal protection debt, data-driven systems implemented at the end of the ML pipeline may lack the safeguards necessary to avoid downstream harm to natural persons.

The Report argues that the coming about of Legal Protection Debt and its accumulation at the end of the ML pipeline can be contrasted through the adoption of a Legal protection by design approach. This implies the overcoming of a siloed understanding of legal liability that mirrors the modular character of ML pipelines. Addressing legal protection debt requires ML practitioners to adopt a forward looking perspective. Such perspective should situates the stage of development in practitioners are involved in the context of the further stages that take place both upstream and downstream the pipeline. The consideration of the downstream stages of the ML-pipeline shall, as it were, back propagate and inform the choices as to the technical and organisational measure to be taken upstream: upstream design decisions must be based on the anticipation of the downstream uses afforded by datasets and the potential harms that the latter may cause. Translated into a legal perspective, this implies that the actors upstream the pipeline should take into consideration the legal requirements that apply to the last stages of the pipeline.

The Report illustrates how data protection law lays down a set of legal requirements that overcome modularity and encompass the ML pipeline in its entirety, connecting the actors upstream with those downstream. The GDPR makes controllers responsible for the effects of the processing that they carry out. In section 2, the Report shows how the GDPR provides the tools to mitigate the problem of many hands in ML-pipelines. The duties and obligations set by the GDPR require controllers to implement by design safeguards that conjugate the need to address downstream harms with the necessity to comply with the standards that govern scientific research. In this perspective, the Report shows that the obligations established by data protection law either instantiate or harden most of the requirements set by the Open science and Open data framework and also the best practices emerging within the ML-community.

In section 2.1. the report illustrates the core structure of the regime of liability to which controllers are subject under the GDPR. Such regime of liability hinges upon controllers’ duty to perform a context-dependent judgment. Such judgment must informs controllers’ decisions as to the measures to be adopted to ensure compliance with all the obligations established by the GDPR. Such judgment must be based on the consideration of the downstream harms posed by the processing.

In essence, the duty to anticipate and address potential downstream harms requires controllers to adopt a forward-looking approach. In order to ensure compliance with the GDPR, controllers must engage in a dynamic, recursive practice that addresses the requirements of present processing in the light of the future potential developments. At the same time, the planning effort required by the GDPR is strictly connected with the compliance with obligations set by other normative frameworks. In this sense, compliance with the GDPR and compliance with obligations such as those imposed by the Open science and Open data framework go hand in hand. Compliance with the GDPR is a pre-requisite for complying with Open science and Open data framework. Simultaneously, the perspective of open access and re-usability of datasets affects the content of the obligations set by the GDPR.

As a result, the consideration of "what happens downstream" - i.e., the potential uses of datasets, potential harms that the latter may cause, further requirements imposed by other normative frameworks – back propagates, determining the requirements that apply upstream.

In section 2.2. we show how the compliance with the documentation obligations set by the GDPR can contrast the accumulation of a documentation debt and ensure controllers' compliance with the obligations established by other normative framework, such as Open Data and Open Science. The overlapping between the documentation requirements established by such different frameworks shows firstly that a serious approach to the compliance with the GDPR can provide the safeguards necessary to contrast the accumulation of a documentation debt. In this way, compliance with the documentation obligations set by the GDPR can prevent the accumulation of other forms of technical debt and, eventually, of legal protection debt. At the same time, the convergence between the requirements set by the GDPR and those established by the FAIR principle and the Horizon DMP template shows how the performance of the documentation obligations established by the GDPR can also facilitate compliance with requirements specific to data processing conducted in the context of scientific research.

A correct framing of the practices of dataset creation, curation and release in the context of research requires to make an effort towards the integrity of the legal framework as a whole, taking into consideration the relations between Open data, Open science and data protection law. First, it is first important to stress that compliance with data protection law represents a pre-requisite for the achievement of the goals of Open Data and Open Science framework.

In section 2.3. the report analyses the requirements that govern the release and downstream (re)use of datasets. Compliance with the requirements set by the GDPR is essential to avoid that dataset dissemination gives rise to the accumulation of legal protection debt along ML pipelines. Based on the assessment of adequacy and effectiveness required for all forms of processing, controllers can consider the adoption a range of measures to ensure that data transfer are compliant with the GDPR. Among such measures, the Report examines the use of licenses, the providing of adequate documentation for the released dataset, data access management and traceability measures, included the use of unique identifiers.

The Report contains an Annex illustrating the provisions of the GDPR that establish a special regime for the processing carried out for scientific research purposes. We highlight how most of the provisions contained in the GDPR are not subject to any derogation or exemption

in view of the scientific research purpose of the processing. All in all, the research regime provided by the GDPR covers the application of a limited number of provisions (or part of provisions). A processing that is unlawful in that it does not comply with the general provisions set by the GDPR cannot enjoy the effects of the derogations provided by the research regime. The derogations allowed under the special research regime concern almost exclusively the GDPR provisions on the rights of data subjects, while no derogation is possible for the general obligations that delineate the responsibility of the controller. The derogations provided under the special research regime allow controllers to modulate their obligations towards data subjects where the processing of personal data is not likely to affect significantly the natural persons that are identified or identifiable through such data. As it were, the decrease of the level of potential harm makes possible the lessening of the safeguards required to ensure the protection of data subjects. Even in such cases, however, no derogation is allowed with respect to the requirements different than those concerning the rights of data subject. This circumstance makes manifest that the system established by the GDPR aims at providing a form of protection that goes beyond the natural persons whose personal data are processed at that time by controllers.