HumanE AI Net:

The HumanE AI Network

Grant Agreement Number: 952026
Project Acronym: HumanE AI Net

Project Dates: 2020-09-01 to 2023-08-31
Project Duration: 36 months

D6.1 Strategic Research Agenda

Author(s): Paul Lukowicz
Contributing partners: John Shawe-Taylor, James Crowley, Antti Oulasvirta, Virginia Dignum, George Kampis.
Date: Mai 10, 2022
Approved by: Paul Lukowicz
Type: Report ®
Status: final
Contact: Paul.Lukowicz@dfki.de

Dissemination Level		
PU	Public	X

DISCLAIMER

This document contains material, which is the copyright of *HumanE AI Net* Consortium parties, and no copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the *HumanE AI Net* Consortium as a whole, nor a certain party of the *HumanE AI Net* Consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

DOCUMENT INFO

0.1 Authors

Authors	Institution	e-mail
George Kampis (GK)	DFKI	George.Kampis@dfki.de
Paul Lukowicz (PL)	DFKI	Paul.Lukowicz@dfki.de
Antti Oulasvirta (AO)	Aalto University	antti.oulasvirta@aalto.fi
John Shawe-Taylor (JST)	UCL	j.shawe-taylor@ucl.ac.uk
James Crowley (JC)	INRIA	James.Crowley@inria.fr
Virginia Dignum (VD)	UMU	virginia.dignum@umu.se

0.2 Document History

Revision		
Date	**Lead Author(s)**	**Comments**
30.09.2020	GK	Empty template
10.05.2022	PL	Offline draft
10.05.2022	GK	Final formatting

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This document described the research agenda of the HumanE AI project in terms of the evolution with respect to the research questions described in the proposal. We focus on research questions that are directly related to the projects vision of AI that enhances human capabilities and empowers citizens both in individual and collective/social level while observing ethical and fundamental rights concerns "by design". We leave the work on more generic AI research agenda to collaboration with roadmapping activities within VISION and our collaboration with the CLAIRE roadmapping effort. We also closely collaborate with the EILSE project towards joint research agenda items that combines our AI-human based angle with ELISE more fundamental ML oriented vision (currently a joint call for proposals of microprojects is being defined).

Most significant evolution of the research agenda has taken place at the interface of the individual WPs as result of synergies between the different communities (core AI, HCI, Ubiquitous computing, Social Science, Complexity Science) producing which we consider significant, critical insights on what is needed to move towards a vision of truly human centric European AI. Described in detail in section 3 these include:

1. A hierarchical framework to provide a taxonomy of research problems for collaborative AI systems. Solutions to problems at any level can build on techniques and solutions developed at lower levels. The framework is proposed as a research roadmap, grouping related challenges into subcategories according to the information that is processed and the nature of the interaction. This facilitates formulation and comparative evaluation of competing techniques.
2. An understanding how the question of a common ground and shared representations relates to different types of interaction and what are key directions that we need to explore to facilitate our vision of human centric AI. In particular, we emphasize the role that the concept of narrative, leveraging recent advances in NLP and selfie supervised multimodal representation learning can play across the WPs.
3. An extension of the concept of trustworthy and explainable AI from a definition focused on technical aspects to a user-oriented approach that emphasizes systems that act and interact in a way that people and the society feel comfortable trusting and using.
4. The insight that we need to work on a research methodology and infrastructure that brings together the different cultures of the discipline involved. This includes a definition of evaluation standards and experimental methodologies as well as the creation of data sets and tools.

Within the individual WP research agendas (see section 4) the adaptations focus on incorporating new developments in the respective fields and synchronising with the overall crosscutting adaptations outlined above.

We consider this research agenda to be a "living document" that will be continuously updated as new insights and ideas arise from the project work and overall progress in the field.

2. INTRODUCTION

The project is committed to a vision of Human Centric, Trustworthy AI with European Values and the translation that vision into innovation that will make the European economic competitive and the European way of life viable in the future. As unique "selling points" we

- focus on AI that enhances human capabilities and empowers citizens
- consider the role and effect of AI with respect to both: individuals and the society as a whole
- do research as to how ethical and fundamental rights relate to AI, "Question Zero" (where to integrate AI solutions, where to abstain) and protection by design

We address the above aspects from an interdisciplinary perspective that leverages the unique composition of the consortium: building on all key areas of AI with a strong focus on human-computer Interaction and contributions from social science, law, philosophy and complexity science.

The research agenda presented in this document reflects the above vision. Thus, we do not propose a "standard", generic AI research agenda based on items that can be found on any current roadmap. Nor do we reject such problem, but rather we seek to build on interaction with other networks through the VISION-research agenda activity. We have also been active in the CLAIRE roadmapping activity and are closely cooperating with ELISE when it comes in depth ML related issues. In this document we focus on questions that are situated at the interface of AI, HCI, social science and complexity science and are critical for AI systems closely cooperating with humans and interacting with society. Such questions were already outlined in detail in the proposal which, within WPs 1–5 was structured as a research agenda in itself. The vast majority of the questions from the proposal remain valid at this stage of the project. In this document we concentrate on extensions and modifications of these questions that resulted from work and analysis in the first period of the project. The most important are the ones resulting from synergies between the different work packages and the communities involved in the network. This type of interdisciplinary synergy is a core aspect of HumanE AI Net. We outline those cross WP insights and new challenges in section 3 as the main contribution of this deliverable. This is followed by a task-by-task analysis of the modifications/extension of specific, narrower "vertical" research questions in section 4. This includes some new agenda items that we propose to add to the respective WPs.

3. TOP-LEVEL CONCLUSIONS AND CROSS-WP TOPICS

After the first 18 months we are confident the overall research agenda described in the proposal remains valid and highly relevant for Human Centric AI. However, as was expected, there are some adjustments, extensions that became apparent during the work so far. The main insights resulted from the interactions between the different communities involved in the project and ai concern cross cutting topics which is why we describe them in a dedicated section. We believe that these are insights that are highly valuable not just for project re-adjustment but also to the community as a whole.

3.1 UNDERSTANDING HUMAN-AI COLLABORATION (COLLABORATIVE AI).

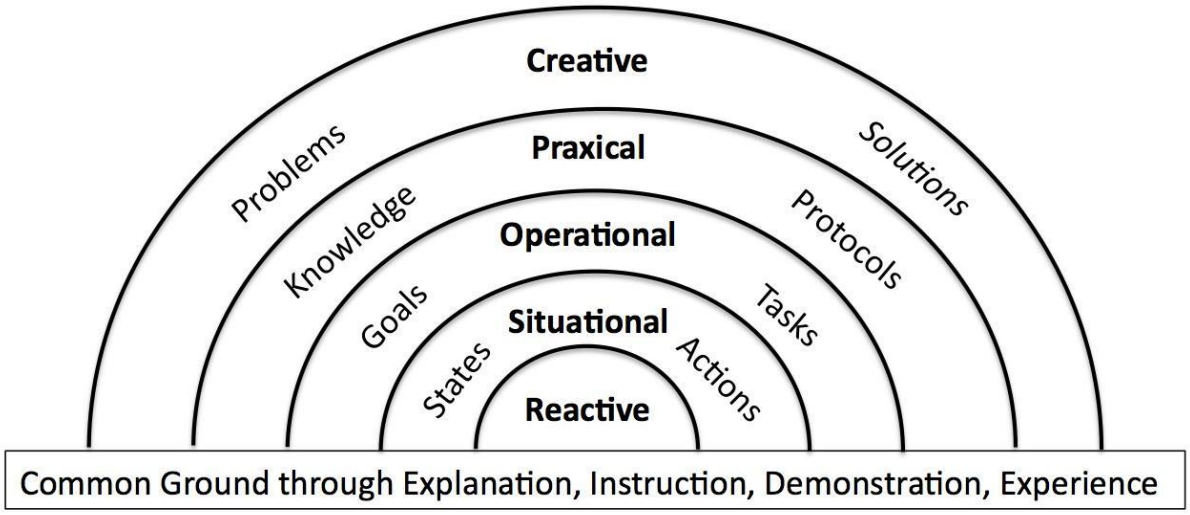

Figure 1 A hierarchy of capabilities for research on collaboration with intelligent systems. Collaboration at each level builds on abilities at the lower levels. Common ground for collaboration requires shared understanding of Situations, Goals, Preferences and Problems. Common ground is reached through explanation, instruction, demonstration and experience.

An important step on the way to systems that can flexibly enhance human capabilities is a framework that can deliver transparent collaboration between intelligent agents, be they artificial or human. We are mainly considering situations in which the overall goals are aligned, but there may be a need to coordinate individual goals and roles in specific tasks. We are conscious that there is no theoretical framework that can adequately capture the scenario at this level of generality, but the model of collaborative intelligence systems outlined above provides a delineation of the different levels at which collaborations can be carried forward. We are therefore proposing a new topic that will explicitly develop new theoretical models for multi-agent collaboration. Such a topic spans WPs 1,2 and 3 touching on WP 4. Collaboration of any kind requires communication between agents in order that they can coordinate their roles and the division of tasks as well as conveying information pertinent to the execution of those tasks.

Theories and experiments from Cognitive Science [Johnson-Laird 89], Ergonomics [Endsley 2000] and Multi-modal Human-Computer Interaction [Oviatt 2017] show that humans observe, model, act and interact using multiple modalities with multiple temporal scales and multiple frames of references. Accordingly, human-AI collaboration can be organized as a hierarchy of perception-action cycles, each with specific representations for information. We refer to the levels in this hierarchy as reactive (sensory-motor), situational (spatio-temporal), operational (task-oriented), practical (experience-based) and creative, as illustrated in Figure 1.

1. **Reactive collaboration** assumes a form of a tightly coupled interaction where the actions of each agent are immediately sensed and used to trigger actions by the other. A classic example is controlling a pointer on the screen of a computer using a mouse. Reactive collaboration between people and machines requires that machines sense and act with a similar time scale as the human. Sensory-motor reflexes in humans occur over a time scale of 80 to 300

milliseconds with reaction times determined by the number of neural layers between the sensing organ and the muscle activation units. Effective interaction between humans and machines requires that the temporal and physical properties of the machine be tuned to the sensory-motor reflexes of the human collaborator.

2. **Situation Aware collaboration** refers to an interaction where perception and action are mediated by shared awareness of a situation. Situation awareness has long been recognized as a core competence for intelligent behaviour, as well as survival in critical environments. The term can be traced to the early 20th century, where situation awareness was identified as a crucial skill for crews in military aircraft. Situation awareness has been recognized as a foundation for successful decision-making across a broad range of domains including law enforcement, navigation, healthcare, emergency response, military command, and self-defense. Inadequate situation awareness has been identified as one of the primary factors in aviation accidents attributed to human error [Endsley 99].
3. **Operational Collaboration.** The operational level concerns the planning and execution of tasks. Information at the operational level includes the current and desired situations; their expression as goals and sub-goals; and tasks sub-tasks and plans of actions that can be used to attain the desired situation. This level can also concern actions that can be used to attain or maintain a stable situation, as well as detection of threats and opportunities.
   Operational collaboration requires sharing authority. Authority may be shared with a strict hierarchy, where one agent has the power to over-rule the actions of the other as with an aircraft or maritime crew. Authority is often shared using a protocol where each agent has a primary authority over a particular task domain, with a possibility of accepting delegation of authority in other domains. Authority may also be shared equally where each agent is free to initiate tasks according to its understanding of the common goals and current situation, as can occur in some forms of team sports such as football or ice hockey.
4. **Practical collaboration** refers to the exchange of knowledge about how to attain goals and maximize value based on experience or training. Human society exists because of our capability to share experience and coordinate activities. Full collaboration with intelligent systems will require similar abilities: the capability for humans to communicate and share knowledge and experience with intelligent systems and the capability for intelligent systems to communicate and share knowledge with humans.
5. **Creative collaboration** refers to a form of interaction where two or more partners work together to solve a problem or create an original artifact. This could range from elaboration of a theory or model to explain a phenomenon to creation of a performance, painting or sculpture, to the design of a tool or system. In the most effective forms of creative collaboration, each partner evaluates the observations and analysis of the partner in order to offer constructive criticism or to reinforce and build on emerging insights. When two partners work well together, a form of creative resonance emerges in which each partner improves and builds on the ideas of the other.

Each level of the hierarchy concerns interaction with distinct forms of information: Sensory-motor signals for the reactive level, entities and relations for the situational level, tasks and plans for the operational level, domain specific knowledge about how to perceive and act for the practical level, and problems, hypotheses and solutions at the creative level.

The framework is designed to provide a taxonomy of research problems for collaborative systems rather than a system architecture. Solutions to problems at any level can build on techniques and solutions developed at lower levels. The framework is proposed as a research roadmap, grouping related challenges into subcategories according to the information that is processed and the nature of the interaction. This facilitates formulation and comparative evaluation of competing techniques.

However, given that the framework takes its inspiration from the RCS reference architecture for robotics control, as well as from cognitive models of biological systems, it is conceivable that it may also serve as a reference model for designing systems, providing a functional decomposition for collaborative intelligent systems. However, this is speculation. To the best of our knowledge, no actual systems have been constructed using such a model.

Natural Language understanding and generation permeate at all levels. Reading, writing, listening and speaking have substantial sensory-motor components. Much of natural language communication is about describing situations. For example, the verb of a sentence is a predicate describing the relation between the subject entity and one or more object entities. Natural language can be used for coordinating actions during operational collaboration. Natural language is highly effective and widely used for communicating practical knowledge and for creative collaboration.

Enabling technologies for collaboration with intelligent systems have potential for significant societal impact and wealth generation. An obvious example is in computer games and virtual worlds (the metaverse) where enabling virtual characters with abilities for situation understanding, operational collaboration and creative problem solving can have enormous financial impact. Similarly, virtual musical groups with simulated players that can sense and react musically to a musician, can engender creativity by empowering musicians to explore new forms of music without the clash of egos that can occur in real musical bands.

Chatbots are another area where an enabling technology for collaborative problem solving could provide enormous wealth generation. Most current chatbot technologies rely on pre-programmed linguistic patterns to generate plausible natural language responses to queries, while restricting responses to pre-programmed answers or interpretations of results from search engines. Empowering chatbots with an ability to creatively explore solutions with users would have an enormous appeal, for example by providing virtual customer service agents for online commerce that could guide and assist users in planning purchases in complex specialty areas such as home appliances, wines or vacation planning. Augmenting such agents with a technology for social interaction coupled with abilities to inspire pleasure and confidence has an enormous potential for social impact and wealth creation.

Technologies to permit humans and intelligent systems to collaboratively analyse problems and determine solutions augmented with explanations and courses of

possible actions can have enormous impact for social impact and scientific understanding. Finding common ground at all levels is key.

3.2 Common Ground and Shared Representations

Effective communication, collaboration, and trust all depends on the stability of the involved partners to relate to a common understanding of the world. This includes understanding of the situations, understanding of the effects of actions or events and the understanding of the manner of attaining objectives. Thus, a key problem that human centric AI needs to address is bridging the gap between the human and machine “understanding”, including relating human world models and AI/ML representations built from multimodal input data. While the question of building world models from multimodal data is already at the centre of WP 2 the notion of shared representations goes further, towards the ability of dynamically establishing shared understanding of complex situation between AI agents and different humans that are suitable for various types of collaboration (see point 1 above). This is an aspect that cuts across WPs 1,2 and 3 requiring expertise from core ML, symbolic AI, perception and HCI.

In a recent paper in Nature [Dafoe 2021] the authors argue for a science of cooperative intelligence based on machine abilities to understand, communicate and interact with people, under the guidance of norms and institutions. The authors referred to this as "finding common ground" with AI systems. Common ground is a metaphor, generally understood as a basis for mutual understanding that can be found or established in negotiations. Finding common ground and the related problem of mutual understanding provide an interesting perspective on research challenges for collaborative AI.

Common ground begins by agreeing on the facts of a situation: the verifiable observations of the entities and relations that describe a situation. This requires a shared ability to detect entities and relations as well as a shared vocabulary of terms so that information can be exchanged. Established approaches for machine perception use supervised learning to pre-train object detectors with labelled training data from a pre-defined set of categories. This approach limits perception to a closed set of pre-defined entities and relations, thus limiting the set of situations that can be modelled or communicated. An important challenge for situation-aware collaboration is to provide a means to perceive and communicate information about new entities and new relations. This requires an ability to learn from demonstration, explanation and interaction.

Learning to recognize entities can be particularly challenging for functional categories of objects based on affordances. Affordances are the qualities or properties of an object that define possible uses for the object and indicate how it can be used [Gibson 77]. Functional categories are sets of entities that share affordances. For humans, learning to recognize and name functional categories is facilitated by experience. Without reference to the experience of sitting, machines are reduced to superficial detection of chairs based on shape and appearance.

Common ground also requires agreement on how actions can change situations. Consider, for example, the problem of learning to perform an action by observation. For humans, learning from observation is greatly aided by a phenomenon known as mirroring. Observing the performance of an action or the display of an emotion can trigger the feeling of performing the action or having the emotion. Researchers in cognitive neuroscience explain mirroring as the result of a common sensory-motor

neural code used for both perception and action that enables a person to feel an action as it is performed [Prinz 97]. Observing another person performing an action activates the sensory-motor system associated with the action, creating a feeling similar to performing the action, facilitating retention and learning. Mirroring can also trigger memories of previous episodes of the action making it possible to predict consequences and potential outcomes. Providing intelligent systems with a capability to learn novel policies for mapping perception to action from observation without a capability for mirroring is a major open challenge for machine learning.

Mirroring may also play a role in predicting the intentions of a collaborator, facilitating efficient coordination without the need for explanation and enabling monitoring of actions to detect and prevent errors. The incapacity to mirror actions and emotions is a barrier for finding common ground with intelligent systems that may be partially offset by explanation.

3.2.1 Narratives

Narratives are already part of WP1 and WP2 research agendas and have been involved in several MPs. Narratives can be seen as an operational form of a shared representation. We devote a dedicated section to Narratives out of the insight about their importance for the HuamnE AI Net vision on many levels. In this sense we see narratives as a core research topic cutting across WPs 1,2,3 and to a degree 4.

Narrating is one of the most human things we do. Narratives provide the structure by which individuals and communities understand the world in which they live. They tell individuals which elements and processes are related to each other, how to structure their experience, and how to evaluate other individuals, objects and processes. Consecutive actions or sets of objects may be seen as related or unrelated, depending on the narrative in which they become embedded. Narratives allow humans to make sense of the world and communicate their insights and experiences.

Narrative embedding relies on conceptual representation of objects and actions. Humans relate to observed objects and actions by matching them to known narratives. This gives meaning to those objects and actions. For example, a dog may be embedded into a family narrative, “when I was visiting my aunt, her dog was resting on her lap…”, and a hunting narrative, “dogs surrounded the bison” – in each of these narratives the meaning of a dog will be very different.

Narratives represent a human understanding of causality, which is subjective, non-binary, not rational, and based on complex causal relations. Humans communicate causality by telling stories, which present the consequences of actions, decisions, and events. These causal relations are generalized into more general narrative schemas. Individual stories are merged with other stories of a narrative community defining narrative schemas (e.g., how things are and how the world around us works) which constitutes an important element of a group or a societal culture. In this context narratives transmit values; they are the building blocks of ethics. Values and ethics of almost all religions are conveyed as stories. For children, the world of values and ethics is constructed to a significant degree by stories: fairy tales. Beyond values, narratives provide means of accumulating, transmitting and storing experience. Both the formal ways of transmitting experience (e.g., in education) and informal ones rely on stories.

Stories and narratives play a central role in the human understanding of the world, learning about the world, and communicating about the world, including in particular,

reasoning about and communicating about values, ethics and culture. In contrast to current approaches in AI the narratives approach is focusing on enabling intelligent systems that develop an understanding of contextualized phenomena and that can explain actions in complex real-world scenarios.

3.3 HUMAN/SOCIAL VIEW. THE ANGLE OF EXPLAINABILITY AND TRUSTWORTHINESS.

Explainability and trustworthiness, which today in AI are defined mostly in technical terms (system performing according to formal specifications, mapping decisions made by the systems to parts of the input space etc) must be recast in terms of Human Computer Interaction and social aspects. Thus, the question is not (only) about the system performance with respect to some sort of formal, technical metric (e.g., accuracy) but if the way system acts and interacts in a way that makes users included to see it as trustworthy. Similarly, explanations of system functionality must be cast in terms that correspond to mentals models that the user has of the respective situation/task and satisfy the user's personal level of required understanding in a given situation. In this sense both

3.3.1 Explanations

An explanation can be defined as a statement or narrative that makes something clear. For situation modelling, the "something" is the situation, including the underlying entities and relations as well as associated actions and intentions. For operational collaboration, an explanation can provide a description of the sequence of intended actions that can take a situation to a desired state, as well as a description of the sequence of intermediate situations that can be used to ensure the proper execution of actions and operations. For practical collaboration, explanations can be used to share knowledge about how to obtain information and coordinate actions based on habits and customs. Explanations can also facilitate agreement on protocols for interaction and collaboration, facilitating coordinated action and recognition of intention. Explanations can be used to describe hypotheses for creative collaboration. An ability to generate and interpret explanations is key for all levels of collaboration.

Explanations can be structured as narratives, describing a sequence of situations including the actions that drive the sequence, and alternative outcomes that can result from unsuccessful actions or external intervention. Such a narrative structure can substitute for a lack of experience by providing grounding for interpreting instructions. Explanations can be used to compensate for the lack of experience when providing instructions.

Explanations can be used to diagnose and learn from the results of operations "after the fact" when operations fail to provide a desired outcome. An explanation can help identify whether the failure was due to incomplete or inaccurate model of the situation or the result of erroneous assumptions or some other cause. A narrative that explains the understanding of the situation and the reasons for selecting actions can be used to learn from the failure and improve operations for the future.

Explanations can be formulated as responses to the Quintilian questions Who, What, Why, When, Where, and How. Specifying these elements is key to establishing a shared situation model and a shared agreement on operational plans and authority, whether describing past, present or future situations. "What" and "where" describe the entities and relations that compose a situation. "Why" describes the desired or goal

situations. "How" concerns the sequence of actions or operations that can be used to reach the goal. "When" describes the conditions under which the actions and operations can be performed. "Who" assigns the operational authority to perform the actions or establishes a protocol for determining authority during the operation.

Explanations provide a powerful technique for sharing and negotiating situation models, operational plans, practical knowledge and creative solutions. Developing technologies for dynamically generating explanations and for interpreting explanations are an important challenge for collaboration with intelligence systems

3.3.2 Trust

Trust in collaboration means (1) behaving in such a way that depends on another agent in a risky task while (2) having a belief consistent with this behaviour. Beliefs emerge via interaction. The key to trustworthy AI is helping agents form accurate trust-related beliefs that can serve as a basis of reducing risk. We believe that “trusting” is not a passive belief-formation process but an active process where users seek to act in a way beneficial to them under risk. The joint human-automation optimum is more likely achieved if both agents’ beliefs are correctly calibrated (no overtrust, undertrust). Beliefs that form via interaction are critical: According to prior work, beliefs related to integrity, benevolence, competence or ability, and risk are central and communicated by the policy of the agent. Trust-related beliefs are multi-faceted, they include:

- Goal belief: does this agent have a goal to help/harm me?
- Disposition belief: does this agent have a disposition to act in a way that could help/harm me?
- Competence belief: does this agent have the competence to actually achieve that?
- Situational belief: does this agent enter situations where it might do that?

3.4 Research methodology and infrastructure for Human Centric AI.

During the initial months of the project, it became increasingly apparent that there is a pressing need to define an overreaching research methodology that bridges the differences between the individual disciplines needed to address core problems of human centric AI. The primary concern are evaluation metrics and methods that combine bare technical performance metrics with user acceptance, usability and social aspects (see 3.3 above). This goes beyond the well-known cost factors that weight different types of errors according to their significance towards subtle, dynamic, situation and user specific assessment that takes into account the impact on the effectiveness of the system within the different types of interaction defined in 3.1. A related issue is experimental methodology including ethical aspects of data collection and experimentation.

In parallel with the methodology definition, we need to provide tools and infrastructure such as data sets, evaluation scripts, repositories, base line evaluation sets and benchmarking support. Finally, the research methodology aspect needs to be incorporated in education, in particular, at the Ph.D. level.

4. CONCLUSIONS (WORK PACKAGE BY WORK PACKAGE)

4.1 HUMAN-IN-THE-LOOP MACHINE LEARNING, REASONING, AND PLANNING

The aim as stated in the proposal is:

"Allowing humans to not just understand and follow the learning, reasoning, and planning process of AI systems (being explainable and accountable), but also to seamlessly interact with it, guide it, and enrich it with uniquely human capabilities, knowledge about the world, and the specific user's personal perspective. "

On this high level of abstraction, the aim remains unchanged as a **core specific direction of HumanE AI Net** in the area of learning and reasoning. Key top-level adjustments/extensions of are:

1. There is work needed to identify effective benchmarks for measuring the effectiveness of learning/reasoning techniques in the context of human AI interaction or social AI systems. Such benchmarking methodologies could be a major contribution of the project and might also be developed with industrial partners (see also 3.4).
2. As the field is progressing at very high speed during the time since the conception of the proposal new developments have emerged that need to be taken into account. On example is tranformers, which have been for a couple of years but are only recently being fully recognized for their potential in many different areas, including multimodal perception.
3. The learning and reasoning work needs to be better considered in the context to the Human-AI interaction framework described in 3.1.

The developments within the specific concrete areas and be summarized as below.

4.1.1 Linking symbolic and sub-symbolic learning

4.1.1.1 Original Research Goals

The original aim here is the "construction of hybrid systems that combine symbolic and statistical methods of reasoning". A specific methodology stipulated in the proposal is is the consideration of narratives — which are particularly natural representations for humans that might well offer a fruitful common ground with machine representation, an insight that goes back to early work in AI on scripts.

4.1.1.2 Selected Microprojects/Results

- **Frank van Harmelen**, "Neural-Symbolic Integration: explainability and reasoning in KENN". Defining a computational framework for integrating symbolic and sub-symbolic, through adding a layer that converts neural network outputs incorporating known constraints. Backpropagating through these layers enhanced learning in the sense of replacing some of the training data with known constraints. Different 'co-norms' were shown to improve learning.
- **Haris Papageorgiou,** "Combining symbolic and sub-symbolic approaches - Improving neural Question-Answering-Systems through Document Analysis

for enhanced accuracy and efficiency in Human-AI interaction". Question answering using a neural system building on top logical components to combine information from different sources relevant to the original question.

4.1.1.3 Direction of Adjustments/Extensions

There has recently been tremendous success in terms of interfacing large scale language models (e.g., BERT) as semantic interfaces between human understandable knowledge and ML systems. A well-known example is that of Zero Shot Learners, where classes for which no training data "seen" by the ML system can be correctly classified by relating the ML system internal representation to semantic spaces derived from NLP models. Along the same lines we are increasing seeing systems (e.g., CLIP) that jointly learn language and image (sound, sensor, etc.) representations leveraging language knowledge to improve the sub-symbolic learning performance.

4.1.2 Learning with and about narratives

4.1.2.1 Original Research Goals

The original aims were defined as: "We will investigate the use of narratives to provide human-understandable descriptions for complex situations, and sub-symbolic representations. We also will research how narratives can be adapted as a bridge between human reasoning and understanding, on the one hand, and internal AI representation on the other."

The developments within the specific concrete areas and be summarized as follows.

4.1.2.2 Selected Microprojects/Results

- **John Shawe-Taylor,** "Educational Recommenders with Narratives". Representing the semantics of educational materials in terms of the wikipedia topics that are referenced and using this representation to learn the appropriateness of different content for individual learners. The narratives correspond to sequences of materials that can guide a learner in approaching novel learning goals.
- **Chiara Ghidini**, "Discovering Temporal Logic patterns as binary supervised learning". To discover explicit temporal representations from executions that can be considered as symbolic forms of possible narratives. These can then inform humans in order to help them make sense of what is happening in the data and check whether the narratives are compliant with what was anticipated.
- **Guido Caldarelli,** "Creation of stories and narrative from data of Cultural Heritage". The goal is to use AI and Complex Networks methods to extract information about the structure of the society in the past as reconstructed from historical archives.

4.1.2.3 Direction of Adjustments/Extensions

A core issue is leveraging the recent (evolution) in NLP (see also 4.1.1) to building and using narratives. Another important aspect is linking "classical" text-based narratives with multimodal representations such as videos, sound, and sensor data.

4.1.3 Continuous and incremental learning in joint human-AI systems

4.1.3.1 Original Research Goals

A core concern is the use of hybrid incrementally modifiable representations in joint human-machine learning and planning. One example from reinforcement learning is to learn an intelligible abstraction of the state-space (the world) and the possible transitions, and then learn a reward function over this abstract model. Another the use of hybrid representations in generating explanations based on shared models between humans and machines.

4.1.3.2 Selected Microprojects/Results

- **Dilhan Thilakarathne,** "Can we use ML tasks as a proof of work in consensus algorithms?"
- **Davor Orlic, "**X5LEARN: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site interface".
- **Mehdi Khamassi, "**Coping with the variability of human feedback during interactive learning through ensemble reinforcement learning". The main result is a combination of model-based and model-free reinforcement learning using a meta-controller. Experiments have been performed in several scenarios including interactive robot learning and cooperation tasks. Article submitted.

4.1.3.3 Direction of Adjustments/Extensions

This research direction remains highly relevant. However, it needs to be better aligned with work in WP2 and 3 in the light of the interaction framework concept defined in 3.1.

4.1.4 Compositionality and automated machine learning (Auto-ML)

4.1.4.1 Original Research Goals

Enable the combination of symbolic and statistical AI methods and further extend them with theoretical models that allow continuous adaptation. A core goal devising methods for automating the development, deployment, and maintenance of AI systems that are performant, robust, and predictable, without requiring deep and highly specialised AI expertise. The key to achieving this vision of automated AI (or AutoAI) is our proposed approach for rendering AI systems interpretable by learning to decompose them into simpler components, which can automatically identify key structure in the solution, hence rendering it more robust and explainable.

4.1.4.2 Selected Microprojects/Results

- **Joao Gama**, "Online Deep-AUTOML".
- **Uwe Köckemann,** "AI Integration Languages: a Case Study on Constrained Machine Learning". We have implemented the moving targets algorithm in the AIDDL framework for integrative AI. This has benefits for modelling, experimentation, and usability. On the modelling side, it enables us to provide applications of "moving target" as regular machine learning problems extended

with constraints and a loss function. On the experimentation side, we can now easily switch the learning and constraint solvers used by the “moving targets” algorithm, and we have added support for multiple constraint types. Finally, we made the “moving targets” method easier to use, since it can now be controlled through a small model written in the AIDDL language.

4.1.4.3 Direction of Adjustments/Extensions

There is a need to develop a theoretical framework for compositionality that could inform practical algorithms that can guide and to some extent automate composition of relevant sub-systems, perhaps developing an auto-compositionality theme.

4.1.5 Quantifying model uncertainty

4.1.5.1 Original Research Goals

For AI to interact meaningfully with humans it must use the vocabulary and semantics of probabilistic arguments in a way that is accessible and understandable to humans. However, uncertainty quantification is not just important as a vocabulary of communication, it is also a vital component if an agent is to weigh different alternative interpretations of a situation, to assimilate information from different sources, and to make decisions about what new information would be most useful in disambiguating a concept or question.

4.1.5.2 Selected Microprojects/Results

- **Loris Bozzato,** “Reasoning on Contextual Hierarchies via Answer Set Programming with Algebraic Measures”.
- **Christian Müller**, “Uncertainty Handling in Highly Automated Driving: Beyond Data”. The project looked at policies for self-driving vehicles that can extrapolate existing data to deal with uncertainty and approximate the behaviour of human drivers. Expected results include a representation with deep models that can incorporate rules, and two papers

4.1.5.3 Direction of Adjustments/Extensions

This is a highly relevant question that has so far not been sufficiently addressed in the project and needs to be further exploited. It needs to be connected to notion of human oriented trust and explainability (see 3.3) making sure that we define and address uncertainty based not only on objective measure but also based on how it is perceived by users.

4.2 Pillar 2: Multimodal perception and modelling

The aim as stated in the proposal is: “Enabling AI systems to perceive and interpret complex real-world environments, human actions, and interactions situated in such environments and the related emotions, motivations, and social structures. This requires enabling AI systems to build up and maintain comprehensive models that, in their scope and level of sophistication, should strive for more human-like world understanding and include common sense knowledge that captures causality and is grounded in physical reality.”

Specific topics include:

4.2.1 Multimodal interactive learning of models

4.2.1.1 Original Research Goals

We will develop technologies for models that integrate perception from visual, auditory and environmental sensors to provide structural and qualitative descriptions of objects, environments, materials, and processes. Models should make it possible to associate and organize spatio-temporal auditory and visual perception, with the geometric structure of an environment, and the functional and operational properties of objects and structures.

4.2.1.2 Selected Microprojects/Results

- **James Crowley**, "Multimodal Perception and Interaction with Transformers". This micro-project has surveyed tools, data sets, research challenges and performance metrics for experiments in the use of transformers for tasks such as audio-visual narration of scenes, actions and activities, audio-visual gestures, and perception and evocation of engagement, attention, and emotion. Participants have provided a tutorial on the use of transformers for multimodal perception and interaction.

4.2.1.3 Direction of Adjustments/Extensions

Results from the research roadmapping activity indicate that establishing mutual understanding of a situation is an important ability that underlies operational and practical cooperation between people and intelligent systems. Recent results indicate that this problem can be addressed with transformers, for communication from people to systems and for communication from systems to people. We believe that this problem can be addressed using transfer learning using architectures such as BERT or GPT3 with data from the EPIC Kitchens or EGO4D research challenges.

4.2.2 Multimodal perception and narrative description of actions, activities and tasks

4.2.2.1 Original Research Goals

People perceive and understand the world not just as objects and events, but as narratives that situate objects and events within a context and establish causal relationships. Context and causality enable rich descriptions for events that are not directly observable, including hypothetical or abstract events, and events that occurred in the past. Current approaches to action recognition simply detect actions from spatiotemporal signatures and state changes in the environment, without placing the activities in the larger context of an activity or task. We intend to extend this towards more human like, narrative-based description.

4.2.2.2 Selected Microprojects/Results

- **Jan Hajič**, "Multilingual SynSemClass for the Semantic Web (MSSW)". Develops an event-type ontology for actions, activities and states, focused on language interoperability. Such an ontology would serve, together with other ontologies (for entities), for grounding knowledge and situational symbolic representations extracted from input signal(s) (text, speech, visual, …). In turn, these representations together with distributional / neural models will allow for reasoning, constructing responses in dialogue

systems, narrative construction and manipulation, summarization, question answering etc.

4.2.2.3 Direction of Adjustments/Extensions

Recognizing an activity entails representing the activity, together with its environment, timeline, context, etc. with a suitable model, whether neural, statistical or symbolic (or a combination of them), and from the processing point of view, a recognition method combining various inputs (text, speech, visual, haptic, …) and creating such an intermediate representation to use for reasoning, inference, manipulation and other transformations.

4.2.3 Multimodal perception of awareness, emotions, and attitudes

4.2.3.1 Original Research Goals

Going beyond emotions to understanding human intentions, attitudes, and related values is an important topic for psychology, sociology, and philosophy with so far little work within AI. Our approach assumes that comprehensive world models, combined with the ability to seamlessly involve humans in the learning and reasoning process, will be instrumental in addressing this topic. We leverage synergies with the respective activities within HumanE AI Net to develop AI systems that can, at least to a degree, recognize and reason about user motivations, attitudes, and values; meanwhile, the systems' interactions with humans will greatly contribute toward making the vision of a European brand of human-centric AI a reality.

4.2.3.2 Selected Microprojects/Results

- **Mauro Dragoni, "**The knowledgeable and empathic behaviour change coach." This micro-project proposal aims to develop an abstract layer of a conceptual model representing key components relating to maintaining healthy behaviour and supporting behaviour change. The conceptual model will provide a top-level representation of the clinical (from the psychological perspective) enablers and barriers that can be exploited for developing more fine-grained models supporting the realization of behaviour change paths within and across specific domains.
- **Sencer Melih Deniz**, "Neural mechanism in human brain activity during weight lifting". In this project, the change pattern in EEG has been investigated during lifting of different weights and the features in EEG data making difference during lifting a weight has been analysed. Classification between different weights of load cases has been realized by using deep learning-based machine learning methods.

4.2.3.3 Direction of Adjustments/Extensions

We identify as an important research question how do we understand intention and attention of an AI or person.

4.2.4 Perception of social signals and social interaction

4.2.4.1 Original Research Goals

Our goal is to develop methods to endow an artificial agent with the ability to acquire social common sense using the implicit feedback obtained from interaction with people. We believe that such methods can provide a foundation for socially polite HCI, and ultimately for other forms of cognitive abilities.

Knowledge for sociable interaction can be encoded as a network of situations that capture both linguistic and nonverbal interaction cues and proper behavioural responses. Stereotypical social interactions can be represented as trajectories through the situation graph. We will explore methods that start from simple stereotypical situation models and extend a situation graph by adding new situations and splitting existing situations.

4.2.4.2 Selected Microprojects/Results

- No Micro-project has yet been proposed to specifically address this research topic. The recently published Ego4D data set contains data for a research challenge in this area, and we will encourage participants to consider formulating a micro-project in this area using the Ego4D challenge.

4.2.4.3 Direction of Adjustments/Extensions

This remains an important research area that addresses an ability that is important for interaction for collaboration between humans and intelligent systems.

4.2.5 Distributed collaborative perception and modelling

4.2.5.1 Original Research Goals

People have a shared ability to explain observed phenomena and predict future phenomena based not only on direct experience, but on experience learned from others. We need an ability for intelligent systems to learn common sense from experience shared by others. To participate as members of techno-social groups, and engage in collaborative perception and modelling, intelligent systems must be able to represent narratives, understand narratives communicated by other group members, communicate their own knowledge in the form of narratives, and integrate their own narratives with the narratives of other group members

4.2.5.2 Selected Microprojects/Results

- **Andrea Passarella, "**Social AI gossiping". The project aims to understand how to compose, in a fully decentralized AI system, models coming from heterogeneous sources and, in the case of potentially untrustworthy nodes, decide who can be trusted and why. The project focuses on the specific scenario of model "gossiping" for accomplishing a decentralized learning task and on the study of what models emerge from the combination of local models, where combination takes into account the social relationships between the humans associated with the AI.

4.2.5.3 Direction of Adjustments/Extensions

Shared narratives remain a promising approach for research in this area.

4.2.6 Methods for overcoming the difficulty of collecting labelled training data

4.2.6.1 Original Research Goals

Getting sufficiently labelled training data is a core concern for many ML domains. However, for multiple reasons, it is particularly grave when it comes to the perception of complex real-world situations, such as those involving humans, which besides performing actions also engage in social interactions, perceive emotions, and so on.

In the proposal we have broadly stated that we aim to alleviate this problem, focusing at first at data set creation, curation and availability.

4.2.6.2 Selected Microprojects/Results

- **Mathias Ciliberto**, "Collection of datasets tailored for HumanE-AI multimodal perception and modelling". Released an updated version of the Opportunity dataset (OPP++), providing a multi-modal dataset providing anonymised videos and pose data. This advances the development of multi-modal models by providing rich labelled data with actions at different levels. Further work will include a machine learning challenge where not only classification scores will be evaluated, but also explainability as well as sensor data generation from poses.

4.2.6.3 Direction of Adjustments/Extensions

Recent advancements have allowed better sharing of not only datasets, but also trained models. Transfer learning has become more common and methods to bridge the gap between sensor modalities (either combining or translating between them are under development and have taken advantage of big data repositories). Furthermore, we have recently seen a huge advances in Self Supervised Learning (SSL) as well as in Zero (Few) Shot learning methods. While so far we have focused on creating data sets, looking at such methods and adapting them to multimodal perception of the type addressed by the project is a key item on the research agenda. An interesting question is to what degree SSL could be used to create very large "generic" models for various domains of multimodal perception problem (following the lead of NLP with its generic pre-trained language models) together with support for adapting them to various more specific downstream tasks.

4.3 PILLAR 3: HUMAN-AI COLLABORATION AND INTERACTION

The aim as stated in the proposal is "Developing paradigms that allow humans and complex AI systems (including robotic systems and AI-enhanced environments) to interact and collaborate in a way that facilitates synergistic co-working, co-creation and enhancing each other's capabilities. This includes the ability of AI systems to be capable of computational self-awareness (reflexivity) as to functionality and performance, in relation to relevant expectations and needs of their human partners, including transparent, robust adaptation to dynamic open-ended environments and situations. Overall, AI systems must above all become trustworthy partners for human users."

4.3.1 Foundations of human-AI interaction and collaboration

4.3.1.1 Original Research Goals

The project breaks down the study of human-AI relationship into three main types: (i) collaboration, (ii) interaction, and (iii) symbiosis. When studying interaction, we study AI methods that, on the one hand, understand people and can anticipate the consequences of their actions on people, and, on the other, communicate their purposes so as to ground collaboration. This work also involves seeking more natural interfaces to communicate with AI, including multimodal and conversational user interfaces, and augmented reality (AR) interfaces. When studying collaboration, we consider concepts like cooperation, emotional intelligence, collective intelligence, and group cognition. When studying symbiosis, we study emergent properties of AI

systems where people and AI combine their processes, skills, and experiences to achieve something greater together than just by themselves

4.3.1.2 Selected Microprojects/Results

- **Janin Koch**, "Exploring the impact of *agency* on human-computer partnerships". Theoretical and empirical roles of agency in successful human-computer partnerships:
  1) identifying which parameters are relevant to the description of the system agency,
  2) what impact these parameters have on the perceived agency
  3) how to modify them in order to achieve different roles of systems in a process. Achievement: "Younger and Older Adults' Perceptions on Role, Behavior, Goal and Recovery Strategies for Managing Breakdown Situations in Human-Robot Dialogues (https://dl.acm.org/doi/10.1145/3472307.3484679). Follow-up at a Dagstuhl workshop on Human-centred AI (June 2022)
- **Albrecht Schmidt**, "Autobiographical Recall in Virtual Reality (Flo)". Considering VR as an important environment for Human-AI collaboration, the MP explores memories and recall of them generated by VR experiences through physiological parameters. Achievement: Dataset on autobiographic recall in VR and a HumaneAI workshop on Human Memory and AI.
- **Jan Hajič, "**Multilingual Event-Type-Anchored Ontology for Natural Language Understanding (META-O-NLU)". Studies event types -based ontology for NLU, an important topic for human-AI communication and grounding. Demonstrates that adding a language to an ontology for event types is feasible in limited time and resources; generalizes a workflow for other languages. Achievement: SynSemClass database extended by 1500+ German verbs denoting event types and states (version 4 to by published by end of April 2022), paper at LREC 2022
- **François Yvon**, "Evaluating segmentation in automatic captioning systems". Segmentation is important aspect of human-AI understanding in NLI. This project evaluates the quality of the output segmentation, where decisions regarding the length, disposition and display duration of the caption need to be taken, all having a direct impact on the acceptability and readability. Results: Survey of existing segmentation metric, Design of a contrastive evaluation set, Comparison of metrics on multiple languages / tasks.
- **Mohamed Chetouani**, "Proactive communication in social robots: Develop Proactive Communication Models for Social Robots". Achievement: Integrated system that can generate proactive robot behaviour by reasoning on both factors: intentions and predictions. Journal paper to be submitted.

4.3.1.3 Direction of Adjustments/Extensions

At the Venice meeting (in April 2022) we refined this topic towards "Cognitive foundations of human-AI collaboration". The idea is that enabling Human-AI collaboration requires shared understanding of the situations, intentions, responsibilities, and values of interactions. Taking a starting point in theories on human behaviour developed in fields such as cognitive psychology, social psychology, social science, a new foundation for socially intelligent, interactive systems will be developed. Key objectives are:

1. Experiments to evaluate the validity of human behavior theories in Human-AI collaboration

2. New theories and interaction approaches to improve understanding within the different levels of human-AI collaboration
3. Explore causal models that link behavior with cognitive, emotional, and other latent factors to be used for inferring, predicting, planning, and acting without extensive data on an individual (first impression)

From these scientific opportunities at the intersection of AI and HCI arise to study the emergence of "grounding" in interaction arises, in order to understand how people and AI can adapt their behaviour when interacting with each other and evaluate the validity of Theory of mind, and intersubjectivity as emergent properties of interaction.

4.3.2 Human-AI interaction and collaboration

4.3.2.1 Original Research Goals

Given a basic understanding of the way humans approach AI systems, concrete interaction paradigms must be developed. Furthermore, for humans and AI to be able to collaborate toward common goals, they must be able to interact and *understand* each other, establish common ground, and see the other's perspective (thus having a type of Theory of Mind).

4.3.2.2 Selected Microprojects/Results

- **Brian Ravenet,** "Interactive Reinforcement Learning for Humorous Agents". This microproject aims to enrich conversational agents with a humour model. The expected results include a novel humour model, an online game to playfully gather the necessary training data on humorous interactions from users as well as a publication in an AI or AI in games conference or journal. Achievements: Humour models for conversational agents; Paper in International Conference of Journal related to AI and AI in Games
- **Patrick Paroubek, "**DIASER: DIAlog task-oriented annotations for enhanced modeling of uSER." This microproject aims to evaluate the usefulness of current dialog dataset annotations, and to improve on them through explicit user representations, improved annotation consistency, and unification of annotations from multiple datasets. Achievements: A corpus of 37,173 annotated dialogues with unified and enhanced annotations was built from existing open dialogue resources; A paper was accepted at the TALN2021 conference: "Defining and Detecting Inconsistent System Behaviour in Task-oriented Dialogues", Another paper to be submitted to the "Dialogue and Discourse" journal.
- **Antti Oulasvirta, "**Optimal Alerting". A new way to decide when to alert a human user (e.g., driver), taking into account his beliefs, capabilities, and the external situation. Theoretically this will be based on POSG (partially observable stochastic game) developed together with UMPC/CNSR (Gori). Ongoing.

4.3.2.3 Direction of Adjustments/Extensions

We aim to shift the research focus towards "Improving Human-AI Interaction and Collaboration" which is a variation of the original vision. Thus, in order to turn the AI from a passive tool to an active collaborator concrete interaction paradigms must be developed. Depending on the context of use and the common goal of the collaboration, this calls for means to personalize the AI and adapt the interaction techniques and underlying modalities to the situation of the user.

4.3.3 Reflexivity and adaptation in human-AI collaboration

4.3.3.1 Original Research Goals

Our work will entail methods for meta-reasoning between the human and AI system, where they can ask together or to each other “Are we doing the right thing?” or “Is it ethical what we are suggesting?” On the interaction side, our goal is to enhance reflection by having a small dialogue at particular times. Often AI systems are developed to advise or suggest without the opportunity for negotiation or understanding. A recent suggestion is that AI systems should explain their decisions. Our work will develop solutions that determine what to ask and when and how, which at the machine-learning side will combine aspects of active learning, sequential planning, and reasoning

4.3.3.2 Selected Microprojects/Results

- **Mireia Diez Sanchez**, “Adaption of ASR for Impaired Speech with minimum resources (AdAIS)”. This micro-project makes AI based systems more accessible for impaired people. It focuses on the speech recognition part. Achievement: ASR systems were adapted for speech from subjects with dysarthria speech impairment of various degrees. German data comprising only 130 hours of untranscribed doctor-patient German speech conversations.
- **Gilles Bailly**, “Learning Individual Users’ Strategies for Adaptive UIs”. This microproject presented a new model-based approach to predict how an individual user will react to an adaptation of a UI, in particular by taking into account the user’s level of experience with the UI. The model allows adaptive UIs to take more graceful (less disruptive, effort-causing) changes. The benefit was empirically demonstrated in a novel model-based deep RL method for adaptive menu systems. Achievements: CHI’21 full paper. Code and data are released.

4.3.3.3 Direction of Adjustments/Extensions

Reflexivity needs to connect to the topic and task at hand, and relevance. How to respond appropriately: socially, contentwise, “emotionally”, and assess whether past responses were appropriate. Response adapted to the human, co-adapted with the human. Based on goals, intentions, responsibilities, human capabilities. Key objectives from that perspective are

1. Methodology itself: co-created routine where the agent learns with human guidance
2. Trustworthiness: frame and communicate the reasons for taking action, monitor alternative, next actions in case of increasing emergency
3. Creating mutual understanding in a situation (if misunderstanding or disagreement occurs)

4.3.4 User models and interaction history

4.3.4.1 Original Research Goals

We here pursue two important capabilities that user models should have: (1) *forward modelling or* providing a richer and more generalizable account of human behaviour suitable for real-world interactive AI, which has been an issue in cognitive and user models for decades, and (2) *inverse modelling,* or fitting models to individual users. Both are needed for deployment in interactive AI, that must on the one hand update its

model representations with interactions and, on the other, select actions while anticipating their consequences on users (counterfactuality). In addition, the research will develop interaction history trails that can: (1) keep a record of previous encounters so that they can be referred to in subsequent interactions between the users and the AI system and (2) decide on what should be forgotten in a human-AI encounter or interactions (ethically, legally, and morally, to stay feasible).

4.3.4.2 Selected Microprojects/Results

- **Patrizia Fattori,** "Prediction of static and perturbed reach goals from movement kinematics Movements to select 3D targets require the integration of different sensory information". This micro-project aimed to investigate at which point of the movement the final target position can be predicted and whether the accuracy to predict horizontal or sagittal dimensions differ. Achievements: data set of individual movement trajectories + manuscript in preparation + code of recurrent neural network
- **Richard Benjamins, "**Improving air quality in large cities using mobile phone and IoT data". Goal: combining mobility data, mobile phone data, IoT pollution and climate sensor data and Open data to provide actionable insights which can be used in mobility and pollution policy making Achievements: prototype, video, and business and government presentations.
- **Daniel Weimer, "**Connected vehicle simulation for AI-based applications". This microproject aims at creating a simulator for connected vehicles in a realistic traffic environment. This simulator serves as a GDPR-compliant starting point for AI-based connected vehicle applications such as parking spots occupancy prediction based on vehicle data. Ongoing.
- **Virginia Dignum,** "Human-machine collaboration for content analysis in context of Ukranian war". Ongoing.

4.3.4.3 Direction of Adjustments/Extensions

A possible extension of this research area is towards "Human knowledge models and their use". The objective is to study models that can represent human knowledge, their use and adaptation, in particular, to human-AI interaction. One example could be a big pre-trained language model created from all the text the humans ever generated (using web crawled data) and its fine tuning to topic focused dialog systems. Another would be scene characterization model (can be from video, audio, …) and its use for world understanding, in assistants. This area has a strong connection to WP1 as it will involve many questions such as the representation of visual, voice, textual, physical (etc.) data by neural network based embeddings and by symbolic representations (incl. grounding), merging different modality of models to a higher level understanding, language-independent and/or multilanguage models (both distributional/neural and symbolic or a combination), sharing of context among models for different modalities (person id from voice and video, gender recognition from voice and video. Its use in dialog system) and heavy use of transfer learning and large data sets.

4.3.5 Visualization interactions and guidance

Visualization remains an important aspect of interaction between humans and complex systems. Visual analytics (VA) supports the information-discovery process by combining analytical methods (from data mining to knowledge discovery) with interactive visual means to enable humans to engage in an active "analytical discourse" with their datasets. However, for humans/users, who are usually experts in their application domains but not in VA, it is difficult to determine which VA methods to use for particular data and tasks. Guidance is needed to assist humans/users in

selecting appropriate visual means and interaction techniques, using analytical methods, and configuring instantiation of these algorithms with suitable parameter settings and combinations thereof. After a VA method and parameters are selected, guidance is also needed to explore the data, identify interesting data nuggets and findings, and collect and group insights to explore high-level hypotheses, and gain new knowledge.

4.3.5.1 Selected Microprojects/Results

None applicable.

4.3.5.2 Direction of Adjustments/Extensions

We would like to move the research towards “Guiding data exploration using visual human-AI interactions”. Objectives include new visualizations to explore and guide users to better analyse large datasets, rethinking the form of visualization, i.e., expanding the traditional 2D form towards a 3D representation of visualizations, defining and evaluating new interaction techniques using diverse modalities to explore large datasets, evaluating instantiation of assisting algorithms with professionals and novices users and defining the meaning of guidance in human-AI interaction, e.g., recommender systems vs. more exploratory forms of guidance. Scenarios can be found in Finance, Health and many other areas.

4.3.6 Trustworthy social and sociable interaction

4.3.6.1 Original Research Goals

Current systems lack ability for social interaction because they are unable to perceive and understand humans, human awareness, and intentions, and to learn from interaction with humans. Building on the research on the perception of human emotions the modelling of social context and complex, evolving world models we will address key challenges in enabling AI systems to act appropriately within complex social contexts. A second important issue is that the AI systems, when interacting with one or more persons (and possibly other autonomous AI systems), should consider the broader social context in which they interact. For instance, an e-health system should not recommend taking a walk at dinnertime as the whole family gets to the table. It should be aware of practices, narratives, norms, and conventions to fit the interaction within those structures.

4.3.6.2 Selected Microprojects/Results

- **András Lőrincz**. Machine supervision of human activity: The example of rehabilitation exercises Technology review through a given scenario which was physical rehabilitation, requiring body pose estimation and dialog regarding error correction and pain. Achievement: Paper published at ICANN 2022.
- **Aart van Halteren,** Conversational AI for patient reported outcomes This project aimed to investigate how conversational AI agents could help to improve cohesion in virtual team meetings. Specifically, the aim was to investigate how a person's emotion, personality, relationship to fellow teammates, goal and position in the meeting influences how they remember the meeting. Achievement: Dataset and publication prepared.

4.3.6.3 Direction of Adjustments/Extensions

This research line needs to be connected to the definition of human centric trustworthiness and explainability as described in 3.3.

4.3.7 New Potential Directions

We have summarized the above adjustment points within the larger theme of common ground. *Common ground* refers to an understanding of an activity shared between collaborative partners that promotes success in the activity. Our approach in this WP is that common ground *emerges*, i.e., it is established *interactively* in a dynamic process between collaborative partners and is affected by not only overt actions of the partners but also beliefs and other latent factors. The new objective of this pillar is to study and develop techniques and methods that facilitate the joint achievement of common ground in collaborative interactions.

Sub-objective 1: Psychological factors related to common ground, including cognitive (e.g., beliefs, theory of mind, trust) and social psychological (e.g., self-presentation, emotional expression, …).

Sub-objective 2: The design of user interfaces and "embodiments" (e.g., avatars) that promote common ground (e.g., design guidelines, dialogues etc)

Sub-objective 3: The design of communicative capabilities of AI agents to promote the establishment of common ground in collaborative tasks; includes explainable AI, but with the clear purpose of promoting common ground in a shared activity.

Sub-objective 4: Interaction techniques for learning from and teaching a collaborative partner.

Sub-objective 5: "Common sense" and other representational approaches to model knowledge relevant for common ground in collaborative tasks.

4.4 Pillar 4: Societal awareness

Being able to model and understand the consequences of complex network effects in large-scale mixed communities of humans and AI systems interacting over various temporal and spatial scales. This includes the ability to balance requirements related to individual users and the common good and societal concerns. Specific topics include:

This pillar strives to shape the research on the societal dimension of AI, as increasingly complex socio-technical systems emerge, made by interacting people and intelligent agents. As increasingly complex **AI-influenced socio-technical systems** (AI-STS) emerge, made of many interacting people, algorithms, and machines, the social dimension of AI becomes evident. Examples include:

- mobility, with travellers helped by smart assistants to reach their destinations
- public discourse and social media,
- electronic markets, where the diffusion of opinions and economic decisions are shaped by personalized recommendation systems and targeted advertising.

AI-assistants and recommender systems are designed to help individual users cope with information load. The problem is that a crowd of individually "intelligent" people and machines is not necessarily a socially "intelligent" crowd. On the contrary, it can be stupid in many cases, due to collective effects and emergent phenomena. The sum of many individually "optimal" choices is often not collectively beneficial, because individual choices interact and influence each other, on top of common shared

resources: e.g., instability, traffic congestion, pollution, misinformation, and polarization.

For example, it was shown that in e-markets, collaborative filtering algorithms increase individual diversity but at the same decrease collective diversity. The recommender systems cause individuals to discover and buy a greater variety of products, but each individual is pushed to purchase the same set of popular titles, leading to concentration bias at the aggregate level. Navigation systems suggest directions that make sense from an individual perspective but may exacerbate congestion if too many drivers are directed on the same route. Personalized recommendations on social media often make sense to the user, but may artificially amplify polarization, echo chambers, filter bubbles, and radicalization.

Based on these examples, the key observation from complex systems science is that the sum of many individually "optimal" choices is often collectively suboptimal, because individual choices interact and influence each other, on top of common resources.

The emergent phenomena and collective effects of AI-STS and their impact on society are not sufficiently addressed by AI research. This goal requires a step ahead in the trans-disciplinary integration of AI with network and complexity science and (computational) social science.

Key general questions:

- How to model and understand the aggregated outcomes? How to bridge the micro-level effects (users' choices and AI suggestions) with the macro-level effects.
- How to use such modeling and understanding to explore alternative mechanisms and architectures of AI-STS
- Within ethical and legal frameworks and public policy that sets the goals, how to design AI mechanisms that help AI-STS to evolve towards such agreed collective outcomes, e.g.,
    - sustainable mobility in cities,
    - diversity and pluralism in the public debate,
    - fair distribution of resources
- Caveat: beware of techno-solutionism!
- What science can stem from the combination of complexity, (computational) social science and AI?

Specific topics include what follows.

4.4.1 Gray box models of society scale, networked hybrid human-AI systems

4.4.1.1 Original Research Goals

The general challenge is to characterize how the individual interactions of individuals, both humans and AI systems, with their own local models, as well as the social relationships between individuals, impact the outcome of AI models globally and collectively. Using a combination of machine learning, data mining, and complexity theory, we strive at understanding the networked effects of many distributed AI systems interacting together, some (or all) possibly representing human users, therefore comprising a complex human and technical ecosystem. The different layers

of this system are in mutual interaction, producing emergent phenomena which may range from synchronization to collapse.

4.4.1.2 Selected Microprojects/Results

- **Pierluigi Contucci,** *Agent-based modeling of the Human-AI ecosystem*. UNIBO, CEU

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. Achievements: Code: Simulation of delegation of information processing in techno-social groups. Publication: "Human-AI ecosystem with abrupt changes as a function of the composition" P. Contucci, J.Kertesz, G. Osabutey. Accepted to PlosOne https://arxiv.org/pdf/2204.03372.pdf

- **Giulio Rossetti,** *Algorithmic bias and media effects.* CNR, CEU, UNIPI

Description: The project investigates polarization in OSN. The plan is to enhance a previous 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, a study on whether moderate media can be effective in counteracting polarisation is conducted. Achievements: Code: https://github.com/GiulioRossetti/AlgorithmicBias Publication: in preparation.

4.4.1.3 Direction of Adjustments/Extensions

We largely see the goals to remain valid with a stronger focus on Models of the dynamics of social AI processes and feedback loops taking into account the fundamental properties of complex networks (connectedness, clustering, hubs) combined with different forms of AI influence to develop realistic models of aggregated behaviour. A core question is how to combine model-theoretic and data-driven empirical research linking to the goals of WP1 with respect to combining symbolic and sub-symbolic models.

4.4.2 AI systems' individual versus collective goals

4.4.2.1 Original Research Goals

Social dilemmas occur when there is a conflict between the individual and public interests. Such problems may appear also in the ecosystem of distributed AI and humans with additional difficulties due to the relative rigidity of the trained AI system on the one hand and the necessity to achieve social benefit and keep the individuals interested on the other hand. What are the principles and solutions for individual versus social optimization using AI and how can an optimum balance be achieved?

4.4.2.2 Selected Microprojects/Results

- **Jesus Cerquides,** ***"****A tale of two consensus. Building consensus in collaborative and self-interested scenarios".* The aim of the project is to learn a representation with deep models in a way to incorporate rules (e.g., physics equations governing dynamics of the autonomous vehicle) or distributions that can be simply defined by humans in advance. The learned representations from the source domain (the domain whose samples are based on the defined equations/distributions) are then transferred to the target domain with different distributions/rules and the model adapts itself by including target-specific features that can best explain variations of target samples w.r.t. underlying

source rules/distributions. In this way, human knowledge is considered implicitly in the feature space.

- **Mirco Nanni, "**Network effects of mobility navigation systems". The aim of the projects is to study emergent collective phenomena at the metropolitan level in personal navigation assistance systems with different recommendation policies, with respect to different collective optimization criteria (fluidity of traffic, safety risks, environmental sustainability, urban segregation, response to emergencies) Expected outcomes: Code: (Big-) data-driven simulations with scenario assessment. Publication: a scientific paper
- **Frank Dignum.** "Normative behaviour and extremism in Facebook groups". This project investigates whether normative behaviour can be detected in Facebook groups. In a first step, will hypothesize about possible norms that could lead to a group becoming more extreme on social media, or whether groups that become more extreme will develop certain norms that distinguish them from other groups and that could be detected. Simulations and analyses of historical Facebook data (using manual detection in specific case studies and more broadly through NLP) will help reveal the existence of normative behaviour and its potential change over time. Outcomes: Identification of radical behaviour in Parler groups. Characterizing the language use of radicalized communities detected on Parler
- **Frank Dignum. "**Socially aware interactions". The project uses the social context for effective and focused dialogue, geared towards a specific goal that is accepted by all parties in the interactions. The plan is to start with the Dialogue Trainer system that allows for authoring very simple but directed dialogues to train (medical) students to have effective conversations with patients and based on this tool, the mission is to design a system that will actually deliberate about the social context. Outcomes:cCode: Prototype of dialogue system https://github.com/GAIPS/socially-aware-interactions Publication: Socially Aware Interactions: From Dialogue Trees to Natural Language Dialogue Systems. I. Lobo, D. Rato, R. Prada, F. Dignum https://link.springer.com/chapter/10.1007/978-3-030-94890-0_8.
- **Frank Dignum**, "Social interactions with robots". The project shows how social practices can be used to guide human-robot interactions. This provides a social context that can be helpful to adapt the actions of the robot to both the situation and the user. The project was a very first attempt to create a practical implementation and thus can only be seen as a basis on which further work can be done to really take advantage of all aspects of social practices. Achievements: AI Planning with Social Practices for the Pepper robot.
- **Jennifer Renoux**, *"Social dilemma with information asymmetry."* The study carried out during the micro-project will give insight into how an artificial agent may influence a human's behavior in a social-dilemma context, thus allowing for informed design and development of such artificial agent. In addition, the platform developed will be made available publicly, allowing future researchers to experiment with other configurations and other types of feedback. By using a well-development and consistent platform, the results of different studies will be more easily comparable. Achievements: Code: The Pest Control Game experimental platform https://github.com/jrenoux/humane-ai-sdia.git Publication: in preparation.

4.4.2.3 Direction of Adjustments/Extensions

The research direction remains very current and valid. Mode attention should be given to understanding the effects on individual vs collective balance under different AI

mechanisms and the question of how Distributed, Federated, and Decentralized AI (learning, reasoning) influences individual vs collective balance.

4.4.3 Societal impact of AI systems

4.4.3.1 Original Research Goals

How to evaluate the societal impact of competing AI technologies and promote the ones more compliant with the European values? develop AI systems that contribute to improving the quality of and access to information, deal with information noise and fake news, detect and counter manipulation, and deal with information overload What are the possibilities, the risks and the impact of AI on governance, considering the opportunities of AI-assisted participatory technologies? How to understand and model strategies with which AI can enhance public involvement, help foresee social consequences of policies, and facilitate social adaptability to change? How can AI contribute to the handling of the conflict between the different time scales of individual interests, legislation periods, and the solution of global problems?

4.4.3.2 Selected Microprojects/Results

- **Michel Klein,** "Evidence-based chatbot interaction aimed at reducing sedentary behaviour." The project investigates how AI systems can collaborate with humans, specifically focusing on changing a specific behavior. It increases our understanding of how specific interaction forms between an AI system and a human are effective in achieving behavior change. It also investigates to what extent knowledge about health behavior can contribute to designing realistic and effective communication. Achievements: Publication: bachelor thesis
- **Eugenia Polizzi, "**Using Social Norms to counteract misinformation in online communities". The project investigates the role of social norms on misinformation in online communities. Top-down "debunking" interventions have been applied to limit the spread of fake news, but so far with limited power. Recognizing the role of social norms in the context of the misinformation fight may offer a novel approach to solving such a challenge, shifting to bottom-up solutions that help people to correct misperceptions about how widely certain opinions are truly held. This knowledge can help identify new interventions in online communities that help prevent the spread of misinformation. Achievements: Publication: "The voice of few, the opinions of many: evidence of social biases in Twitter COVID-19 fake news sharing" di P. Castioni, G. Andrighetto, R. Gallotti, E. Polizzi, M. De Domenico https://arxiv.org/abs/2112.01304
- **Laura Sartori, "**What idea of AI? Social and public perception of AI." The project addresses issues related to social trust, cohesion, and public perception. It has clarified how and to what degree Ai is accepted by the general public and highlighted the different levels of public acceptance of AI across social groups. Achievements: Publications: A sociotechnical perspective for the future of AI: narratives, inequalities, and human control, in Ethics and Information technology". L. Sartori, Laura, T. Andreas. 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

4.4.3.3 Direction of Adjustments/Extensions

The topic remains of high interest to the project. In the future specific directions to be given more attention include value-driven, responsible AI design, development and deployment of society-scale, complex human-AI systems, co-design that takes into account different stakeholders (companies, users, regulators, …) whose interests may collide, and compromises need to be sought (Pareto-optimality) and the role of narratives, design fictions, realistic simulations, sandboxes for anticipating intended and unintended impacts.

4.4.4 Self-organized, socially distributed information processing in AI-based techno-social systems

4.4.4.1 Original Research Goals

Understand how to optimize distributed information processing in techno-social systems and what are the corresponding rules of delegating information processing to specific members (AI or human). The ultimate goal is to develop and enhance distributed information processing in socio-technical systems so that they provide a platform for common action. To this end, we will study the mechanism of self-organization in socio-technical groups at different scales from common action, e.g., in emergency response to societal movements. In this context, it is also important to understand how to achieve robustness of the human-AI ecosystems with respect to various types of malicious behaviour, such as abuse of power and exploitation of AI technical weaknesses. Ultimately, we will develop principles for designing schemes of AI systems that are robust or resilient to manipulation and are at the same time incentive compatible.

4.4.4.2 Selected Microprojects/Results

- **Paolo Ferragina, "**Pluralistic Recommendation in News". The project aims at designing a Recommender System able to foster pluralistic viewpoints in news pieces suggestions in the European domain. In this scenario, a crucial point of the project is to build a dataset with millions of European news articles labelled by their political leaning, popularity, and distribution area. Additionally, the second point is to define a topic modelling algorithm and a multilingual classifier able to identify the main topics and the political leaning of each article by leveraging AI-based techniques for NLP. Achievements: Dataset: European News with political bias with metadata (around 16 million articles) https://drive.google.com/file/d/1Qq2khT7IM-5_oHSNJhbK_-EATNdOSY-n/view. European News with political bias with metadata plus topic annotation for each article (around 4 million articles) https://drive.google.com/file/d/1KGy-FcLulACK_Fa3Abd9Xr4qaurBnu2S/view

4.4.4.3 Direction of Adjustments/Extensions

This too remains a relevant line of research. Focus in the next period should be given to the following questions:

- How to organize decentralized human-machine collaboration and collaborative decision making in society-scale, complex human-AI systems
- How to design society-scale, complex human-AI systems architectures that adapt in an efficient way to respond to endogenous and exogenous changes
- How trust evolves in self-organizing decentralized society-scale, complex human-AI systems
- How narratives can be used to foster coordination and action
- How to design socially distributed information processing and knowledge management

4.5 Pillar 5: Legal and ethical bases for responsible AI

HumanE AI Net efforts on the legal and ethical bases for responsible AI were set up in the proposal as:

- Identification and development of tools for engineering, design and monitoring of AI systems according to societal values and human rights principles. (Aligned with *RQ2)*
- Investigate the impact of using AI tools on the sectors where they are deployed (e.g. medicine, economics) and on the practitioners, users and other stakeholders using or affected by them.

The overall goal is to boost research aimed at developing methods and methodological guidelines for the entire lifecycle of the AI system: design, field validation with stakeholders (simulations, sandbox), deployment and feedback through continuous oversight. This will include:

- Ensuring that design processes result in systems that are robust, accountable, explainable, responsible and transparent
- Ethics for designers: and making sure that those developing AI systems are aware of their role and impact on the values and capabilities of those systems
- Methods to elicit and align multi-stakeholder values and interest and constraints capable of balancing societal and individual values and rights
- Methods to integrate and validate a combination of different possibly conflicting values (Design for Values), describe dilemmas and priorities, and integrate them into the computational solutions
- Compliance with laws and regulation and with guidelines for ethical AI
- Explainable AI systems in support of high-stakes decision making (e.g., in health, justice, job screening)
- Feedback methods to inform policy makers and regulators on missing elements in current regulations and practices.

In general, the work is still aligned with these objectives. However, given the complex nature of these goals, it becomes clear that any considerations regarding the design processes and the behaviour (or results/effect) of AI applications, the

ethical aspects and their societal impact – including different social perspectives, such as economic, cultural, gender – need to be considered together both during design and development as well as in the use and behaviour of the systems. At the same time, clear guidance with regard to Legal Protection by Design is needed. Moreover, it is necessary to investigate the impact of the AI systems on society, both from a global, society-wide perspective, as well as from the perspective of collectives (groups) and from the perspective of individuals (humans).

4.5.1 Legal Protection by Design (LPbD)

4.5.1.1 Original project goals

Where it concerns our work on legal protection by design (LPbD), our work is divided into two tasks that

- address the question of incorporation of fundamental rights protection into the architecture of AI systems including (1) checks and balances of the Rule of Law and (2) requirements imposed by positive law that elaborates fundamental rights protection (**T5.1**), and
- engage in a kind of constructive technology assessment, interacting with the developers of the projects, teasing out potential risks for the rights and freedoms of natural persons who may suffer the consequences of implementation (**T5.2**).

4.5.1.2 Selected Microprojects/Results

In collaboration with WP2 on microproject Data Curation in Machine Learning (DCML), focusing on the legal framework in the HAI research context (VUB). The report on DCML should be finalised before the summer, e.g., providing guidance on how to deal with existing datasets that may have been collected unlawfully. The aim of our research into the MP on 'Collection of datasets tailored for HumanE-AI multimodal perception and modelling' is to investigate the relevant legal framework of Data Curation for Machine Learning (DCML), notably with regard to data protection. These concern (1) the legal framework for the HAI-NET research context and (2) potential risks to fundamental rights and freedoms in downstream use cases. This should feed into a series of design choices that enable Legal Protection by Design. The first objective is being finalised at this very moment in the form of a report.

4.5.1.3 Direction Adjustments/Extensions

Our work on LPbD is proceeding along the lines set up in the proposal, noting that it requires taking into account two different contexts: the HAI-NET research context and potential deployment contexts. The first requires an assessment the applicable legal framework to the MPs research design, the second requires anticipating the application of the legal framework to reasonably foreseeable use cases.

The LPbD tutorial materials will be made available on the website or the AI4Europe once the AI Act proposal has been consolidated. As results of these tasks, we are currently finalising two publications:

- Chapter on Legal Protection by Design in upcoming LNCS publication (based on a Tutorial in the ACAI2021 advanced course on AI on Human Centred AI, see below)
- Report Data Curation for Machine Learning (DCML) and the GDPR: Includes an analysis of the EU data protection framework for data curation, a case study

on WP2 data curation micro project and prepares for the development of Data Protection by Design principles in the curation of data for Machine Learning.

4.5.2 ELSEC considerations in AI development and use

4.5.2.1 Original project goals

The other main building blocks of pilar 5 concerns the so called 'by design' (**T5.3**) and 'in design' (**T5.4**) approaches, where the first mostly focus on ensuring that the results, or behaviour, of the system is guaranteed to meet ethical and/or societal principles (e.g. 'ethics by design', 'legal protection by design' or 'privacy by design'), whereas the later focus on the design, engineering and deployment processes and the role of stakeholders, and how these take into account democratic, human rights and sustainability principles.

4.5.2.2 Selected Microprojects/Results

- **Jonne Maas**: "The role of designers regarding AI design: a case study". The purpose of this micro-project was to critically reflect on the design of an AI system by investigating the role of the designer. Designers make choices during the design of the system. Analysing these choices and their effective consequences contributes to an overall understanding of the situated knowledge embedded in a system. The reflection concerned questions like "What does it mean for the output of the system what the designer's interpretations are?" "In what way do they then exercise power on this system?" In particular, this micro-project examined a concrete case. It followed the design of an agent-based social simulation that aims at modelling how inequality affects democracy. As tangible outputs we produced an agent-based simulation on how wealth inequality affects political relations and a conference paper critically reflecting on the design of the simulation. Achievements: An agent-based simulation on how wealth inequality affects political relations. Video: The Role of an AI Designer: design choices and their epistemic and moral limitations (see https://play.umu.se/media/t/0_kjvslv8f/411923). More information: https://www.ai4europe.eu/research/research-bundles/role-designers-regarding-ai-design-case-study
- **Dilhan Thilakarathne**: "Validating fairness property in post-processing vs in-processing system". After choosing a formal definition of fairness (we limit ourselves with definitions based on group fairness through equal resources or equal opportunities), one can attain fairness on the basis of this definition in two ways: directly incorporating the chosen definition into the algorithm through in-processing (as another constraint besides the usual error minimization; or using adversarial learning etc.) or introducing an additional layer to the pipeline through post-processing (considering the model as a black-box and focusing on its inputs and predictions to alter the decision boundary approximating the ideal fair outcomes, e.g. using a Glass-Box methodology). We aim to compare both approaches, providing guidance on how best to incorporate fairness definitions into the design pipeline, focusing on the following research questions: Is there any qualitative difference between fairness acquired

through in-processing and fairness attained by post-processing? What are the advantages of each method (e.g. performance, amenability to different fairness definitions)? Achievements: A paper: https://doi.org/10.1007/s10676-022-09636-z More information: https://www.ai4europe.eu/research/research-bundles/validating-fairness-property-post-processing-vs-processing-systems

- **Laura Sartori**: "What idea of AI? Social and public perception of AI**"**. The project aimed to cultivate a multidisciplinary dialogue into AI technologies and the narratives that go around their development, deployment, and usage. The project achieved this aim by producing and advancing a whitepaper. In addition, a pilot survey had been launched to gather naïve users' perception of AI technologies – results not published yet. Achievements: A Sartori, L., Theodorou, A. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics Inf. Technol. 24, 4 (2022). https://doi.org/10.1007/s10676-022-09624-3More information: https://www.ai4europe.eu/research/research-bundles/what-idea-ai-social-and-public-perception-ai
- **Bettina Fazzinga**: "Ethical Chatbots". The purpose of this project is to develop an architecture for AI dialogue systems that is "ethical" in the sense that it is inspired by ethical principles and values. In particular, the architecture combines NLP techniques and computational argumentation to ensure that user data are ethically managed, the reasoning process is consistent, and the answer is explainable. To illustrate the system, we focused on a case study regarding COVID-19 vaccine information.
  Achievements: CLAR Conference Paper, NL4AI Conference Paper . More information: https://www.ai4europe.eu/research/research-bundles/ethical-chatbots

4.5.2.3 Direction of Adjustments/Extensions

With respect to this work, it has become clear that it is necessary to incentivize efforts to integrate both lines of work into a comprehensive framework integrating in- and by-design considerations, which can account for both the identification of social benefit as well as a way to ensure the empowerment of individuals (users, practitioners, …). This should start from Question Zero, that is, the question of whether specific AI systems should be developed and/or deployed, a question that is of key importance in the light of potential infringements of fundamental rights and freedoms.

4.5.3 Support of RRIA and Consolidation as well as coordination of the research agenda

4.5.3.1 Original project goals

WP5 provides ethical and legal support of the RRIA tasks of the other research WPs. This included a 2-day tutorial (VUB) attended by a senior researcher of each partner.

WP5 is also implementing the consolidation and coordination function for the research agenda.

4.5.3.2 Selected Microprojects/Results

In the line of work, no explicit micro-projects were set up, but the current results, described in the remained of this section are originated from several discussions across the members of WP5, in some cases extended to WP4 and WP3.

Research is needed on philosophical frameworks on public trust, including the development of empirical mechanisms to assess said trust (e.g., Public Technology Assessment), such that potential simulations positive and negative impact of the same system can be understood vis a vis differences in perception and effect metrics.

In this line, an important research question is that of identifying who is affected, how large should the temporal, situational and network analysis of the effects of a system be, and who/how can a line be drawn? This is not only a relevant research question but at the same time identifies an important power structure, with associated power imbalances.

A related question concerns the decision of how many (moral, epistemic) values should be included in the development of an AI system? Whereas it is widely recognised that effect needs be accessed (and designed) from multiple perspectives, it is also important to deal with the hardness of over constrained design spaces (‘we cannot have it all’). Again here, the political power to decide on the values to be included directly affects the design options and resulting systems. Approaches to systematically deal with such engineering complexity are largely lacking at the moment.

Towards this issue, and in collaboration with WP4, we are starting work on a novel conceptualisation of AI systems and impact that will provide a shared basis for dialog across disciplines, application sectors and social actors.

This conceptualisation, illustrated by the figure below, is based on three dimensions, and when finalised, will provide the means to characterise the differences and similarities of systems at the intersection of these dimensions, noting that whatever the dimension, legal imperatives must be needed:

- Societal considerations
  - Including ethical, social, economic, cultural, gender dimensions
- Types of stakeholders
  - E.g., Single user or Multiple users, owner or deployer, society at large
  - Development and use cycle, including the stages
  - Engineer (algorithms, interfaces, datasets)
  - Learn (training, setup and adaptation)
  - Act (result of system’s action or decision)
  - Effect (perceived consequences of system actions)
  - Assess (evaluation - external or internal - of the effects of the system)
  - Feedback loop (from assess back to engineer…)

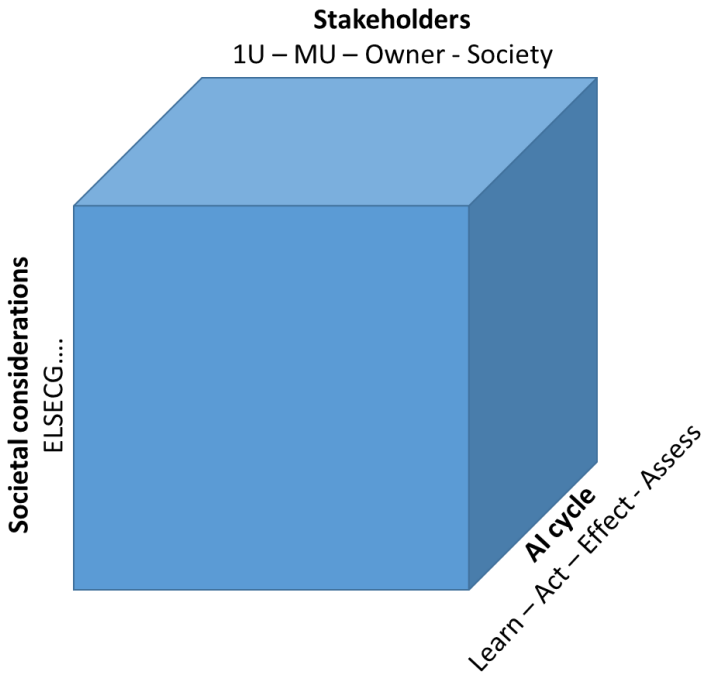

Figure 2 Conceptualization of AI systems.

In parallel, we continue the current work on Legal Protection by Design, including the assessment of the applicable legal framework, in terms of relevant actors, such as data controllers/processors/providers/users/data subjects/affected natural persons, in terms of relevant design decisions (including curation of training datasets, modelling, output) and potential use cases. Key attention to fundamental rights impact assessment (FRIA) and to the development of a set of legal protection by design strategies.

Part of the implementation of this renewed research agenda shows the following research collaborations (mini/micro projects).

- a humane AI approach to misinformation detection in the context of news reporting on the Ukrainian war (based on Russian, Ukrainian, English news sites)
- analysis of the impact of AI on self-determination and democracy through a critical review of election recommender systems
- investigate the social, legal and ethical aspects underlying the design of an AI system (in particular of data-driven machine learning systems) that is under development in one of the other WPs or by one of the industrial members of the HumanE AI Net consortium.

4.5.3.2 Direction of Adjustments/Extensions

An interesting future direction is to go from just ethics, to social etc. Responsibility-in-design? Questions include how to understand, design and develop socio-technical systems that are socially aware and non-discriminatory (connection to WP4!) the role of AI/Human on the cycle. These directions will be developed in the following new mini projects, consolidating and integrating results from the past micro-projects.

Mini project 1: A human-AI approach to misinformation detection in the context of news reporting on the Ukrainian war (based on Russian, Ukrainian, English news sites)

Description: Currently, information that is received by a huge amount of people at the same time from various Internet sources can have a significant impact on important political, economic and social events in different regions, countries and even the world. Automatic or even manual verification of the information for realism and trustworthiness is a very complicated process, especially if there is a lot of the same misinformation in various sources. The mini project aims to produce recommendations and guidelines for validation whether textual information in news websites is misinformation or not that will be obtained on the base of (1) approaches of Supervised Machine Learning and (2) an aligned parallel text corpus that will be created through the scraping Ukrainian, Russian, EU and USA news websites and will be semantically annotated in accordance with event types and event arguments. Expected results:

- A short demo summarizing the microproject and its results
- Annotated corpus of news articles that is aligned in event date and event place.
- A paper published in a relevant journal or conference

Mini project 2: Voting advice applications (VAAs) are increasingly popular throughout the world. Especially in European countries such as Finland or the Netherlands, these applications reach large parts of the electorate. Even VAAs with a small effect size are therefore of fundamental democratic interest. VAAs map the political views and wishes of the electorate onto that of parties or candidates in elections and so reduce the cost to the voters of obtaining and analysing political information. This is of particular benefit to voters of countries with multi-party systems and complex political landscapes. Due to this facilitation of decision making, VAAs have been shown to increase election turnover vis-a-vis abstention.

However, VAAs are not without their challenges. The best matching result depends to a large degree on design voices, such as which model or metrics are being used. Similarly, as voters, we do not answer consistently. Our preferences are impacted by voting method, wording and circumstance. Therefore, application design is of vital importance for fair and independent opinion formation. Aims: This mini project seeks to put in relation the biases of VAAs with the stakeholders funding their development and operation, and to identify methods of mitigating these biases. Expected results:

- A paper published in a relevant journal or conference
- A short video summarizing the microproject and its results