Toward AI Systems that Augment and Empower Humans by Understanding Us, our Society and the World Around Us

Grant Agreement Number: 761758

Project Acronym: HumanE AI

Project Dates: 2019-01-01 to 2019-12-31

Project Duration: 12 months

D2.1 HumanE AI Concept and Research Plan

Author(s): James Crowley

Contributing partners: Oulasvirta Antti, John Shawe-Taylor, Mohamed Chetouani, Barry O'Sullivan, Ana Paiva, Andrzej Nowak, Catholijn Jonker, Dino Pedreschi, Fosca Giannotti, Frank van Harmelen, Jan Hajic, Jeroen van den Hoven, Raja Chatila, and Yvonne Rogers.

Date: November 20 2019

Approved by: Paul Lukowicz

Type: Report

Status: First report

Contact: James.Crowley@inria.fr

Dissemination Level		
PU	Public	

DISCLAIMER

DOCUMENT INFO

0.1 Authors

Name	Institution	e-mail
James Crowley	INRIA	James.Crowley@inria.fr
Paul Lukowicz	DFKI	paul.lukowicz@googlemail.com

With contributions from

Name	Institution	e-mail
Oulasvirta Antti	Aalto, FI	<antti.oulasvirta@aalto.fi>
John Shawe-Taylor	UCL, UK	<j.shawe-taylor@ucl.ac.uk>
Mohamed Chetouani	Sorbonne U. FR	<mohamed.chetouani@sorbonne-universite.fr>
Barry O'Sullivan	UC Cork, UK	<barry.osullivan@insight-centre.org>
Ana Paiva	IST, PT	<paiva.a@gmail.com>
Andrzej Nowak	U Warszawski, PL	<andrzejn232@gmail.com>
Catholijn Jonker	TU Delft, NL	<C.M.Jonker@tudelft.nl>
Dino Pedreschi	CNR, IT	<pedre@di.unipi.it>
Fosca Giannotti	CNR, IT	<fosca.giannotti@isti.cnr.it>
Frank van Harmelen	Vrije Uni Bruxelles, BE	<Frank.van.Harmelen@vu.nl>
Jan Hajic	Univ. Karlova, CZ	<hajic@ufal.mff.cuni.cz>
Jeroen van den Hoven	TU Delft, NL	<M.J.vandenHoven@tudelft.nl>
Raja Chatila	Sorbonne U. FR	<Raja.Chatila@sorbonne-universite.fr>
Yvonne Rogers	UC London,UK	<y.rogers@ucl.ac.uk>

Document History

Revision		
Date	**Lead Author(s)**	**Comments**
23-24 September 2019	James Crowley	Draft of Research Challenges developed during Brussels Meeting.
10-11 Octobre 2019	Paul Lukowicz	Workshop on Research Methods at Den Haag. Important discussions about Challenges and possible micro-projects.
26 October 2019	James Crowley	Revised outline of report, proposed challenges
28 October 2019	James Crowley	Revised structure of report.
30 October 2019	James Crowley	Edited first version of Introduction.
31 October 2019	James Crowley	Revised Introduction
2 November 2019	James Crowley	Worked on Research Challenges using input from Humane AI writing team.
10-12 November 2019	James Crowley	Rewrote section on Humane AI Vision using input from Humane AI writing team.
13-14 November 2019	James Crowley	Edited section on research challenges using input from Humane AI writing team.
15 November 2019	James Crowley	Added section on Research Methods
16 November 2019	James Crowley	Expanded Introduction to describe contents
18 November 2019	James Crowley	Wrote executive summary, edited introduction.
19 November 2019	James Crowley	Added Bibliography.
20 November 2019	James Crowley	Edited section 4 to avoid repetition with another deliverable.

Table of Contents

EXECUTIVE SUMMARY

This report describes results from an initiative to organize a community of researchers and innovators around a research program that seeks to create AI technologies that empower humans and human society to vastly improve quality of life for all. The introduction describes the context and motivation for an initiative to organize the Humane AI community. Section 2 describes the Humane AI Vision to create the foundations for AI systems that empower people and society. This section describes enabling technologies and lists required innovations in a broad spectrum of areas including Machine Learning, Computer Vision, Robotics, Human Computer Interaction, Natural Language Processing and Conversational AI.

Section 3 details the research challenges that must overcome to meet this vision. These include new approaches to learning, reasoning, and planning, new theories and techniques for multimodal perception and modeling to enable intelligent systems to perceive and model humans, and human behaviours, human language, and social interaction, new theories and methods for combined human-machine intelligence, where AI and humans interact and collaborate. Section 3 also describes the challenge of assuring that the symbiosis of humans and AI systems, and the need for new technologies that enable AI systems that are ethical, legal and responsible by design, in accordance with European ethical, cultural and social values.

Section 4 describes methods by which Humane AI can best pursue its vision in the current socio-economic environment. We put forward a program for challenge-based research in which the community documents research challenges and with published performance targets using benchmark problems and data sets. We describe an approach centered on dynamically defined collaborative mini-projects undertaken by small groups of researchers to opportunistically respond to research problems and opportunities as they emerge, without having to wait for the multi year funding cycles.

1. INTRODUCTION

Over the course of the last decade, AI researchers have made ground-breaking progress in hard and long-standing problems related to machine learning, computer vision, speech recognition and autonomous systems. In combination with the increasing availability of planetary scale data from the internet, and extremely powerful computing resources using cloud computing, AI is quickly becoming an integral part of human society, economic activity as well as an indispensable tool of scientific discovery. As impressive as these developments are, and as much as these technologies have already changed our lives, there is a general agreement that what we see today is just the beginning of an AI revolution. There is also a strong consensus that AI will bring changes that will be much more profound than most other technological revolutions in human history. Depending on the course of this technological revolution will take, AI can either empower individuals and society, creating unimaginable opportunities to improve overall human experience and quality of life, or create the tools to destroy society, enslave individuals, and concentrate power and wealth in the hands of a few.

Europe carries the responsibility for reorienting the AI revolution. The choices we face are related to fundamental ethical and cultural values concerning the impact of AI on society, in particular where it refers to the future of labor, evolving social interactions, healthcare, privacy, fairness and security. The ability to make the right choices requires new solutions to fundamental scientific questions in Artificial Intelligence and Human Computer Interaction.

Humane AI represents a community of researchers and innovators who seek to create the conditions for AI technologies that empower humans and human society to vastly improve quality of life for all. This community has set as its goal to develop the scientific and technological foundations needed to shape the AI revolution in a direction that is beneficial to humans and humanity on both an individual and a societal level. The aim is to facilitate AI systems that enhance human capabilities and empower people as individuals and while assuring evolution of a healthy and nurturing society. The HumanE AI community must bring about the mobilization of a research landscape far beyond the direct project funding and create a unique innovation ecosystems that will provide many fold return on investment for the European economy and society.

The HumanE AI community was originally formed to create Flagship initiative to meet the challenge of Human Centred AI. This current report reports preliminary results from a one-year community-wide initiative to organize the community to meet this vision. The contents of this report are adapted from the results of a series of workshops and meetings that have been held over the last year, including workshops organised in Berlin, Brussels and Den Haag. Important inspiration has also been provided from the work of WP 7 of AI4EU on Human Centric AI, as well as a communal writing effort carried out during the period form mid September to mid November 2019.

Section 2 describes the Humane AI Vision to create the foundations for AI systems that empower people and society. This section describes enabling technologies for systems that assist and empower people and society. Section 2 lists required innovations in a broad spectrum of areas including Machine Learning, Computer Vision, Robotics, Human Computer Interaction, Natural Language Processing and Conversational AI.

Section 3 details the research challenges that must overcome to meet this vision. These include new approaches to learning, reasoning, and planning are interactive processes involving close synergistic collaboration between AI systems and people within a dynamic, possibly open-ended real-world environment. This will require new theories and techniques for multimodal

perception and modeling to enable intelligent systems to perceive and model humans, human actions, and behaviours, human attention and awareness, human emotions, human language and human social interaction, as well as real-world human environments. This vision will also require new theories and methods for combined human-machine intelligence, where AI and humans interact and collaborate. We must explore how best to assure that the symbiosis of humans and AI systems can best work together as a society. The Humane AI vision also requires new technologies that enable AI systems that are ethical, legal and responsible by design, in accordance with European ethical, cultural and social values.

Section 4 describes methods by which Humane AI can best pursue its vision in the current socio-economic environment. We put forward a program for challenge-based research in which the community documents research challenges with published performance targets using benchmark problems and data sets. We describe an approach centered on dynamically defined collaborative mini-projects undertaken by small groups of researchers to opportunistically respond to research problems and opportunities as they emerge, without having to wait for the multi year funding cycles. We describe how the Humane AI community can work with the AI4EU one-stop-shop innovation platform, the European Digital innovation hubs and other relevant European and national initiatives.

2. THE HUMANE AI VISION

Turing defined intelligence as human-level performance at interaction. However, original Turing test envisaged interaction via questions typed into a text-only teletype. Modern technology allows us to extend Turing's challenge to systems that exhibit human-level performance at interaction with people, with the physical world and with other artificial systems through multiple modes of perception and action. The goal of Humane AI is to harness the emergence of enabling technologies for human-level interaction to empower individuals and society, by providing new abilities to perceive and understand complex phenomena, to individually and collectively solve problems, and to empower individuals with new abilities for creativity and experience.

The Humane AI community seeks to create the foundations for AI systems that empower people and society. Our goal is to provide enabling technologies for systems that seamlessly fit in with complex social settings and dynamically adapt to changes in our environment to empower people. This will require innovations in a broad spectrum of areas including Machine Learning, Computer Vision, Robotics, Human Computer Interaction, Natural Language Processing and Conversational AI. In particular, HumanE AI, enable the emergence of a new paradigm for Collaborative Humane Computer Interaction based on a convergence of HCI with Machine Learning. Core innovations should include (1) tools for enhancing human cognitive capabilities, channeling human creativity, inventiveness and intuition and empowering humans to make important decisions in a more informed way, (2) AI systems that can intelligently interact within complex social settings and operative in open-ended environments, (3) enabling technologies for explainable, transparent, validated and trustworthy AI systems and (4) new approaches to embed value based ethics and European Cultural, Legal and social values as core design considerations in all AI systems and applications.

HumanE AI must develop the scientific and technological foundations for artificial intelligence that is beneficial to humans and humanity, in accordance with European ethical, social, and cultural values. The core challenge is the development of robust, trustworthy AI systems that understand humans, adapt to human environments, and behave appropriately in social situations. Our overall goal is to develop a technology for artificial intelligence that enhances human abilities and empower individuals and society.

Achieving this vision requires new solutions to fundamental scientific questions—not just within narrow classical AI silos, but at the intersections of AI disciplines such as learning, reasoning, and perception with scientific domains including human-computer interaction, cognitive science, and the social sciences. This will require substantial research in areas such as:

1. **Human-in-the-loop machine learning, reasoning, and planning**. 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.
2. **Multimodal perception and modelling**. 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.

3. **Human-AI interaction and collaboration**. Developing paradigms that allow humans and AI systems including service robots and smart environments to interact and collaborate in a way that enhances human abilities and empowers people.
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.
5. **Legal and ethical bases for responsible AI**. Ensuring that the design and use of AI is aligned with ethical principles and human values, taking into account cultural and societal context, while enabling human users to act ethically and respecting their autonomy and self-determination. This also implies that AI systems must be "under the Rule of Law": their research design, operations and output should be contestable by those affected by their decisions, and a liability for those who put them on the market.

3. RESEARCH CHALLENGES FOR HUMANE AI.

At the core of our concept is to develop the foundations for intelligent systems that interact and collaborate with people to enhance human abilities and empower both individuals and society. Collaboration will require that humans and AI systems work together as partners to achieve a common goal, sharing a mutual understanding of each other's abilities and respective roles. Human-level performance in collaboration will require integration of learning, reasoning, perception, and interaction.

Humane AI must go beyond HCI challenges, to ensure the human maintains control. This includes enabling users to understand how interactions are driven (transparency) and to maintain final control over interaction with AI systems. This also includes addressing compliance with European ethical and social values as a core research problem, and not simply a boundary condition. We must seek new methods to construct compliance for European ethical, legal and cultural values are by design. This will require multidisciplinary collaboration of AI, philosophy, social science, and complex systems.

A number of fundamental gaps in knowledge and technology must be addressed in three closely related areas. The first area is learning, reasoning, and planning methods, which allow for a large degree of interactivity. To facilitate a collaboration between humans and AI systems based on trust and enhancing each other's capabilities, intelligent systems must not only be able to provide explanations at the end of the learning or reasoning task. They also provide feedback on progress and are able to incorporate high-level human input. We refer to such novel methods as "human-in-the-loop learning, reasoning and planning". Second, human aware interaction and collaboration will require multimodal perception of dynamic real-world environments and social settings, including the ability to build and maintain comprehensive models of environments and of humans interacting within such environments. Intelligent systems must share an understanding of a problem's larger context to properly cooperate in developing a solution. In addition, appropriate interaction and collaboration mechanisms must be developed on both the individual and collective level.

3.1 Human-in-the-Loop Machine Learning, Reasoning, and Planning

Human in the loop learning, reasoning, and planning are interactive processes involving close synergistic collaboration between AI system(s) and user(s) within a dynamic, possibly open-ended real-world environment. Key gaps in knowledge and technology that must be addressed toward this vision include the following:

1. Hybrid representations that combine symbolic, compositional approaches with statistical and latent representations [Garcez 2019][van Harmelen 2019]. Such hybrid representations are needed to allow the benefits of data-driven learning to be combined with knowledge representations that are more compatible with the way humans view and reason about the world around them. A wide variety of representations should be investigated by the consortium, including hybrids of logic and neural networks, such as logic tensor networks [Donadello2017], and latent representations of knowledge graphs through embeddings [Wang2017] and narratives [Meghini2019].
2. Methods for leveraging the above representations to not just present humans with explanations based on simple links between the input and output spaces (e.g. [Koh2017]), but to be able to reason about shared internal representations just like humans can intuitively explain to others how they arrive at certain conclusions. This is closely connected to work on human-AI collaboration that will study how to present such reasoning to humans in various situations.

3. Methods for interactively including human understanding in the learning and reasoning process, which is difficult with current data driven approaches. This additional knowledge can take the form of conceptual categories, knowledge of causality, and common-sense knowledge. Infusing such human knowledge into the machine-learning process can increase efficiency, and improve the generalizability and robustness of the results [Marcus2019].

Under this vision, the knowledge of an AI system evolves and is influenced by its behaviour in the world and interaction with humans. We want to facilitate systems that can learn, reason, plan, act, and observe the world, by continuously and cyclically interleaving these activities. This will require addressing a number of problems.

Linking symbolic and sub-symbolic learning

The construction of hybrid systems that combine symbolic and statistical methods of reasoning is widely seen as one of the grand challenges facing AI today. For example, Pearl and colleagues noted, "Our general conclusion is that human-level AI cannot emerge solely from model-blind learning machines; it requires the symbiotic collaboration of data and models" [Pearl2018]. Marcus and colleagues stated, "By pushing beyond perceptual classification and into a broader integration of inference and knowledge, artificial intelligence will advance greatly." [Marcus 2018], [Marcus 2019]. Going further, Darwich noted, "the question is not whether it is functions or models but how to profoundly integrate and fuse function-optimisation with model-based reasoning" [Darwiche2019]. However, as shown in two of our survey papers [Garcez 2019][vanHarmelen 2019], there is no consensus on how to achieve this, with proposed techniques in the literature ranging from graph theory to linear algebra, and from propositional logic and fuzzy logic to continuous differentiable functions.

An interesting approach 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 [Schank1975]. However, to avoid the limitations of earlier work, such scripts must need to be automatically *generated* (see [Jorge2019] for an overview of the state of the art), and we must develop techniques for *using* such scripts for shared human-machine understanding [Bosser2018] and explaining [Jentner2018][Calegari2019].

Learning with and about narratives

People share knowledge by narrating stories. A narrative places the events of a story in a larger context, providing an interpretation of the story that recounts a series of events and their origin and consequences. A narration interprets the story in a manner that can be generalized and used to predict and explain events.

People use narratives to understand phenomena. Narratives make it possible to provide rich descriptions for events that are not directly observable, including prior events, and hypothetical or abstract events. Narratives enable predictions for possible future events, and to reason about how to create or avoid events.

We should investigate the use of narratives to provide human-understandable descriptions for complex situations, and sub-symbolic representations [Urbaniak2018][Gilpin2018]. We should explore how narratives can be adapted as a bridge between human reasoning and understanding, on the one hand, and internal AI representation on the other [Vlek2016]. Specifically, we must address the following questions:

1. How can AI systems process and learn from human knowledge expressed in the form of narratives and stories [Jorge2019]?

2. How can AI systems explain their reasoning, learning, and acquired knowledge in the form of narratives and stories that humans can easily understand and relate to [Pasquali2019][Gervás2019]?
3. How can AI systems and humans jointly create, adapt, and interpret narratives or stories as a means of interactively reasoning and learning together [Bosser2018]?

Continuous and incremental learning in joint human-AI systems

A requirement for AI systems with humans in the loop is the use of hybrid representations in **joint human-machine learning and planning**. An early example of this in reinforcement learning is [Garnelo2016]. Rather than the typical opaque representations usually learned in deep reinforcement learning systems, the goal 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, rather than the latent representation. More recent examples by consortium members are in [Toro Icarte 2018] and [Lever2016].

A second challenge is the use of hybrid representations in **generating explanations based on shared models between humans and machines.** This implies upgrading the knowledge-discovery process with the capability of generating high-quality machine-learning models equipped with their own human-comprehensible description, which in turn requires a novel blend of mathematical and statistical models with logic and causal inference and reasoning. Work by consortium members such as [Tiddi 2015] shows how background knowledge in the form of very large knowledge graphs can be used to generate intelligible explanations that are not constructible from data alone. [Guidotti 2019] exploits auditing methods of machine-learning models to generate explanation rules reconstructing both factual and counterfactual knowledge. Other work exploits symbolic representations such as Inductive Logic Programs to explain the neural network-generated labels of objects in images [Yang2019].

The above can provide the basis for learning with humans in the loop, e.g., by exploiting rich human feedback ("this is wrong *because*…"), exploiting implicit feedback (by obtaining feedback from behavior, voice, and face), through imitation, and via active learning (the machine asking the human "Should we explore **this**?").

Compositionality and automated machine learning (Auto-ML)

Major breakthroughs in recent AI developments have come when well-understood learning components have been composed to create more complex behaviors and systems as, for example, in AlphaGo, where a deep learner analyzing the board value is combined with a reinforcement learner implemented by using a deep learner to estimate the value function and a probabilistic method of prioritizing the exploration of the search space. These compositions are typically ad hoc and heuristic, requiring trial and error to deliver stable solutions. HumanE AI must develop a theoretical foundation for the composition of learning components, be they symbolic or sub-symbolic, enabling the reliable engineering of systems that can deliver specified complex cognitive behaviours. We should enable the combination of symbolic and statistical AI methods and further extend them with theoretical models that allow continuous adaptation.

The compositional approach to delivering AI systems has a number of advantages apart from the obvious inspiration of general software engineering. In addition to reducing more complex problems to well-understood components, it renders systems more transparent in that the decisions of one component can be traced to outputs of others via well-understood functionality. A key approach to unlocking the potential of the compositional approach is its link to optimization. Overall objectives for the cognitive system can be translated into optimization criteria that respect various constraints: there is then a natural correspondence between distributed strategies for solving the optimization problem and decompositions of the cognitive

system. This also underpins our proposed approach for rendering AI systems interpretable by learning to decompose them into simpler components, an approach that can identify structure in the solution, hence rendering it more robust and explainable.

The recent success of general-purpose algorithm configuration and selection methods—and notably, the rise of automated machine learning already leverages this insight. HumanE AI should build on this foundation, devising methods for automating the development, deployment, and maintenance of AI systems that are efficient, robust, and predictable, without requiring deep and highly specialized AI expertise. The key to achieving this vision of automated AI is to render intelligent 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.

Quantifying model uncertainty

Likelihood estimation forms an important part of both our understanding of the world and the way in which we communicate that understanding to others. To interact meaningfully with humans intelligent systems must use the vocabulary and semantics of probabilistic arguments in a way that is accessible and understandable to humans. Uncertainty quantification is also a vital component of communication for evaluating alternative interpretations of a situation, for assimilating information from different sources, and for making decisions about what new information would be most useful in disambiguating a concept or question.

A project on HumanE AI must investigate methods for both assessing and quantifying uncertainty of individual models, but also of the ways in which this can be inferred when models and/or information are combined, hence propagating measures of uncertainty through composite systems. Uncertainty will be important at all of the aforementioned levels, from assessing the confidence of individual estimations to the likelihood of logical relations or narratives in a particular context.

There are a variety of methods for estimating blackbox uncertainty, such as Bayesian posterior distributions estimated for example in dropout models for deep learning networks, or more precise measures such as conformity that can guarantee accurate percentile bars that hold with high confidence [Barber et al., 2019] . Extending such approaches to composite and dynamical systems should be an important focus to inform decisions made by an agent, either to increase its information or alternatively trade information gain with expected success, as in bandit-style algorithms. Such uncertainty estimates may also link with hard or soft constraints that must be placed on a system in order for its behaviour to be “safe” or “desirable.” The best way to monitor and model the uncertainty should be investigated with approximate reasoning techniques as well as separate modeling “watching” networks. The approaches may be important in driving lifelong learning algorithms in which uncertainties that can determine which models need refinement and/or verification from new data, and might also be sought through interaction with humans by asking for clarifications.

3.2 Multimodal Perception and Modelling

To interact and collaborate with people, intelligent systems must be able to perceive and model humans, human actions, and behaviours, human attention and awareness, human emotions, human language and human social interaction, as well as real-world human environments.

Human interaction and human collaboration depend on the ability to understand the situation and reliably assign meanings to events and actions. Actions can have quite different meanings, depending on context. People infer such meanings either directly from subtle cues in behaviour, emotions, and nonverbal communications or indirectly from the context and background

knowledge. This requires the ability to sense subtle behaviour, and emotional and social cues, as well as an ability to automatically acquire and apply background knowledge to provide context. Acquisition must be automatic because such background knowledge is far too complex to be hand-coded. HumanE AI must provide this foundation by building on recent advances in multimodal perception and modelling sensory, spatiotemporal, and conceptual phenomena.

Multimodal interactive learning of models

Perception is the association of external stimuli to an internal model. Perception and modeling are inseparable. Human ability to correctly perceive and interpret complex situations, even when given limited and/or noisy input, is inherently linked to a deep, differentiated, understanding based on human experience. Current limitations of computer perception are rooted in an inability to acquire and use such background knowledge.

We must 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. Such models are required to organize and provide context for perception of objects, events, and actions. 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.

Multimodal perception and narrative description of actions, activities and tasks

As explained above, 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. Situating observations within a context will be required to predict the intended and actual consequences of the action, and explanations for the purpose of the action.

For example, monitoring of manipulation activity requires recognition of manipulation actions in the context of an activity. The activity context provides constraints that can be used to focus attention on the objects and materials to be manipulated, and to disambiguate recognition results that are uncertain or ambiguous. This disambiguation applies to recognition of actions as wells objects and materials. The use of activity context can reduce both the error rate and the computational cost for action recognition.

A manipulation activity can be formalized as a process, modeled as a series of state transitions. With this approach, the activity is monitored as a series of states, where the process state is the composition of the states of the individual objects. This process state is referred to as a situation [Johnson-Laird 89]. The situation model provides context for the action, making it possible to describe the action as part of story [Genette 72] with context information about why and how the action was performed. This story may then be interpreted as part of a narrative, associating contextual information that makes it possible to explain the action and predict its consequences. The results may be used to drive a natural language generation (NLG) tool to communicate and interact with a human collaborator.

Multimodal perception of awareness, emotions, and attitudes

Human awareness is constrained by limits to working memory and perceptual abilities. Modelling awareness is required to permit a system to predict human abilities and construct explanations. Awareness can be perceived from fixation, head orientation, posture, and vocal

interjections, as well spoken language interaction. Emotions play a fundamental role in human reason, and can be perceived from physiological signs such as micro-expression, heart rate, posture, self-touch, prosody, and paralinguistic expressions. Attitude condition (how humans react to phenomena) can be determined from patterns of reactions, as well as direct spoken language interaction.

Much of human activity is reactive and unconscious. At the most basic level, sensory signals directly drive human muscles and emotional responses at the signal level in a tightly coupled interaction. Multiple sensor modalities, including tactile, visual, and auditory may be combined in such signal-level interaction. Systems that interact and collaborate with humans must appropriately respond to such signals.

So far, little work in AI has sought to provide understanding human intentions and attitudes. Some previous work has focused on agent-based modeling [Georgeff98], with few results in real-time recognition in interactive, real-world scenarios. Comprehensive world models, combined with the ability to involve humans in the learning and reasoning process, may provide an approach to provide intelligent systems with such abilities.

Perception of social signals and social interaction

Most research on perception of human interaction tends to focus on recognizing and communicating linguistic signals. However, much of human-human interaction is nonverbal and highly dependent on the social context. A technology for situated interaction will require abilities to perceive and assimilate nonverbal social signals, to understand and predict social situations, and to acquire and develop social interaction skills. Brezeal and colleagues [Breazeal2016] have recently surveyed research trends in social robotics and its application to human-robot interaction (HRI). They argue that sociable robots must be able to communicate naturally with people using both verbal and nonverbal signals, and engage users on both cognitive and emotional levels to provide effective social and task-related services.

To achieve our vision for HumanE AI, we must develop methods to endow intelligent systems 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 interaction, and ultimately for other forms of cognitive abilities. We propose to capture social common sense by training the appropriateness of behaviours in social situations. A key challenge is to employ an adequate representation for social situations.

It may be possible to encode knowledge for sociable interaction 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 must explore methods that start from simple stereotypical situation models and extend a situation graph by adding new situations.

Distributed collaborative perception and modeling

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. Sharing of information provides a background context that is accepted within a culture and provides a powerful foundation for reasoning and communication through common sense. Human narratives convey information concerning what sequences of behaviors are required in specific social situations. Narratives are the source of prediction for how other actors will behave in a specific situation, as well as determining the agent's appropriate reaction to these behaviors, along with the consequence of these actions. Narratives also are used to explain the behavior of 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 modeling, 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.

Methods for overcoming the difficulty of collecting labeled training data

Getting sufficiently labeled training data is a core concern for many ML domains. For example, much of the recent progress in computer vision and language processing has been related to the availability of huge public datasets (e.g., the 1-million-picture ImageNet dataset [Deng2009]), which enabled public ML challenges (e.g., the Large-Scale Visual Recognition Challenge with ImageNet [Russakovsky2015]). 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. Some reasons for these challenges include the following:

- Annotation, often executed by visualizing video recording of experiments, can be extremely slow. Some researchers report that annotating 10 minutes of video may take as much as 10 hours of time, when very fine-grained and accurate annotations are required [Roggen2010]. Other researchers surveyed publicly available datasets and reported the costs of creating them, with many public datasets costing in the hundreds of thousands of dollars [Welbourne2014].
- Annotations often carried from video recordings cannot be used when natural behavior must be captured, as such recordings are likely to influence how naturally people behave, and in a number of situations it may not be possible to collect such datasets for ethical reasons, which limits collection to in-lab "naturalistic" emulation of everyday scenarios.
- Annotating datasets require someone previously enumerating a set of elements to identify (e.g., interesting situations or actions) within a recording. In long-duration datasets—which are especially used for unsupervised and open-ended perception and modeling—an a priori exhaustive enumeration of such interesting situations is infeasible, and post hoc re-annotation from videos may not be possible, depending on the nature of the experiment, informed consent, and ethical considerations.
- Finally, interactive systems, where an activity-aware system interacts with humans (e.g., in a human-robot interaction task) will also influence people's behavior dynamically. Therefore, such experiments cannot be captured easily in a static dataset.

These challenges open a wide number of fundamental research areas to enable the collection of larger scale and more realistic datasets, which will be explored in this project. Some of the approaches that will be pursued include:

- Leveraging the large availability of online datasets, and devising methods to transform such datasets to make them suitable for the sensor modalities available. For instance, recent work has shown that sensor data can be seamlessly transformed across modalities (e.g., between RGB images and wearable sensors [Fortes2019] or depth sensors and wearable sensors [Banos2012]). These can be combined with the availability of online datasets that are potentially of different modalities (e.g., YouTube data that comprises textual annotations to be converted to data suitable for wearable sensors).
- Leveraging crowdsourcing to annotate datasets, with AI approaches to support the efficient annotation (e.g., identifying relevant time segments), improve robustness (inter-rater reliability), and design the technical infrastructure to integrate these approaches (e.g., [Satybaldiev2019]).

- Exploiting human-in-the-loop learning, such as letting users provide annotations at their own pace through a combination of active learning (e.g., prompting users about their activities) and semisupervised learning. A primary challenge is to identify when it is most valuable to prompt a user, as a combination of information gain and minimal distraction.

3.3 Human AI Interaction and Collaboration

Beyond considering the human in the loop, the goal of human-AI is to study and develop methods for combined human-machine intelligence, where AI and humans work in cooperation and collaboration. To achieve this, we must investigate principled approaches to support the synergy of human and artificial intelligence, enabling humans to continue doing what they are good at but also be in control when making decisions.

Researchers at Stanford University have proposed that AI research and development should follow three objectives: (i) to technically reflect the depth characterized by human intelligence; (ii) improve human capabilities rather than replace them; and (iii) focus on AI's impact on humans [Li2018]. There has also been a call for the HCI community to play an increasing role in realizing this vision, by providing their expertise in the following: human-machine integration/teaming, UI modeling and HCI design, transference of psychological theories, enhancement of existing methods, and development of HCI design standards.

Foundations of human-AI interaction and collaboration

HumanE AI must address: (i) interaction, (ii) collaboration, and (iii) symbiosis. For interaction with intelligent systems, we can consider several different approaches to provide more natural channels for communication and tightly coupled perception-action, including multimodality, conversation, and augmented reality. At the level of collaboration, we must consider concepts like cooperation, awareness of attention and emotion and collective intelligence. For symbiosis, we should 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.

The knowledge that HCI can bring to bear on these three forms of human-AI include user modeling, inference, and machine-learning methods suitable for interactive settings with humans, deep empirical research and the design of interaction techniques, and user interfaces for interaction with artificially intelligent partners. Empirical methods in HCI can provide a way of discovering the mental representations people develop when using AI systems, their expectations when interacting with an AI system, and their acceptance of the decisions it suggests or actions it makes itself. Understanding these aspects better is crucial for humans to be able to collaborate with AI, but also for AI methods such as inverse or machine theory of mind.

Additional questions that must be answered include: How is user acceptance and the adoption of AI systems affected by the cultural and social background of the user? What is the overall effect (short and long term) of interacting with intelligent systems on humans and the environment?

A core capability in human-AI interaction is *understanding* of human partners. While present-day ML research mostly approaches this as a classification or prediction task in supervised or unsupervised learning, we seek a new foundation from theories of human behavior. In particular, we believe that models and theories from computational psychology [Sun2008], computational cognitive sciences [Kriegeskorte2018], and computational social sciences [Lazer2009] can underpin artificial understanding of human behavior. This research calls for plausible models of human behavior that we will use for artificial agents that can—thanks to causal models that link behavior with cognitive, emotional, and other latent factors—better infer, plan, and act without extensive data on an individual [Lake 2017]; this, of course, also presupposes high-quality

language capabilities in both speech and text domains (T3.6), to analyze the speech and/or text to a formalized representation suitable to provide "input" to the aforementioned theories and models that will then be able to arrive at the true understanding of human behavior.

We will need to assemble the complementary expertise of our consortium members to open new avenues to explore the three types of human-AI and address associated research with these questions. One such example is a study of social practices on greeting rituals.

Human-AI interaction and collaboration

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. This is sometimes referred to as a Theory of Mind. As such, there are several research questions:

- When should the intelligent system's processes be externalized so that system functionality provides the appropriate level of transparency for the user?
- How should relevant information about internal processes of the AI system be represented, to make it intuitively understandable to the user?
- How can humans communicate comprehension and provide recommendations as a form of dialogue with respect to AI reasoning?
- What types of interaction are best suited for different situations?

Our approach to answering these questions combines theoretical analysis with empirical user-centered design. First, in terms of *theoretical analysis*, we will analyze interactive problems as games and decision problems. For example, we will use Markov Decision Processes, which can be solved in simulation or in some cases analytically. These will be simplified such that we can infer conditions under which information disclosure between the two partners does or (or does not) work. Second, in terms of *empirical user-centered design*, these ideas will be developed in a user-centered manner with pre-studies of the particular applications, and evaluated empirically with representative user groups.

Reflexivity and adaptation in human-AI collaboration

Key questions for systems where humans and AI work with each other synergistically to support each other as partners in co-creation are:

- How do humans and machines continuously adapt to each other and the context?
- How can machines understand their impact on humans before taking action?
- How do we design self-aware systems that can monitor and self-diagnose their interactions with the environment and other humans to self-improve their interactions?

Our approach is to build on meta-reasoning methods, wherein the behavior of the artificial agent is supervised at a higher level, which the consortium has explored earlier, for example, in ubiquitous computing.

In particular, 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.

User models and interaction history

User models can be divided according to how humans mind is represented in interaction (e.g., neural, mathematical, simulation, RL-based, and Bayesian) [Kriegeskorte2018], and which factors are included (e.g., cognitive, physiological, emotions, and motivational). We here pursue two important capabilities that user models should have: (1) *forward modeling,* or providing a richer and more generalizable account of human behavior suitable for real-world interactive AI, which has been an issue in cognitive and user models for decades, and (2) *inverse modeling,* or fitting models to individual users. Both are needed for deployment in interactive AI, which must on the one hand update its model representations with interactions and, on the other, select actions while anticipating their consequences on users (counter-factuality) [Lake2017]. Recent user models have also used reinforcement learning, wherein the state-space quickly explodes with longer user history, or embeddings (Rabinowitz 2018) that collapse multidimensional behavior to an uninterpretable lower-dimensional representation. We need to develop model-based approaches make it possible to combine inferential and learning capabilities with explicitly specified structures.

In addition, we should explore the use of 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).

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.

Guidance has its roots in HCI and can be seen as a mixed-initiative process. As Ceneda and colleagues note, "Guidance is a computer-assisted process that aims to actively resolve a knowledge gap encountered by users during an interactive visual analytics session". A mixed-initiative process is an approach whereby both humans and systems can "take the initiative" and contribute to the process. The central elements are the time, degree, and type of involvements. Guidance is a dialogue between humans and systems in which humans provide—implicitly or explicitly—their own needs and issues as input and the system provides possible answers to alleviate problematic situations.

Natural language processing and conversational AI

NLP is an enabling technology for several, if not most, areas of the HumanE AI, and is particularly relevant to perception, interaction, and HCI areas, with fundamental methods being tackled in machine learning. Natural language is a natural form of communication between humans, in both spoken and text form.

Today, as a rule, NLP uses deep-learning techniques. The main focus in this proposal, however, is to move away from the highly successful "blackbox" approaches, to connect the high performance of the neural network paradigm with symbolic methods—especially in the area of semantics and understanding, where existing, human-understandable databases and ontologies are used to approximate world knowledge. Speech and text analysis and generation are a necessary component of communication and interaction, such as in dialogue systems of all sorts. Similarly, when a computer must generate a narrative (to explain reasoning or arguments), natural language generation may be used; and conversely, any human input must be tackled first by an natural language understanding component.

The main areas of research should be (1) analyzing natural language speech and text beyond the current state of the art; (2) NLG from planned, formally represented communication; (3) explaining "why" in deep general understanding systems; (4) multilingual issues in all of the above, and machine translation where needed for cross-lingual understanding and communication; and (5) creating and unifying an ontology where needed (e.g., on event types).

A key research theme building current techniques for natural language processing is to enhance human reflection on actions via a dialogue with the intelligent system. This can be seen as a form of conversational AI. Often intelligent systems are developed to advise or suggest without the opportunity for negotiation or understanding.

Trustworthy social and sociable interaction

Reeves and Nass argue that a social interface may be the truly universal interface [Reeves 98]. 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 modeling of social context and complex, evolving world models we will address key challenges in enabling AI systems to act appropriately within complex social contexts.

Breazeal has proposed a hierarchy of four classes of social robots, from socially evocative to sociable. As one moves progressively up the hierarchy, robots' abilities to engage in social interaction increase. Within this hierarchy, socially evocative robots are designed to encourage people to anthropomorphize technology to interact with it. Socially communicative robots use human-like social cues and communication modalities to facilitate interactions with people. Socially responsive robots are able to learn from their interaction and social partners. Sociable robots are socially participative, and maintain their own internal goals and motivations.

Kendon [Kendon 75] argues for understanding social interaction as a form of dialogue. With this view, prosody and gestures are seen as annotations to the linguistic contents of interaction, serving to guide attention as well as to communicate non-linguistic signals. Pentland [Pentland 2005] proposes an approach based on social signalling of attitude and attention, using such vocal cues as amplitude, frequency, and timing of prosodic and gestural signals. Such unconscious signals provide important cues about social situations and social relations that are not available in measures of affect. The importance of such signals is one of the reasons that we propose extending our investigation beyond visual perception into acoustic and tactile perception modes [Ta 2015].

When interacting with one or more persons intelligent 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. Social sciences have for decades studied emergent properties of social groups. However, techno-social systems' emergent properties are much less understood.

3.4 Societal AI

As increasingly complex socio-technical systems emerge, consisting of many interacting people and intelligent and autonomous systems, AI acquires an important societal dimension. A key observation is that a crowd of (interacting) intelligent individuals is not necessarily an intelligent crowd. On the contrary, it can be idiotic in many cases, because of undesired, unintended network effects and emergent aggregated behavior. Examples abound in contemporary society. Anyone who has used a car navigation system to bypass a traffic jam knows. Each navigation system generates recommendations that make sense from an individual point of view, and the driver can easily understand the rationale behind the recommendations. However, the sum of decisions made by many navigation systems can have grave consequences on the traffic system as a whole: from the traffic jams on local alternative routes to ripples propagating through the system on a larger spatial scale, to long-term behavior changes that may lead to drivers permanently avoid certain areas (which can have a negative economic impact on disadvantaged neighborhoods), or artificially increase the risk of accidents on highly recommended roads.

The interaction among individual choices may unfold dramatically into global challenges linked to economic inequality, environmental sustainability, and democracy. In the field of opinion formation and diffusion, a crowd of citizens using social media as a source of information is subject to the algorithmic bias of the platform's recommendation mechanisms suggesting personalized content. This bias will create echo chambers and filter bubbles, sometimes induced in an artificial way, in the sense that without the personalization bias the crowd would reach a common shared opinion. Again, a recommender system that makes sense at an individual level may result in an undesired collective effect of information disorder and radicalization.

Aggregated network and societal effects and of AI and their (positive or negative) impacts on society are not sufficiently discussed in the public and not sufficiently addressed by AI research, despite the striking importance to understand and predict the aggregated outcomes of sociotechnical AI-based systems and related complex social processes, as well as how to avoid their harmful effects. Such effects are a source of a **whole new set of explainability, accountability, and trustworthiness issues,** even assuming that we can solve those problems for an individual machine-learning-based AI system. Therefore, we cannot concentrate solely on making individual citizens or institutions more aware and capable of making informed decisions. We also need to study the emerging network effects of crowds of intelligent interacting agents, as well as the design of mechanisms for distributed collaboration that push toward the realization of the agreed set of values and objectives at collective level: sustainable mobility in cities, diversity and pluralism in the public debate, and fair distribution of economic resources.

We therefore advocate the emergence of *societal AI* as a new field of investigation of potentially huge impact, requiring the next step ahead in transdisciplinary integration of AI, data science, social sciences, psychology, network science, and complex systems.

Graybox models of society scale, networked hybrid human-AI systems

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 that may range from synchronization to collapse.

Naturally, several questions regarding the considerable challenges emerge: How can systems be modeled adequately and predict these networked effects? What are the typical scenarios of system evolution? What are the relevant mechanisms and quantities to control to prevent a system from unpredicted/harmful behavior? How can researchers design collaborative, distributed learning and data-mining methods for AI systems that are motivated by the social mechanism for accumulating "common knowledge" and "collective wisdom" without unnecessary, unsustainable, and harmful centralized collection of raw personal data? What are the best ways to design and manage such a complex system, so that it behaves in a way that is compliant with ethical principles, while dealing with the Collingridge dilemma, in which designers must select solutions at the beginning from a broad variety of possibilities, with little information about the perception of the suggested solutions, while proceeding in the process and accumulating feedback in which the degree of available freedom for the design shrinks.

AI systems' individual versus collective goals

Social dilemmas occur when there is a conflict between individual and public interest. 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 keeping 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?

As already illustrated, these complex systems should work on fulfilling collective goals (or requirements). However, requirements change over time, as they also change from one context to another. How can we design and manage such complex socio-technical systems that **adapt to our evolving requirements**? How can we maintain humans' control in such systems to ensure that it is the humans' requirements and values that are being considered?

A related question is how to design mechanisms that support distributed socio-technical systems made of self-interested agents. Such systems should be both efficient and ethical. In other words, the challenge is to develop mechanisms that will result in the system converging to an equilibrium that complies with the European values and social objectives (e.g. income distribution) but without unnecessary losses in efficiency. Interestingly, AI can play a vital role of enhancing desirable behaviors in the system, e.g., by supporting coordination and cooperation that is, more often than not, crucial to achieve any meaningful improvements. Thus far, teaching and learning in repeated strategic situations were already studied theoretically [Camerer2002] and the experiments were conducted involving human players [Hyndman2012]. Importantly, however, AI technologies bring many new possibilities to the table [Crandall2018, Peysakhovich 2018] because, unlike with human players, we now have a unique chance to design both not only the rules of interaction but also some of the participants as such. Our ultimate goal is to build a scheme of a socio-technical system in which AI not only cooperates with humans but if necessary helps them to learn how to cooperate as well as other desirable behaviors.

Societal impact of AI systems

How to **evaluate societal impact** of competing AI technologies and promote the ones more compliant with the European values? As one of the possible approaches, explore how to construct **in vitro (controlled) experiments** of the interaction between AI technologies and humans, in order to select the technological setup most suitable for the ethical standards.

Understand and model the way algorithms and AI technologies **reinforce/generate certain undesirable human behaviors** and emerging societal phenomena, like producing echo-chambers, opinion polarization, and bias amplification at the collective level. Ultimately, the goal is to 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.

Artificial intelligent technologies have the potential to substantially change the relation between the governing and governed. New opportunities open to obtain feedback and make predictions about the effect of intended measures and several new channels for participation in decision making will emerge. 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, 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?

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

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). It has been already argued for long that social influence is the most fundamental and pervasive social process [McGuire 1985]. For instance, mutual influences among individuals underlay formation of public opinion [Nowak 1990], group decisions, and actions [DeDreu2008]. At the group level, social influence is tantamount to *distributed, optimizing information processing*. In particular, in this process, individuals optimize their decision-making and judgment by delegating information processing to potential sources of influence. In this context, the Regulatory Theory of Social Influence [Nowak 2020] specifies four factors —*trust*, *coherence*, *issue importance*, and *own expertise*—that play a critical role in the processes of determining the target's choice of sources and the level of abstraction in the information sought from these sources. Beyond maximizing the cognitive efficiency of the target and the quality of his or her outcomes, these processes also enhance the functioning of the social group in which the target is embedded, because the most expert on the topic and the most reliable group members gather and process the information. Our intention is to use the research on human groups [Petty 1986; Mullen 1994; Nowak 2020, Nowak 2017] as a starting point of designing **AI members of socio-technical groups**, which can improve the functioning of the group in reaching optimal decisions and judgments. AI agents need to understand their role in distributed information processing social systems, be aware of the competence and reliability of group members, the importance of the issue at hand and their own limitations. Another challenge is to design the rules by which on the basis of this knowledge, AI agents can decide which information to process themselves and which to delegate to humans, and who in the group is most capable of processing which information. We propose to design mechanisms that enable AI agent to easily and as naturally as humans estimate trustworthiness of both human and other AI members of the group and to use trust estimates as guidance for optimal information processing in social groups.

The ultimate goal is to develop 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 behavior, 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.

3.5 AI Ethics, Law and Responsible AI

Responsible AI is about the processes by which AI is developed (ethics in AI design and development), accountability for the results of AI system deliberation (ethics by design) and making sure that those developing AI systems are aware of their role and impact on the values and capabilities of those systems (ethics for designers) [Dignum2019]. Design methods, verification techniques, and codes of conduct are all aspects that need to be developed alongside the computational design of algorithms [van den Hoven 2015].

Every AI system should operate within an ethical and social framework in understandable, verifiable and justifiable ways. Such systems must in any case operate within the bounds of the rule of law, incorporating fundamental rights protection into the AI infrastructure. Theory and methods are needed for the Responsible Design of AI Systems as well as to evaluate and measure the 'maturity' of systems in terms of compliance with legal, ethical and societal principles. This is not merely a matter of articulating legal and ethical requirements, but involves robustness, and social and interactivity design. Concerning ethical and legal design of AI systems, we will clarify the difference between legal and ethical concerns, as well as their interaction [Hildebrandt 2020], and ethical and legal scholars will work side by side to develop both legal protection by design and value-sensitive design approaches. The focus here is the prioritization of ethical, legal, and policy considerations in the development and management of AI systems to ensure responsible design, production and use of trustworthy AI. This requires integration of engineering, policy, law and ethics approaches. The following are the fundamental issues to be addressed by a research roadmap on ethics for human centered AI Systems.

- Adequate account (and criteria of adequacy) of what a moral value is, e.g. a Topos (used in moral narratives), especially in the context of the human AI interaction (with reference to some prominent values in the Ethics & AI debate, e.g. accountability, privacy, fairness.)
- Adequate account for Designing for Human Values and Ethical principles (as non-functional requirements) in the field of AI. Design for Values, value hierarchies, functional decomposition of non-functional (moral) values.
- Methods for measuring values and norms in the human-AI ecosystem as required by an agile approach to designing for values.
- Understanding how values can change (or their balance is modified) as a side effect of complex interaction between humans and AI systems in a complex socio-technical ecosystem, also with respect to the above mentioned value hierarchy.
- Emergence and resolutions of value conflicts by design (epistemic power of machine learning versus data protection, explainability, responsibility vs adaptability (and emergent properties).
- Theory and methods to deal with ethical dilemmas and value prioritization, ensuring that such decisions are open, transparent and amenable to argumentation and participation of a wide range of stakeholders.
- Moral importance of epistemic conditions for responsibility for design and use of AI systems (e.g. contextuality of notions such as 'understanding' 'explaining' and making 'transparent' the working of deep learning).
- Understand the relation of Humane-ness, human centeredness, human dignity in the application of AI.

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.
- The major research challenges are articulated in the following subsections.

Legal Protection by Design (LPbD)

Legal aspects will entail a focus on the preconditions for ethical conduct, for instance but not only: (1) accountability of those who take risks with other people's rights, freedoms and interests (by processing their data or targeting them based on data-driven inferences), (2) effective and meaningful transparency concerning the logic of automated decision systems that enables safe and meaningful interaction with such systems (both in the case of online search, social media and commerce and in the case of real-world navigation as in Internet of Things and robotics), (3) actionable purpose limitation to enable users (inhabitants) of AI environments to foresee the behavior of the myriad systems they may encounter (from connected cars to self-executing insurance contracts based on real-time data-driven input and care robots for the elderly), (4) reliable proportionality testing in the context of impact assessments, balancing the interests of providers against the rights, freedoms and interests of users or third parties that will suffer the consequences (whether algorithmic or data protection or safety and security impact assessment), in a way that enables them to contest such assessments, (5) built-in human-machine interaction that allows users or those targeted to exercise their fundamental rights and freedoms (enabling access to meaningful information, withdrawal from invasive targeting, detecting and contesting prohibited or unfair discrimination, and violations of the presumption of innocence).

This Legal Protection by Design (LPbD) entails the incorporation of fundamental rights protection into the architecture of AI systems. This plays out at two levels. -

1. This first concerns are the checks and balances of the Rule of Law, notably a concrete and effective set of interventions at the level of the research design, the subsequent development of core code, choice of programming language, foreseeable system behaviors, design of the APIs, and various types of interfaces. This concerns the choice architecture instituted by law that confronts (1) developers, (2) manufacturers, (3) sellers, (4) users (e.g. service providers, governments), and (5) end-users, providing them with leeway, proper constraints, transparency, accountability and foreseeability.
2. The second level concerns requirements imposed by positive law that elaborates fundamental rights protection, such as the GDPR, non-discrimination legislation, labour

law and the more. Here, the point is to follow up on concrete legal norms (e.g. the right to withdraw consent as easily as it has been given), translating them into technical requirements and specifications when developing applications. LPbD differs from ethics by design because it concerns legal obligations that are democratically legitimated and enforceable under the Rule of Law.

The choice of which norms must be built-in therefore does not depend on the ethical inclinations of e.g. developers or service providers, but on **constitutional preconditions** for ethical behavior (e.g. ensuring that those who act ethically will not be pushed out of the market), and on **enforceable law**.

LPbD differs from mere techno-regulation, 'legal by design', or 'compliance by design' because it does not aim to nudge or technologically enforce e.g. administrative law, trying to turn legal obligations into technical measures, but instead aims to build transparency, accountability and contestability into these systems enabling protection against "compliance by design."

The concrete results should consist of:

1) a listing of relevant design principles (including a detailed cross-disciplinary review by computer scientists and legal scholars) that concern the research design of machine-learning applications,
2) a set of case-studies demonstrating how LPbD principles can be integrated into the architecture of AI systems,
3) a dedicated assessment of how these principles interact with HMI design, suggesting new research lines on the cusp of machine learning, HMI, and law.

"Ethics by design" for autonomous and collaborative, assistive AI systems

Methods will be investigated that aim to understand how values can be *wired* into sociotechnical systems and what it means to do so. These may include (but are not limited to) studies **in compliance, security, data protection and privacy by design, fairness, explainability**, and how to implement these in combination with AI techniques and algorithmic governance through formal analysis and representation of regulatory principles, allocating rights, distributing liability, and ensuring legal protection by design.

A core challenge concerns the shaping of AI technologies and ecosystems, comprising autonomous and collaborative, assistive technology in ways that express shared moral values and ethical and legal principles as expressed in (but not limited to) binding EU legal treatises. This involves understanding, developing, and evaluating reasoning abilities of artificial autonomous systems (such as artificial agents and robots).

Even though AI systems are such that they allow us and even encourage us to defer to humans for decision making and performing actions that have grave moral impact, AI systems are artefacts and therefore are neither ethically or legally responsible. Individual humans or human corporations should remain the moral (and legal) agent. We can delegate control to purely synthetic intelligent systems without delegating responsibility or liability to them. To this effect, computational and theoretical methods and tools will be investigated, that support the representation, evaluation, verification, and transparency of ethical deliberation by machines with the aim of supporting and informing human responsibility on shared tasks with those machines.

Research is needed to discern suitable constraints on system behavior, and to elicit desiderata on the representation and use of moral values by AI systems. Furthermore, we need to provide design principles for meaningful human control over autonomous AI systems. This includes, but is not limited to, ensuring that privacy is respected, diversity is fostered in our communities,

discrimination and biases are avoided, societal and environmental well-being is respected, and basic rights and liberties are guaranteed.

An important topic is to boost research on developing tools for **discrimination and segregation discovery**, as well as **discovery and protection of novel vulnerabilities**. AI-based complex sociotechnical systems may amplify human biases present in data. Further, they may also introduce new forms of biases. As a result, AI-based systems may produce decisions or have impacts that are discriminatory or unfair, both under a legal or ethical perspective. Auditing AI-based systems is essential to discover cases of discrimination and to understand the reasons behind them and possible consequences (e.g., segregation). It may be that decisions informed by AI systems could have discriminatory effects, even in the absence of discriminatory intent. Moreover, discriminatory decisions take place on an individual in isolation, and **segregation** is the result of interactions among people in complex sociotechnical systems, nowadays largely governed by AI. Bias in AI systems can result in **both discrimination and forms of segregations.**

Explanation for high-stakes decision making. Decision making is essentially a sociotechnical system, where a decision maker interacts with various sources of information and decision-support tools, a process whose quality should be assessed in terms of the final, aggregated outcome—the quality of the decision—rather than assessing only the quality of the decision-support tool in isolation (e.g., in terms of its predictive accuracy and standalone precision). It is therefore important to develop **tools that explain their predictions in meaningful terms**, a property rarely matched by AI systems available in the market today.

The explanation problem for a decision-support system can be understood as **"**where" to place a boundary between what algorithmic details the decision maker can safely ignore and what meaningful information the decision maker should absolutely know to make an informed decision. Thus, an explanation is intertwined with trustworthiness (what to safely ignore), comprehensibility (meaningfulness of the explanations), and accountability (humans keeping the ultimate responsibility for the decision).

In this context, several questions emerge: what are the most critical features for explanatory AI? Is there a general structure for explanatory AI? How does an AI system reach a specific decision, and based on what rationale or reasons does it do so? Explanations should favor a human-machine interaction via meaningful narratives expressed clearly and concisely through text and visualizations, or any other human-understandable format revealing *the why*, *why-not*, and *what-if*.

Following the same line of reasoning, the AI predictive tools that do not satisfy the explanation requirement should simply not be adopted in high-stakes decision making, also coherently with the GDPR's provisions concerning the "right of explanation" (see Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR, which require data controllers to provide data subjects with information about "the existence of automated decision-making, including profiling and, at least in those cases, *meaningful information about the logic involved*, as well as the significance and envisaged consequences of such processing for the data subject.") The research challenges will intertwine greatly with the one of a "human-in-the-loop" line.

"Ethics in design"—methods and tools for responsibly developing AI systems

The real value of an AI system for decision support (e.g., based on machine learning, but not necessarily) is not in merely proposing an estimation on the probability that a certain relevant event will occur, or that the event is classified under a certain category, but requiring that guarantees are given that the system is developed and used in proper and verifiable ways.

This requires methods and tools for the **value-based design and development of AI systems** that ensure (a) the analysis and evaluation of ethical, legal, and societal implications; (b) the participation and integrity of all stakeholders as they research, design, construct, use, manage and dismantle AI systems; (c) the governance issues required to prevent misuse of these systems; and (d) means to inspect and validate the design and results of the system, such as formal verification, auditing, and monitoring [Durán2018].

Accountability. Accounting includes governance of the design, development, and deployment of algorithmic systems, which takes into consideration all stakeholders and interactions with socio-technical systems. Mitigating includes introducing techniques for data collection, analysis, processing that incorporate and acknowledge the systemic bias and discrimination that may be present in datasets and models; formalizing fairness objectives based on notions from the social sciences, law, and humanistic studies; building sociotechnical systems that incorporate these insights to minimize harm on historically disadvantaged communities and empower them; and introducing methods for decision validation, correction, and participation in co-designing algorithmic systems.

The aim is to boost research on theories, methods, and tools for trustworthy AI approaches, including ethics by design and ethics in design. This will ensure that AI systems are developed in a responsible, verifiable, and transparent way, while guaranteeing that their behavior is aligned with human values and societal principles such as privacy, security, fairness, or well-being. Naturally, users' requirements, legal requirements, and ethical requirements change over time, which necessitates dynamic, continuous evaluation and feedback throughout the system's entire lifecycle, thereby allowing participants to adapt their systems to their ever-evolving requirements.

4. RESEARCH METHODS FOR HUMANE AI

4.1 Challenge Based Research

We recommend that Humane AI employ a challenge-based research approach for developing the enabling technologies required for its vision. The consortium should maintain a continually evolving research road map that can be used to provide descriptions, specifications and examples for each required ability and research problem. For each ability, the community should identify and define technological challenges along with test data sets and measurable performance indicators to measure success. The research community should then be challenged to explore alternative technological approaches that satisfy the required ability. This is a well established scientific approach that is aligned with scientific practices in many other fields that combine technology development with scientific understanding.

4.2 Collaborative Microprojects

To maximize impact within the available resources, HumanE AI should focus on critical gaps in knowledge and technology at the interstice between the competences of the involved centers of excellence. This can be accomplished by organizing the research activities around the concept of "collaborative micro-projects."

Collaborative Micro-projects would involve a small group of researchers from several centers of excellence working together at a single location for a limited period of time to address a specific problem or provide a specific tool or dataset. HumanE AI can use micro-projects to go beyond networking, capacity building, and dissemination activities to implement key components of the proposed research agenda, organised dynamically within each workpackage.

Micro-projects should be required to produce a tangible result, such as a scientific publication, dataset, toolbox, demonstrator, or integration of a toolbox into the AI4EU. Each WP could have dedicated funds for micro-projects, which could be distributed through a lightweight internal proposal system based on quality and contribution to the WP agenda. Micro-projects could also be encouraged between WPs and could have the possibility of including external partners through appropriate mechanisms.

5. BIBLIOGRAPHY

[Banos 2012] Banos, O., Calatroni, A., Damas, M., Pomares, H., Rojas, I., Sagha, H., Mill, J.D.R., Troster, G., Chavarriaga, R. and Roggen, D., 2012, June. Kinect = imu? learning mimo signal mappings to automatically translate activity recognition systems across sensor modalities. In *2012 16th International Symposium on Wearable Computers* (pp. 92-99). IEEE.

[Bosser 2018] Bosser, A.-G., Bitoun, A. A., Legras, F., and Diéguez, M. (2018). Co-constructing subjective narratives for understanding interactive simulation sessions. In INTWICED@ AIIDE.

[Breazeal 2002] Breazeal, C. (2002). Designing sociable robots. MIT Press, Cambridge, MA.

[Breazeal 2016] Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). Social robotics. In Springer handbook of robotics, pages 1935–1972. Springer.

[Calegari 2019] Calegari, R., Ciatto, G., Dellaluce, J., and Omicini, A. (2019). Interpretable narrative explanation for ml predictors with lp: A case study for xai. In 20th Workshop From Objects to Agents (WOA), volume 2404, pages 105–112.

[Camerer 2002] Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. Journal of Economic theory, 104(1):137–188.

[Crandall 2018] Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., Rahwan, I., et al. (2018). Cooperating with machines. Nature communications, 9(1):233.

[Deng 2009] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE.

[Dignum 2019] Dignum, V., Responsible Artificial Intelligence, Springer, 2019

[Donadello 2017] Donadello, I., Serafini, L., and Garcez, A. D. (2017). Logic tensor networks for semantic image interpretation. arXiv preprint arXiv:1705.08968.

[Duran 2018] Duran, J. M. and Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds and Machines, 28(4):645–666.

[Fortes 2019] Rey, V.F., Hevesi, P., Kovalenko, O. and Lukowicz, P., 2019, September. Let there be IMU data: generating training data for wearable, motion sensor based activity recognition from monocular RGB videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (pp. 699-708). ACM.

[Garcez 2019] Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., and Tran, S. N. (2019). Neural-symbolic computing: An effec- tive methodology for principled integration of machine learning and reasoning. arXiv preprint arXiv:1905.06088.

[Garnelo2016] Garnelo, M., Arulkumaran, K., and Shanahan, M. (2016). Towards deep symbolic reinforcement learning. arXiv preprint arXiv:1609.05518.

[Gavrila 1999] Gavrila, D. M. , The visual analysis of human movement: A survey, Computer vision and image understanding 73 (1), 82-98, 1999

[Genette 2014] Genette, G. (2014). Discours du récit. Le Seuil.

[Georgeff 1998] Georgeff, M., Pell, B., Pollack, M., Tambe, M., and Wooldridge, M.. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages* (pp. 1-10). Springer, Berlin, Heidelberg, July 1998.

[Gervas 2019] Gervas, P., Concepci´on, E., Le´on, C., M´endez, G., and Delatorre, P. (2019). The long path to narrative generation. IBM Journal of Research and Development, 63(1):8–1.

[Gilpin 2018] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of inter- pretability of machine learning. In

2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pages 80–89. IEEE.

[Guidotti et al., 2019] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Gi- annotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5):93.

[Hildebrandt 2020] Hildebrandt, M., 11. Closure: on ethics, code and law; [Access: Nov 11 2019] Online https://lawforcomputerscientists.pubpub.org/pub/nx5zv2ux

[Hyndman 2012] Hyndman, K., Ozbay, E. Y., Schotter, A., and Ehrblatt, W. Z. (2012). Convergence: an experimental study of teaching and learning in repeated games. Journal of the European Economic Association, 10(3):573– 604.

[Jentner 2018] Jentner, W., Sevastjanova, R., Stoffel, F., Keim, D. A., Bernard, J., and El-Assady, M. (2018). Minions, sheep, and fruits: metaphor- ical narratives to explain artificial intelligence and build trust. In Workshop on Visualization for AI Explainability at IEEE.

[Johnson-Laird, 1989] Johnson-Laird, P. N. (1989). Mental models of meaning. MIT Press, Cambridge.

[Jorge 2019] Jorge, A. M., Campos, R., Jatowt, A., and Nunes, S. (2019). Information processing & management journal special issue on narrative ex- traction from texts (text2story): Preface.

[Kendon 1975] Kendon, A., Harris, R.M., & Key, M.R. (Eds). (1975). Organization of behavior in face to face interaction. The Hague, Netherlands: Mouton

[Koh 2017] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1885–1894. JMLR. org.

[Kriegeskorte 2018] Kriegeskorte, N. and Douglas, P. K. (2018). Cognitive computational neuroscience. Nature neuroscience, 21(9):1148– 1160.

[Lake 2017] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and brain sciences, 40.

[Lazer 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.- L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. Science, 323(5915):721–723.

[Lever 016] Lever, G., Shawe-Taylor, J., Stafford, R., and Szepesv´ari, C. (2016). Compressed conditional mean embeddings for model-based rein- forcement learning. In Thirtieth AAAI Conference on Artificial Intelligence.

[Marcus 2018] Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

[Marcus 2019] Marcus, G. and Davis, E. (2019). Rebooting ai. Ran- dom House Publishing.

[McGuire 1985] McGuire, W. J. (1985). Attitudes and attitude change. the handbook of social psychology, pages 233–346.

[Meghini 2019] Meghini, C., Bartalesi, V., Metilli, D., and Benedetti, F. (2019). Introducing narratives in europeana: A case study. International Journal of Applied Mathematics and Computer Science, 29(1).

[Nowak 2020] Nowak, A., Vallacher, R., Rychwalska, A., Roszczyska- Kurasika, M., Ziembowicz, K., Biesaga, M., and Kacprzyk-Murawska, M. (2020). Target in control: Social influence as distributed information processing. tba Springer.

[Nowak 2017] Nowak, A., Vallacher, R. R., Zochowski, M., and Rychwal- ska, A. (2017). Functional synchronization: The emergence of coordinated activity in human systems. Frontiers in psychology, 8:945.

[Pasquali 2019] Pasquali, A., Mangaravite, V., Campos, R., Jorge, A. M., and Jatowt, A. (2019). Interactive system for automatically generating tem- poral narratives. In European Conference on Information Retrieval, pages 251–255. Springer.

[Pearl 2018] Pearl, J. (2018). Theoretical impediments to machine learn- ing with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016.

[Pentland 2005] Pentland, A. and Madan, A. (2005). Perception of social interest. In Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI).

[Petty1986] Petty, R. E. and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In Communication and persuasion, pages 1–24. Springer.

[Peysakhovich 2018] Peysakhovich, A. and Lerer, A. (2018). Towards ai that can solve social dilemmas. In 2018 AAAI Spring Symposium Series.

[Rabinowitz2018] Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S., and Botvinick, M. (2018). Machine theory of mind. arXiv preprint arXiv:1802.07740.

[Reeves1996] Reeves, B. and Nass, C. I. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge university press.

[Roggen2010] Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A. and Doppler, J., 2010, June. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)* (pp. 233-240). IEEE.

[Russakovsky 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), pp.211-252.

[Satybaldiev 2019] Satybaldiev, A., Hevesi, P., Hirsch, M., Rey, V.F. and Lukowicz, P., 2019, September. CoAT: a web-based, collaborative annotation tool. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (pp. 814-818). ACM.

[Schank1975] Schank, R. C. and Abelson, R. P. (1975). Scripts, plans, and knowledge. In IJCAI, pages 151–157.

[Sun 2008] Sun, R. (2008). The Cambridge handbook of computational psychol- ogy. Cambridge University Press.

[Ta2015] Ta, V.-C., Johal, W., Portaz, M., Castelli, E., and Vaufreydaz, D. (2015). The grenoble system for the social touch challenge at icmi 2015. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 391–398. ACM.

[Toro Icarte 2018] Toro Icarte, R., Klassen, T. Q., Valenzano, R., and McIlraith, S. A. (2018). Teaching multiple tasks to an rl agent using ltl. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pages 452–461. International Foundation for Au- tonomous Agents and Multiagent Systems.

[Urbaniak 2018] Urbaniak, R. (2018). Narration in judiciary fact-finding: a probabilistic explication. Artificial Intelligence and Law, 26(4):345–376.

[Van den Hoven 2015] Van den Hoven, Jeroen; Vermaas, Pieter E; van de Poel, Ibo (Ed.) (2015) Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains, Springer

[VanHarmelen 2019] van Harmelen, F. and Teije, A. t. (2019). A boxology of design patterns for hybrid learning and reasoning systems. arXiv preprint arXiv:1905.12389.

[Vlek2016] Vlek, C. S., Prakken, H., Renooij, S., and Verheij, B. (2016). A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law, 24(3):285–324.

[Wang2017] Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering, 29(12):2724–2743.

[Welbourne 2014] Welbourne, E. and Tapia, E.M., 2014, September. CrowdSignals: a call to crowdfund the community's largest mobile dataset. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (pp. 873-877). ACM.

[Yang2019] Yang, Y. and Song, L. (2019). Learn to explain effi- ciently via neural logic inductive learning. arXiv preprint arXiv:1910.02481.