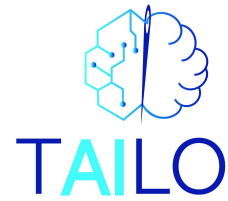


Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR

Grant Agreement Number 952215

Foundations, techniques, algorithms and tools to for social AI v.1

Document type (nature)	Report
Deliverable No	D6.1
Work package number(s)	WP6
Date	30/05/2022
Responsible Beneficiary	IST
Author(s)	Ana Paiva
Publicity level	Public
Short description	A report on advances developed within T6.1, T6.2, T6.3 and T6.4

History			
Revision	**Date**	**Modification**	**Author**
1	06/06/2022	-	Ana Paiva
1	2023-01-10	Resubmitted on formal grounds	Ana Paiva

Document Review		
Reviewer	**Partner Acronym**	**Date of report approval**
Giuseppe De Giacomo	Uni Roma	220607
Joaquin Vanschoren	TU/e	220607

Table of Contents

Summary

This document describes recent work by the TAILOR partners involved in WP6, that aims to build trustworthy Social AI by integrating reasoning, learning and optimization mechanisms in contexts where more than one agent is present.

Social and organisational aspects of AI have become major research topics in our field, embracing the areas of multi agent-systems, human-agent interaction, network systems, game theory, social learning, and more. Social AI focuses on techniques to build AI that emerges, is situated, and is able to perform in social contexts, and eventually support the creation of hybrid populations that may include not only agents (AI systems) but also humans. But as AI is developed with the goal of acting in social contexts, is distributed, and placed within hybrid populations of humans and machines, several challenges emerge. How can agents communicate, negotiate and reach agreements in a trustworthy manner? How can agents take into account others (including humans) and establish trustworthy relationships among them? How do networks of agents and humans evolve, and does trust play a role and evolves as well? How are teams created and maintained in hybrid populations of humans and agents? Can we trust systems where AI is distributed? And what foundational methods do we have to guarantee trust in them?

Many of these challenges have been addressed during these first eighteen Months of collaboration in the WP6 of the TAILOR network. We are embracing these challenges by combining efforts from the community to increase the knowledge and expertise to promote and develop “trustworthy social AI”.

This deliverable is the first deliverable of this work being done in WP6, establishing some foundations for the field, discussing techniques, algorithms and tools to build and evaluate trustworthy social AI. We also report on the networking activities carried out as a network to achieve these outcomes.

Introduction

About the document

This document represents deliverable D6.1 in TAILOR WP6, 'Social AI: learning and reasoning in social contexts. It provides the first version of the work carried out in WP6. The document is the result of the different tasks in the workpackage involving different organisations and researchers in Europe associated with the TAILOR project.

The document is organised as follows: first, we start with the major idea and definitions within the field, posing some of the main questions, and providing an overview of research topics that characterise the social aspects of trustworthy AI.

Then we discuss some of the challenges related with the 4 scientific tasks and provide some examples of work done in these tasks, by discussing concrete results from the partners contributing to the project.

Then we provide a brief overview of how we work together, and report on the different activities that we have organised in the past year and a half.

Finally, we provide some future research agenda (towards a roadmap of trustworthy AI) and how the community can get together to achieve our goals.

We also provide a list of recent publications by our partners.

Organisation

The following people have been involved in the Deliverable:

Name	Partner ID
Francisco Santos	IST-UL
Alberto Sardinha	IST-UL
Carles Sierra	IIIA-CSIC
Michael Wooldridge	UOX
Ann Nowe	VUB
Vito Trianni,	CNR
Wico Mulder ,	TNO
André Meyer-Vitali	DFKI
Sarit Kraus	BUI
Tom Lenaerts	VUB
Elias Domings	VUB

Part 1. Towards Trustworthy Social AI

On the 3rd of August 1994, Prof. Barbara Grosz, one of the most influential researchers in Artificial intelligence (AI), delivered her AAAI 1994 presidential address on the topic of "Collaborative Systems"[1] proposing a new vision where AI should be collaborative. At the time, AI was suffering from the second AI winter, where funding had decreased significantly, credibility and trust was at its minimum, and many researchers in other fields considered that AI would not be able to deliver anything of value. In spite of this, AI researchers were confident that AI technology could play a major role, and obtain effectiveness levels that would allow it to make a huge impact. The inspiring work by B. Grosz and collaborators gave the field

[1] Grosz, B. J. (1996). Collaborative systems (AAAI-94 presidential address). *AI magazine*, *17*(2), 67-67.

a refreshing and important vision, by showing the need for AI to be situated in the environment, able to use data in a dynamic way, capable of interacting with humans, and most importantly, make decisions collaboratively.
Some years later, the field of multi-agent systems was growing, and the first edition of the landmark book on multiagent systems by M. Woolridge established a roadmap for the field of multiagent systems[2]. There, two main visions were proposed: (1) agents as a paradigm for software engineering; and (2) agents as a tool for understanding human societies.

Almost thirty years since that memorable presidential address, AI has become increasingly more present in our daily lives. A myriad of settings became the stage for AI applications, such as factories, roads, houses, hospitals and even schools. Given these new contexts, A. Paiva recently stressed that AI-powered machines must now place humans at the centre and are designed to interact with humans naturally: AI is becoming social[3]

We believe that there is now a place for a vision anticipated by Grosz and reiterated by Wooldridge and Paiva, that regards AI situated in social contexts, and agents are AI entities that cooperate and communicate in hybrid populations of humans and agents.

But such a diverse use of AI also fosters change, especially in the way we behave and how we interact with each other and with machines. It is essential to reflect upon AI's impact on humans' societies and consider its effects, e.g., on supporting more collaboration, social action, and prosocial behaviour. Thus, the presence of machines in social settings raises the need for a better understanding of their effect on social interactions and how they may be used to influence human behaviours.

The TAILOR network, and this WP in particular, combines work by many scientists exploring the use of AI techniques towards a better understanding of social interactions in nature and societies, but also how to create AI (agents) that places social behaviour at the core and, while engaged in human settings, fosters cooperation and collective action.

1.1 Social AI what it is?

Social AI focuses on techniques to build AI (agents) that models, emerges, is situated, and able to perform in social contexts, acting in populations that may include not only agents but humans as well.

This entails different dimensions of study:

[2] Wooldridge, Michael. *An introduction to multiagent systems*. John wiley & sons, 2009.
[3] Paiva, A. (2022, March). From Social to Prosocial Machines: A New Challenge for AI. In *27th International Conference on Intelligent User Interfaces* (pp. 2-2).

- D1. AI used for understanding social interactions, cooperation, coordination, organisations, and norms;
- D2. AI that exhibits social competencies - Agents/AI that are able to interact in a social manner (including social perception, understanding, groups dynamics, etc).;
- D3. AI for modelling strategic decision-making through game theoretical approaches (both non-cooperative and cooperative game theory);
- D4. AI that captures the dynamics of social interactions in large simulated and hybrid societies;
- D5. AI that performs in social contexts and impacts the social environment we live in.

1.2 Domain applications for Social AI

Social AI systems are being used in many domains. In healthcare we see that AI systems are in dialogue with humans to detect and analyse cancer cells, as well as systems that suggest diagnoses in more general clinical settings. In addition, social AI is more and more supporting humans in self-care and prevention.

Precision agriculture and dairy farming is an ever growing application for AI, which can also be a tool to optimise production, make predictions, and support farmers. To do so, systems need to be distributed, cooperative, and interact with humans (farmers, operators). There, humans should be able to collaborate with machines by tuning the model parameters in the AI systems that are used for crop production and cattle management. Also the traffic and transport sector uses interaction/dialogue based mechanisms in their traffic management systems.

Social AI systems are also used in the energy sector, e.g. citizens optimise the energy consumption in buildings, decide when to share their cars. In the near future, buildings will share information between each other to learn and collaborate in energy management towards the local power grids. This will be extended at smart city level and beyond (e.g. connection to windmill parks or heat nets) and efficient building occupancy management. Leading to a big interconnected but distributed network of socio-technical systems. Different applications in this sector have been carried out by partners in the network. In particular in the use of agent-based simulation for policy making in urban planning (see Task 6.4 for a short description). Experiments have started in the field of law enforcement, for example to use federated reasoning mechanisms to gain a better understanding of debt problems or resolve cold cases.

One of the generic areas of application of social AI, useful in all industrial domains, is modelling and simulation. It concentrates on observing the behaviour of humans or systems of agents in order to better understand that behaviour in terms of derived rules and patterns in the underlying mechanisms (e.g. modelling negotiation mechanisms, decision making or group formation). The insights obtained, i.e. the retrieved rules and models, can be used in simulations and implemented in real-life applications. One sees this already in e.g. multi robot task allocation in search and rescue contexts, traffic management and control in smart cities or advanced planning in digital manufacturing factories.

The area of media is another sector where AI is making its impact. Like in the TV entertainment sector, humans pass preferences and systems classify and personalise their interaction. Content production can be seen as a collaborative process carried out by AI agents (systems) and humans.

1.3 How to achieve trustworthy Social AI

One of the main outcomes expected from TAILOR is the capacity of providing the scientific basis for Trustworthy AI. The economic potential of AI and algorithms is huge, but it will only be successful if it meets the safety, security and ethical requirements posed by our society. The goal is to guarantee the creation of explainable, fair, safe and accountable systems. Explainability, Safety, Fairness, Accountability, Privacy, and Sustainability are the dimensions of Trustworthy AI that are necessarily intertwined with the foundation themes of TAILOR.

So, one of the challenges we have been considering in this workpackage is: *How can we achieve trustworthy Social AI?*

Trust is a complex multidimensional process that does not encompasses the dimension "competence" but it also captures different other phenomena. In social contexts, "trust" describes the process by which humans establish relations with others or entities, attribute them some characteristics, and as a result, something is expected (an outcome O), and hopefully, achieved[4][5].
Trust emerges as a result of beliefs that we hold about the others (other agents), the environment and the objects and products there-within. For example, we, humans, believe that the kitchen chef is able to prepare us a fantastic meal as she is an expert in cooking, so, we trust her to prepare such a meal for a special guest we have in the house. Different levels and types of trust have been identified, namely, executive trust and moral trust. Executive trust, means an agent T (the trustor) believes that the other agent (the trustee) is able to perform a task. Moral trust is somehow different, and is related with the belief that the trustee is guided by moral principles, such as honesty, sincerity, benevolence.

Trust is also associated with a risk. Without the risk that the trustee will not achieve the results, there is no need for trust, that is, there is no trust involved. The trustor needs to be in a position of vulnerability in relation to the trustee performing some action.

So, in general, **trust can be defined as the trustor willingness to be vulnerable to the actions of the trustee**. However, when we go beyond an AI system and a human, and consider many potential agents, some more competent than others, or consider the relationship between humans and the system, new challenges emerge.

[4] Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).
[5] Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In *Trust and deception in virtual societies* (pp. 55-90). Springer, Dordrecht.

Thus, to discuss trust in social contexts where AI is developed, we need to consider several dimensions: (1) trust between humans and a multi-agent system; (2) trust between agents in a system; and (3) trust inherent to a socio-technical system (trustworthiness).

So, how do we guarantee trust in these contexts? One way is to bridge the gap from formal methods, verification as well as validation to the way multi-agent systems are engineered, used, and reinforced, as to guarantee trust. This aspect has been addressed in our 4 tasks. Another way is by considering that AI is a collaborative partner to humans that has the social competencies to provide explanations for understanding its decision making process [Miller2019] (addressed in task 6.1). Another approach is to provide (automated) mechanisms to study and understand the complex behaviour observed in a potentially mixed population of artificial and human agents, starting from learned subsymbolic representation of behaviour (policy), finding symbolic categorisations (reaching agreement on abstractions of the policy), to allow for reasoning, communication, explanation and verification.

Our four scientific tasks (T6.1-T6.4) contribute to addressing some of these challenges.

Part 2. Scientific challenges and work carried out

This part overviews the scientific contributions that have been made by different partners, while addressing the challenges proposed.

We have adopted a generic framework for social AI where agents constitute the members of a networked hybrid society. Each agent (which can be a human) is endowed with the capability to perceive the social context (social perception), interact with other agents through social signals (signalling), communicate, delegate, negotiate and eventually cooperate with each other. The agents should be able to act upon the world (link to WP5) and their decision making is based on some model of representation (link to WP4). As agents act, trust relationships emerge (link to WP3).

Figure 2.1 – Overview of a generic framework for Social AI: each agent is captured as an entity that is able to perceive the world and act upon it, and its decision making can be a result of different techniques and algorithms. Each agent must perceive others, communicate, negotiate and make strategic decisions.

This work was carried out within 4 interconnected scientific tasks:

- Task 6.1 - Modelling social cognition, collaboration, argumentation and teamwork
- Task 6.2 - Theoretical models for cooperation between agents
- Task 6.3 – Learning from others
- Task 6.4 - Emergent Behaviour, agent societies and social networks

Furthermore, the work also addressed concrete real world applications (reported in Task 6.5).

2.1 Foundations for modelling social cognition, collaboration and teamwork (T6.1)

One vision of social AI is AI that plays the role of a collaborative partner to humans. This leads to the broad exploration around "Human-agent teams", a widely used term referring to groups containing at least one human and one autonomous agent (or autonomous system), that form an alliance and work together towards achieving a common goal. It is generally accepted that as agents work together with humans, they should be governed by the same principles that underlie human-human collaboration[6] and as such, human-agent teams are very much inspired by human teams. Yet, it is not clear if human-agent teams will work at all. First, the capabilities of the agents in the teams are often limited, not only in concrete tasks execution, but most importantly in their capabilities for social interactions. Agents so far still do not truly understand others, are unable to interact in a natural way, to understand the intentions of others, or to put themselves in their position (exhibiting a Theory of Mind capability).

One essential aspect of teamwork is collaboration. Collaboration according to Roschelle and Teasley is a "mutual engagement of participants in a coordinated effort to solve a problem together,"[7]. For example, a team of doctors and nurses working in a surgery to operate a patient. Or a team of firefighters, medics, and civil population combating a fire, are all examples of collaborative situations, where the main goal requires the actions and competencies of the diverse team of members.

Collaboration is essential for intelligent behaviour, and as machines are placed in these social settings, they are expected to be able to collaborate with others, and form a team. According to B. Grosz[8], "focusing on the scientific underpinnings of collaborative AI has two main advantages: first it allows for the development of theories and formalizations that are needed to build collaborative systems". These fundamental questions and theories embrace problems and raise questions to different fields of research in AI, namely NLP, Robotics, ML, Planning, Reasoning, and so on. Secondly, the results that can be achieved when grounding research on theories about collaboration may lead to a significant impact not only in AI and computer science but also in other areas, such as social sciences, health education, logistics, criminal justice and many others. The range of domains of application for this approach is vast. Additionally, there has been a recent realisation in "the AI community that new AI systems built for this day and age need to be inherently social"[9].

Moreover, the competencies that AI has may be excellent in one task but rather poor in another. And human partners may be the opposite. For example, a robot helper in a building may be very competent in knowing who inhabits each room of the building, and able to move

[6] Rich, Charles, and Candace L. Sidner. "COLLAGEN: When agents collaborate with people." In *Proceedings of the first international conference on Autonomous Agents*, pp. 284-291. 1997.

[7] Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning* (pp. 69-97). Springer, Berlin, Heidelberg.

[8] Grosz, B. J. (1996). Collaborative systems (AAAI-94 presidential address). *AI magazine*, *17*(2), 67-67.

[9] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. *Nature*, *593*, 33–36.

in the corridors swiftly, but it may not be able to move between floors as it does not have the power to go up and down stairs, nor the arms to call an elevator. Humans, on the other hand, do not know who is who in the building, but are perfectly capable to take the elevator to the 7th floor, and help the robot to do the same.

So, collaboration assumes that:
- there are different participants (often with different competences and knowledge);
- there is mutual engagement of the participants;
- there is a problem that all want to solve; and
- there is a coordinated effort to solve that problem together.

In this task we have been studying ways to model an agent's **cognitive capabilities** that integrate individual knowledge and behaviour with knowledge available to and from other agents (possibly obtained at different times and from different perspectives).

Our recent work has tackled these questions of collaboration from three perspectives [10] [11] [12] [13] [14] [15]. The first is concerned with "agent-agent" collaboration, and leverages *norms* and *rules* as constructs that, when implemented on a multiagent system, can help foster cooperative and socially beneficial interactions among agents. The main contribution in this direction is the development of a computational model of the Institutional Analysis and Development (IAD) framework[16], a well-established theory from the social sciences and policy analysis literature that outlines the universal components that make up any social interaction. Within the IAD framework, one of the main components that structure a social interaction are the rules in place. Furthermore, rules are relatively easy to change in the short term, facilitating for a team to adapt to new conditions or prioritise the achievement of a new goal.

Following this lead, we have developed the Action Situation Language (ASL),[17] a logical language implemented in Prolog that allows to write in a structured syntax the rules that a team of agents is pondering on implementing. The ASL is complemented by a game engine that takes as input the description of an interaction and automatically builds a model of the resulting interaction as an extensive-form game, which can later be analysed using standard game-theoretical solution concepts. This way, a community of agents can draft new rules,

[10] Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C. (2021). Towards a Competence-Based Approach to Allocate Teams to Tasks. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1504–1506.
[11] Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C. (2022a). Building Contrastive Explanations for Multi-agent Team Formation. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 516–524.
[12] Georgara, A., Rodríguez-Aguilar, J. A., Sierra, C., Mich, O., Kazhamiakin, R., Palmero-Approsio, A., & Pazzaglia, J.-C. (2022b). An Anytime Heuristic Algorithm for Allocating Many Teams to Many Tasks. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1598–1600.
[13] Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C. (2021). Towards a Competence-Based Approach to Allocate Teams to Tasks. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1504–1506.
[14] Montes, N., Osman, N., and Sierra, C. (2021). *Enabling Game-Theoretical Analysis of Social Rules* (Vol. 339, pp. 90–99). IOS Press.
[15] Montes, N., Osman, N., and Sierra, C. (2022). *Combining Theory of Mind and Abduction for Cooperation under Imperfect Information [Under review]*.
[16] Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.

[17] https://www.ai4europe.eu/research/ai-catalog/ngames

examine their effects in an automated fashion, and assess whether their adoption is desirable. The decision to adopt a new set of regulations can be made from the perspective of personal and/or team gains (and trade-offs among these), and the social benefits of the most likely outcomes. A publication detailing the technical aspects and examples using the ASL tool is currently under review. A conference paper[18] (Montes 2021) presents some preliminary results.

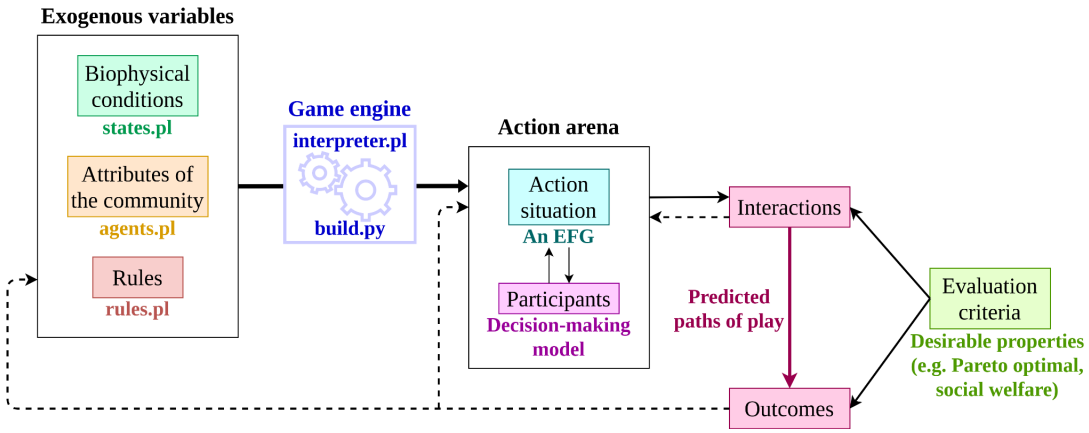


Figure 2: Outline of the IAD framework. Adapted from Ostrom 2005.

Second, we are at the preliminary stages of developing a cognitive model for teams of agents in cooperative domains characterised by imperfect information, i.e. where agents do not have complete access to the current state of the system and hence must rely on their peers to act correctly[19]. This agent model is based on the combination of Theory of Mind (ToM) and abductive reasoning. Generally, ToM refers to the cognitive ability to put oneself in the shoes of others and reason about their mental attitudes, such as their beliefs, intentions, emotions, and so on. Meanwhile, abduction is a logical reasoning paradigm that computes explanations from observations made in the environment [20] by inferring what information constitutes a valid basis for the observed knowledge to hold true.

In our agent model, ToM is utilised by observer agents to adopt the perspective of an acting agent who has just performed some action. Abduction, then, is used to derive the knowledge that the acting agent may have been relying upon in order to decide on the action they have just executed. This abduced knowledge takes the form of explanations that can then be added to the observer agent's knowledge base to be leveraged during their own decision-making. We have successfully implemented this agent model using Jason, an agent-oriented programming language based on the Belief-Desire-Intention (BDI) architecture. We have tested our implementation in the Hanabi game domain, a cooperative card game that has recently

[18] Montes, N., Osman, N., and Sierra, C. (2021). *Enabling Game-Theoretical Analysis of Social Rules* (Vol. 339, pp. 90–99). IOS Press.
[19] Montes, N., Osman, N., and Sierra, C. (2022). Combining Theory of Mind and Abduction for Cooperation under Imperfect Information [Under review].
[20] Denecker, M., and Kakas, A. C. (2002). Abduction in Logic Programming. *Computational Logic: Logic Programming and Beyond, Essays in Honour of Robert A. Kowalski, Part I*, 402–436.

attracted a lot of attention from the AI community[21], with satisfactory preliminary results. Further work in this direction will explore the trade-offs between the computational requirement and the performance gains of employing deeper recursion levels in our ToM-abduction agent model, as well as provide a full domain-independent open-source implementation.

Apart from "Agent-Agent" and "Human-Agent" teams, social AI can be of great assistance to boost the performance of *human collaboration*. It is commonly accepted that putting together the right people to jointly work as a team on some task is a hard and time-consuming thing to do. Human resources in companies, managers in organisations and institutions, or even teachers at schools usually spend a lot of working hours in order to find a combination of people that not only can cope with the task at hand but also can stick together as a group; let alone when there is need for more than one such teams to be formed. People usually adopt heuristics that allow them to spot potentially good teams, which over the years have been theoretically established in scientific areas such as Organisational Psychology and Social Sciences. In this light, social AI can gather findings from the aforementioned scientific fields regarding human collaboration, and assist people that need to form teams by considering as many of these findings as possible to speed up the procedure.

In this task we have been also studying the problem of ***human team formation and task allocation,*** which is the formation of human teams that need to be matched with tasks to solve. Many real-world problems require allocating teams of individuals to tasks. For instance, building teams of people to perform projects in a company[22], or grouping students to undertake school projects[23]. These problems have in common that they are allocation of many teams to many tasks (with size constraints), that usually permits no overlaps. That is, each individual can be part of at most one team, each team can be allocated to at most one task, and each task must be solved by at most one team (at a time). We have illustrated our results in the domain of education motivated by the hard and time-consuming procedure of allocating student teams to school projects or internship programs. Currently, teachers and education authorities obtain such allocations mainly by hand, but given the combinatorial nature of the problem, manual allocation requires a large amount of work. Moreover, a manual allocation is very likely not to find a good solution given the size of the problem.

Our study regards the development of an anytime heuristic algorithm that forms teams and matches the teams with tasks considering findings from Psychology and Social Sciences. Our algorithm moves along four dimensions that influence a team's performance: (i) team's collectively acquired competencies / skills / knowledge with respect to the task to be solved; (ii) balance of team members' in terms of personality[24]; (iii) team's interest (collectively) towards

[21] Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., and Bowling, M. (2020). The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, *280*, 103216.

[22] Sa Silva, I. E., & Krohling, R. A. (2018). A fuzzy sociometric approach to human resource allocation. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1-8.

[23] Andrejczik, E., Bistaffa, F., Blum, C., Rodríguez-Aguilar, J. A., & Sierra, C. (2019). Synergistic team composition: A computational approach to foster diversity in teams. *Knowledge-Based Systems*, 182, 104799.

[24] Belbin, R. (1993). *Team Roles at Work: A Strategy for Human Resource Management*. Butterworth-Heinemann.

the task to be solved[25]; and (iv) team's social cohesion[26] . One of the main components of our approach is that we adopt the concept of similarity among different competencies and the use of structure competence ontologies such as the ESCO ontology (https://esco.ec.europa.eu/en). Our algorithm exploits the four dimensions mentioned above, and combines them in order to form an effective team for each task at hand. We have been using this algorithm to form teams in university classes in order to tackle a semester project. Two conference papers (extended abstracts[27] [28]) presenting the main aspects of our work and outlining our algorithm have been published; while another publication (journal paper) presenting our findings from experimenting with schools is currently under review.

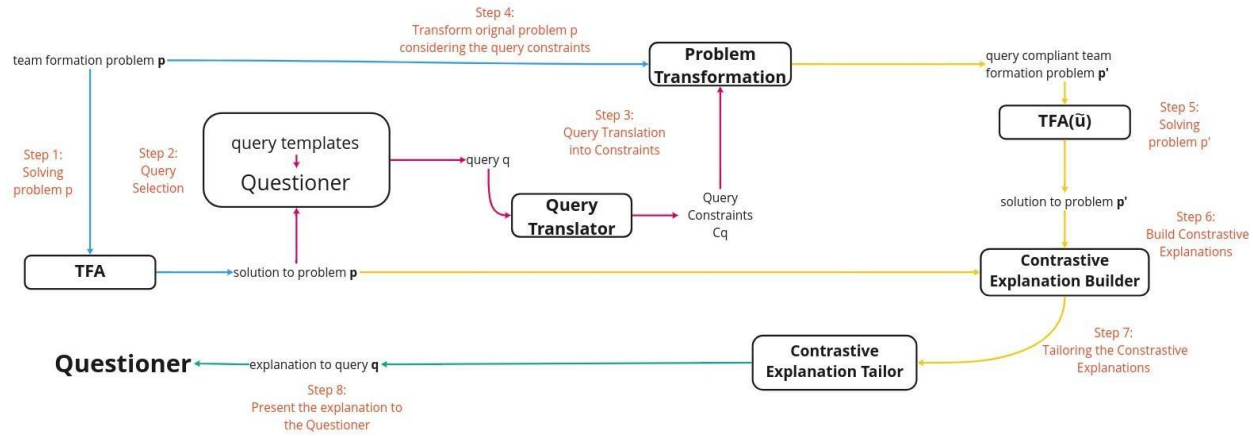


Figure 3: General Justification Algorithm for Team Formation.

One step further, recognising the importance of earning the trust of users we have been working towards a general framework to provide justifications (or explanations) for team formation and task allocation. This framework provides a collection of thirteen intuitive and meaningful questions that cover the main points of interest regarding team formation scenarios. Given this question collection, we have developed a general justification algorithm (illustrated in Fig.3) that wraps existing team formation algorithms and builds contrastive explanations. Such explanations answer to "what would have happened if…" kind of questions and justify why one solution is better than another. alternative one. Finally the explanations built are being tailored to highlight different perspectives by focusing on (i) a small subset of participants, (ii) each individual task, or (iii) the overall matching of teams to tasks. A conference paper[28] detailing our algorithm for contrastive explanations for team formation scenarios, and presenting preliminary results of our work has been published.

[25] Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The Motivation to Work*. John Wiley & Sons.
[26] Randall, L. H., & Kuhnert, K. W. (1993). Using Sociometry to Predict Team Performance in the Work Place. *The Journal of Psychology*, *131*, 21-32.
[27] Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C. (2021). Towards a Competence-Based Approach to Allocate Teams to Tasks. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1504–1506.
[28] Georgara, A., Rodríguez-Aguilar, J. A., Sierra, C., Mich, O., Kazhamiakin, R., Palmero-Approsio, A., & Pazzaglia, J.-C. (2022b). An Anytime Heuristic Algorithm for Allocating Many Teams to Many Tasks. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1598–1600.

2.2. Theoretical models for cooperation between agents (T6.2)

The importance of theoretical models, in particular the use of game theory, population dynamics, and multiagent systems to study strategic decision making in Social AI is testified by the numerous high-level publications that have enriched the field in the last 20 years, which go well beyond standard multiagent systems and AI publication venues given its interdisciplinary flavour and implications.

In this task we present results from three main processes that have been addressed from the partners: delegation, cooperation and explanation. In multiagent systems, agents may delegate tasks into others. When humans are involved, this delegation process is dependent on the trust relationships between humans and agents.

A delegation problem is defined to capture a situation where a "principal" has to delegate decisions to a set of agents[29]. The principal, himself, has his own interests in terms of decision making. Thus, delegation must be done in a way that, if all the agents to whom decisions have been delegated make their respective decisions rationally, the principal's goal will be achieved in equilibrium. Once the decisions are delegated, the agents will act "selfishly, rationally, and independently" in pursuit of their own preferences, and yet, guaranteeing that the goal of the principal is achieved.
A formalisation of this delegation problem is done using Boolean games. In a Boolean game agents are players, and each player i is assumed to have a goal (γi), represented as a propositional formula γi over some set Φ of Boolean variables intuitively represent the space of potential choices/strategies by all the agents. Each agent controls some of these variables, a subset Φi of the total variables Φ, with the idea being that such variables Φi are under the unique control of that particular player i. Using Boolean games as a way to capture this problem, two types of delegation were defined: strong delegation, and weak delegation. Intuitively, strong delegation requires that the objective pursued by the principal is satisfied in all Nash equilibria of the Boolean game that results from an allocation, whereas the weak allocation one requires that one allocation exists such as the goal is satisfied in at least one Nash equilibrium of the Boolean game. More recently, Dunne and colleagues[30] studied how this principal delegation problem compares to an alternative delegation model, a distributed delegation problem, which captures a more cooperative setting, where the agents have to assign responsibilities among one another in the absence of a principal.

Situations where the decision making can be modelled as a Boolean game and the decision broken into a set represented as Boolean variables, delegation is thus equated as the problem of finding the best allocation for agents to make decisions that guarantees that their rational decisions will lead to the goal to be achieved in equilibrium. Note, that this problem of delegation has also been addressed by looking at ways by which people are able to delegate into AI systems (see discussion in part 2.3).

[29] Kraus, Sarit, and Michael J. Wooldridge. "Delegating Decisions in Strategic Settings." In *ECAI*, vol. 12, pp. 468-473. 2012.
[30] Dunne, Paul E., Paul Harrenstein, Sarit Kraus, and Michael Wooldridge. "Delegating Decisions in Strategic Settings." *IEEE Transactions on Artificial Intelligence* 1, no. 1 (2020): 19-33

Another challenging problem to address is cooperation of self-interested agents that need to face a joint enemy. Multi-defender Stackelberg Security Games (MSSG) have recently gained increasing attention in the literature for studying these challenges. Coordination and cooperation between the defenders in such games can increase their ability to protect their assets, but the heterogeneous preferences of the self-interested defenders often make such cooperation very difficult. However, the solutions offered to date are highly sensitive, wherein even small perturbations in the attacker's utility or slight uncertainties thereof can dramatically change the defenders' resulting payoffs and alter the equilibrium. Matzuri et al introduced a robust model for MSSGs[31], which admits solutions that are resistant to small perturbations or uncertainties in the game's parameters. Mutzari et al presented a formal definition of the notion of robustness, as well as the robust MSSG model[32]. There are two approached for modeling cooperation in multi-agent problems: non-cooperative setting and cooperative settings. For the non-cooperative settings they proveed the existence of a robust approximate equilibrium in any such game, and provide an efficient construction thereof. For the cooperative setting, they proved that any such game admits a robust approximate alpha-core, provided an efficient construction thereof, and proved that stronger types of the core may be empty. Interestingly, the robust solutions can substantially increase the defenders' utilities over those of the non-robust ones.

Another important topic for trustworthiness concerns the capability to generate explanations by an AI system. In fact, explanation is necessary for humans to understand and accept decisions made by an AI system when the system's goal is known. It is even more important when the AI system makes decisions in multi-agent environments where the human does not know the systems' goals, since they may depend on other agents' preferences. In such situations, explanations should aim to increase user satisfaction, taking into account the system's decision, the user's and the other agents' preferences, the environment settings and properties such as fairness, envy and privacy. We studied the problem of distilling a policy learned by a deep RL agent, hereby generating explanations that can gradually zoom in to reveal more details[33], and two problems of Explainable decisions in Multi-Agent Environments (xMASE): explanations for multi-agent Reinforcement Learning and justifications for social-choice mechanism outcome. For each case we presented an algorithm for generating the explanations and reported human experiments that demonstrate the benefits of providing the resulting explanations for increasing human satisfaction from the AI system.

[31] Dolev Mutzari, Jiarui Gan and Sarit Kraus. Coalition Formation in Multi-defender Security Games, AAAI 2021.
[32] Mutzari Yonatan Aumann and Sarit Kraus Robust Solutions for Multi-Defender Stackelberg Security Games, IJCAI 2022.
[33] Coppens, Y., Steckelmacher, D., Jonker, C. M. & Nowe, A., "Synthesising Reinforcement Learning Policies Through Set-Valued Inductive Rule Learning", 13 Apr 2021, Trustworthy AI - Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers. Heintz, F., Milano, M. & O'Sullivan, B. (eds.). 1 ed. Cham: Springer International Publishing, p. 163-179 17 p. (Lecture Notes in Computer Science; vol. 12641).

For explanation of social-choice mechanism outcomes, in Suryanarayana et al.[34] [35] proposed a methodology for automatically generating explanations based on desirable mechanism features found in theoretical mechanism design literature is presented. Human experiments reveal that explanations affect both average satisfaction from and acceptance of the outcome in such settings. In particular, explanations are shown to have a positive effect on satisfaction and acceptance when the outcome (the winning candidate in our case) is the least desirable choice for the participant. A comparative analysis with human generated explanations reveals that the automatically generated explanations result in similar levels of satisfaction from and acceptance of an outcome as with the more costly alternative of crowdsourced explanations, hence eliminating the need to keep humans in the loop. Furthermore, the automatically generated explanations significantly reduce participants' belief that a different winner should have been elected compared to crowdsourced explanations.

For explaining multi-agent Reinforcement Learning, Boggess et al. [36] presented novel methods to generate two types of policy explanations for MARL: (i) policy summarization about the agent cooperation and task sequence, and (ii) language explanations to answer queries about agent behavior. Experimental results on three MARL domains demonstrate the scalability of the proposed methods. A user study shows that the generated explanations significantly improve user performance and increase subjective ratings on metrics such as user satisfaction.

To complement these theoretical approaches, partners at VUB have also been combined with development work, and we have created a public game theory library for multiagent systems simulations. VUB and IST partners are developing a new, efficient C++/Python public library that provides fast implementations in C++ of the Monte-Carlo simulations and the most recent analytical approaches necessary to estimate many important indicators such as stationary or strategy distributions associated with massively large multiagent systems. The results of this effort are currently under review [Domingos et al, 2022] and we expect to add these tools to the AI4EU depository.

2.3 Learning in Social Contexts (T6.3)

As the number of agents increases in our everyday environment, many scenarios (e.g., healthcare, search-and-rescue teams, warehouse management) will require agents to learn how to collaborate with humans within a social context. Hence, learning within a social context is a critical element for social AI. In particular, our work focuses on how agents can learn to cooperate within a social context, whereby human-agent teams have to achieve a common goal. In this task, we have been exploring two stands of work.

[34] Suryanarayana, Sharadhi Alape, David Sarne, and Sarit Kraus. Information Design in Affiliate Marketing. Autonomous Agents and Multi-Agent Systems 35,(2):1-28, 2021.
[35] Sharadhi Alape Suryanarayana, D. Sarne, S. Kraus. Justifying Social-Choice Mechanism Outcome for Improving Participant Satisfaction AAMAS 2022.
[36] Kayla Boggess, Sarit Kraus and Lu Feng. Toward Policy Explanations for Multi-Agent Reinforcement Learning, IJCAI 2022.

The first strand explores *hybrid team interactions for multi-party decision-making* in simulated environments where agents are represented as active digital twins and humans participate either interactively or by modelling their (social) behaviour. The deployment in critical trustworthy real-world use cases is the ultimate goal for hybrid intelligent systems. Real life situations sometimes require a coordinated and combined use of different approaches, E.g. Shortcomings in the AI have to be solved by designing teamwork that allows the human to take over.

The second strand explores *ad hoc teamwork* within human-agent teams. A typical ad hoc teamwork scenario considers an agent that needs to cooperate with unknown teammates without being able to resort to any coordination mechanism. Hence, this agent has to learn how to collaborate on the fly and help the teammates to complete the team's task.

State-of-the-art algorithms for ad hoc teamwork can, in theory, be used to allow agents to collaborate with humans on-the-fly, without any pre-coordination protocol. However, they are not tailored for the specific challenges of human-agent collaboration. For instance, some works rely on the environment's reward signals, while others assume that an agent can observe the teammates' actions. Unfortunately, these assumptions may not hold in real-world human-agent interaction settings. Partners from IST have been proposing a novel contribution that addresses some of these limitations above[37]. For instance, our Bayesian Online Prediction for Ad hoc teamwork (BOPA) algorithm enables a robot to learn how to collaborate on the fly with human teammates by relying only on state observations. Our future work builds on this and takes partial observability into account.

2.4 Emergent Behaviour, agent societies and social networks (T6.4)

The main question one needs to consider as we move into a new kind of society where machines become more autonomous, involves the role that such machines will play in the dynamics of cooperation of the whole society, and if altruism can be altered or fostered as one of the emergent properties of such society. Or if, on the contrary, the presence of autonomous machines will make us humans, less humane, and cooperation will end up being less important and more rare. Is it possible that simple AI agents, or bots can have a teaching role contributing to a change in human's strategies when interacting with other humans?

The question is certainly an intricate one, and its answer may be conditioned by many different factors coming into place on the part of the machines (AI), such as the types of systems, how trustworthy they are, and what type of behaviour is the one leading to effective changes in a society. Furthermore, and given the large behavioural space that machines have now, the major question one should ask is what type of behaviour is needed from the machines to make a difference in such a future society to make it more cooperative and prosocial. This question

[37] Neves, A., & Sardinha, A. (2022). Learning to Cooperate with Completely Unknown Teammates. arXiv preprint arXiv:2205.03289.

can be addressed in different ways, allowing more theoretical to more experimental approaches.

As discussed previously in Section 2.2, delegation is an important process intrinsically related to trust. One agent (the trustor) delegates a task to another agent if it trusts it. So, to understand the conditions for this delegation to occur by humans, we have presented an experimental study of human delegation to autonomous agents and hybrid human-agent interactions centred on a public goods dilemma. We aimed to understand experimentally whether the presence of autonomous agents has a positive or negative impact on social behaviour, fairness and cooperation. Our results show that cooperation increases when participants delegate their actions to an artificial agent that plays on their behalf. Yet, this positive effect is reduced when humans interact in hybrid human-agent groups. Further experiments and theoretical approaches are being developed in close collaboration among several TAILOR partners (VUB & IST) [38] [39].

Furthermore, and in the context of cooperation studies, Social AI can also offer fundamental contributions to a better understanding of the principles behind the evolution and self-organisation of cooperation in nature and societies. For instance, it is known that distinct cooperation mechanisms have been proposed and identified in practice. Yet, they have been mostly studied independently. In a Rome/Lisbon partnership, we resorted to multiagent systems combined with game theory to study the conditions in which self-organised choices in large populations evolve towards the use of direct reciprocity, signalling, or their combination. We show that signalling alone leads to higher levels of cooperation than when combined with reciprocity while offering additional robustness against errors. Specifically, successful strategies in the realm of direct reciprocity are often not selected in the presence of signalling, and memory of past interactions is only exploited opportunistically in the case of earlier coordination failure. Differently, signalling always evolves, even when costly. In the light of these results, it may be easier to understand why direct reciprocity has been observed only in a limited number of cases among non-humans, whereas signalling is widespread at all levels of complexity. This suggests that the interaction between different cooperation-promoting mechanisms may be detrimental, suggesting that careful modelling must be performed in order to design agents that can display trustworthy Social AI[40] .

Signalling is a key element in many collective decision making problems, especially when a population must identify the most valuable solution to a problem among many distracting alternatives. In a collaboration between Rome and Brussels, a new evolutionary study is shedding light on the emergence of different signalling systems that can provide both positive and negative feedback mechanisms in a evolving population, with the goal of explaining the different possible strategies observed in nature when multiple non-exclusive alternatives are

[38] Fernández Domingos, E., Terrucha, I., Suchon, R., Grujić, J., Burguillo, J. C., Santos, F. C., & Lenaerts, T. (2022). Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports*, *12*(1), 1-12.

[39] Domingos, Elias Fernández, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C. Burguillo, Francisco C. Santos, and Tom Lenaerts. "Delegation to autonomous agents promotes cooperation in collective-risk dilemmas." *arXiv preprint arXiv:2103.07710* (2021).

[40] Martinez Vaquero, Luis A., Santos, Francisco C., Trianni, V (2020) Signalling boosts the evolution of cooperation in repeated group interactions. J. R. Soc. Interface 17:20200635

provided for a decision problem[41]. A similar consensus problem has been studied by structuring the decision process in a hierarchy of sequential decisions, showing that the hierarchical organisation can improve both accuracy and speed of the decision making process with respect to non-hierarchical approaches. Indeed, by serialising the decision process in a hierarchy, it is possible to save time in evaluating poor options, focusing the collective effort in the most valuable direction[42].

Signalling dynamics can also have a substantial impact whenever humans interact with social agents. Indeed, also here, decision-making can be shaped by complex non-verbal communication. Interactions comprising machines and agents can profoundly impact decision-making (both in actions directed at those machines and towards other humans). In particular, positive social behaviours (such as cooperation and prosocial behaviours) may be elicited through the interaction with socially interactive agents that can invoke positive emotions from their users. As many studies have proposed, these complex social behaviours can transcend the limited interaction domain and lead to actual cooperation and prosocial behaviours directed at other humans[43].

In this context, WP6 teams have contributed with several new results, but also a new survey on the impact of social agents portraying emotions and empathy in human cooperation[44]. Moreover, we also tried to assess the interplay of emotion expressions with distinct important cooperation mechanisms, including direct reciprocity, reputation-based systems and associated social norms. By doing so, we proposed a new class of emotion-based social norms, where emotions are used to forgive those that defect but also punish those that cooperate. These findings emphasise the importance of emotion expressions in fostering cooperation in society, while testifying the relevance of non-verbal communication in Social AI applications[45], but also suggests the need to address the role of cognitive complexity associated with such type of strategic decisions [46].

Despite these advances, humans face very complex dilemmas in which we (but also autonomous agents) face difficulties achieving cooperation. Humans and agents often converge to sub-optimal, risk-dominant solutions where everyone defects. In this line of research we investigated the consequences of risk diversity, wealth inequality, and uncertainty in collective goals and the time to solve them, among other vital issues. To this end, we resort to the tools of multiagent reinforcement learning, evolutionary game theory and behavioural experiments. As examples of key messages obtained, we have shown how uncertainty about

[41] Martinez Vaquero, Luis A., Reina, Andreagiovanni., Trianni, Vito (2020) On the evolutionary origins of signalling in colletive decision making. In preparation.
[42] Oddi, Fabio and Trianni, Vito (2022) Best-of-N collective decisions on a hierarchy. Proceedings of ANTS 2022 (submitted)
[43] Paiva, A., Santos, F., & Santos, F. (2018, April). Engineering pro-sociality with autonomous agents. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
[44] Oliveira, Raquel, Patrícia Arriaga, Fernando P. Santos, Samuel Mascarenhas, and Ana Paiva. "Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour." *Computers in Human Behavior* 114 (2021): 106547.
[45] de Melo, C. M., Terada, K., & Santos, F. C. (2021). Emotion expressions shape human social norms and reputations. *Iscience*, *24*(3), 102141.
[46] Santos, Fernando P., Jorge M. Pacheco, and Francisco C. Santos. "The complexity of human cooperation under indirect reciprocity." *Philosophical Transactions of the Royal Society B* 376, no. 1838 (2021): 20200291.

the time available to solve a collective risk problem prompts early generosity and polarised outcomes. We, therefore, established a link between uncertainty and the emergence of polarised behaviours. This connection has been proved both theoretically and experimentally with human experiments[47] [48]. Furthermore, we have also shown how risk diversity and wealth inequality significantly reduce overall cooperation and hinder collective target achievement and how these sources of diversity can be used to leverage cooperation in some particular scenarios [49] [50] .

From a higher-level perspective, Social AI approaches can also be useful in the governance of socio-technological systems. With the introduction of Artificial Intelligence (AI) and related technologies in our daily lives, fear and anxiety about their misuse and their inherent biases incorporated during their creation have led to a demand for governance and associated regulation. Yet regulating an innovation process that is not well understood may stifle this process and reduce benefits that society may gain from the generated technology, even with the best intentions. Brussels and Lisbon partners have examined this problem theoretically, resorting to a novel innovation dilemma. We identified the conditions under which innovation races may trigger detrimental consequences, and suggest potential regulatory approaches that combine soft law mechanisms with either a peer or governmental sanctioning systems. Overall, this line of research provides an original dynamic systems perspective of the governance potential of enforceable soft law techniques or co-regulatory mechanisms, showing how they may impact developers' ambitions in the context of concrete Social AI applications to governance and political science [51].

2.5 Applications and Impact (T6.5)

The aim to empower human users of AI systems becomes paramount when considering coordination in hybrid teams of humans and autonomous agents in which its members strive to make use of their complementary capabilities.

Agents empower humans with providing their complementary capabilities, such as fast and precise information exchange and analysis of large data sets. Humans maintain the overall

[47] Domingos, Elias Fernández, Jelena Grujić, Juan C. Burguillo, Georg Kirchsteiger, Francisco C. Santos, and Tom Lenaerts. "Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization." *Iscience* 23, no. 12 (2020): 101752.
[48] Domingos, E.F., Grujić, J., Burguillo, J.C., Santos, F.C. and Lenaerts, T., 2021. Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *Simulation Modelling Practice and Theory*, *109*, p.102299.

[49] Merhej, R., Santos, F. P., Melo, F. S., & Santos, F. C. (2021, May). Cooperation between independent reinforcement learners under wealth inequality and collective risks. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 898-906)
[50] Merhej, Ramona, Fernando P. Santos, Francisco S. Melo, Mohamed Chetouani, and Francisco C. Santos. "Cooperation and learning dynamics under risk diversity and financial incentives." In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 908-916. 2022.
[51] Santos, Francisco C., Luís Moniz Pereira, and Tom Lenaerts. "A Regulation Dilemma in Artificial Intelligence Development." In *ALIFE 2021: The 2021 Conference on Artificial Life*. MIT Press, 2021.

responsibility for decision making, although the agents can also be assigned some elements of responsibility, e.g 24/7 support in verifying, validating and approving proposals for decisions.

The impact of social AI systems on our society becomes clear when looking at how humans and machines can complement each other's strengths, for which the origin lies in the interaction between them. Hybrid team interactions for multi-party decision-making can be explored in simulated environments where agents are represented as active digital twins and humans participate either interactively or by modelling their (social) behaviour. However, critical real-life applications are the ultimate goal for purposeful hybrid settings in the real world.

Partners in TAILOR are involved in a number of applied research projects. In the domain of urban sustainability in particular, TNO is working on interactive smart buildings. The buildings interact with each other to optimise their energy consumption and reduce their CO2 footprint. They share global goals in managing energy peak load on the power grid. The setting involves multiple types of stakeholders, such as its residents, employees, building owners, construction engineers, energy companies and last but not least, policy makers. Learning takes place in a hybrid setting where the AI in the buildings form a hybrid team with humans in order to optimise the climate management systems of each building. Workshop papers on the HHAI2022 conference are being published.

The research activities of slovak.AI were focused on the automated detection of online fake news and overcoming the negative effect of misinformation filter bubbles in adaptive systems, mostly the social media and text based approaches[52] , [53], [54], [55]. We have explored the auditing methods (agent based) for recommender systems[56]. We have also explored embodied approach to conceptual knowledge and distinguish between embodiment and grounding[57].

Furthermore, TU Delft studies agent societies in several data-driven case studies, leading to published papers. In the CityAI lab they study the future liveability of cities around the world.

[52] Miroslav Blšták, and Viera Rozinajová. "Automatic question generation based on sentence structure analysis using machine learning approach". Natural Language Engineering, 1-31. DOI: 10.1017/S1351324921000139, 2021.
[53] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova."An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. Fifteenth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, 1–11. DOI: 10.1145/3460231.3474241, 2021.
[54] Jakub Simko, Patrik Racsko, Matus Tomlein, Martina Hanakova, Robert Moro & Maria Bielikova , A study of fake news reading and annotating in social media context, New Review of Hypermedia and Multimedia, 27:1-2, 97-127, DOI: 10.1080/13614568.2021.1889691, 2021.
[55] Michal Kompan, Peter Gaspar, Jakub Macina, Matus Cimerman and Maria Bielikova. , Exploring Customer Price Preference and Product Profit Role in Recommender Systems, in IEEE Intelligent Systems, doi: 10.1109/MIS.2021.3092768, 2021.
[56] Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova.
Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading. Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, 411–414. DOI: 10.1145/3450614.3463353, 2021.
[57] Reinboth, T. and Farkaš, I., , Ultimate Grounding of Abstract Concepts: A Graded Account. Journal of Cognition, 5(1), p.21. DOI: 10.5334/joc.214, 2022.

The studies include crumbling social cohesion, income inequality, overcrowding of public spaces, and unhealthy local environments caused by factors such as heavy traffic and noise pollution. The research concentrates on the pivotal role that the urban environment in combination with human behaviour plays in tackling such challenges. The research capitalises on advances in machine learning and on the wealth of data available now at a city level. The established theories on planning and behaviour, are combined to contribute to the development of more attractive and liveable cities.

We believe that in the next period, and in collaboration with WP8, further concrete applications will emerge.

Part 3- Overview of Activities

3.1 Setting up the scene: links between WP6 and other workpackages

In a network as large as this, the work performed in WP6 has to be seen in combination with the other workpackage, in particular in what concerns its link in terms of scientific challenges The aim behind the work in Trustworthy AI is to define methods, processes, algorithms to build artifacts that are able to autonomously act upon our world or make decisions that are considered not only smart but companies and humans trust them. WP 6 is dedicated to social aspects of such construction, but it interconnects heavily with all the other ones. As shown in Figure 3.1

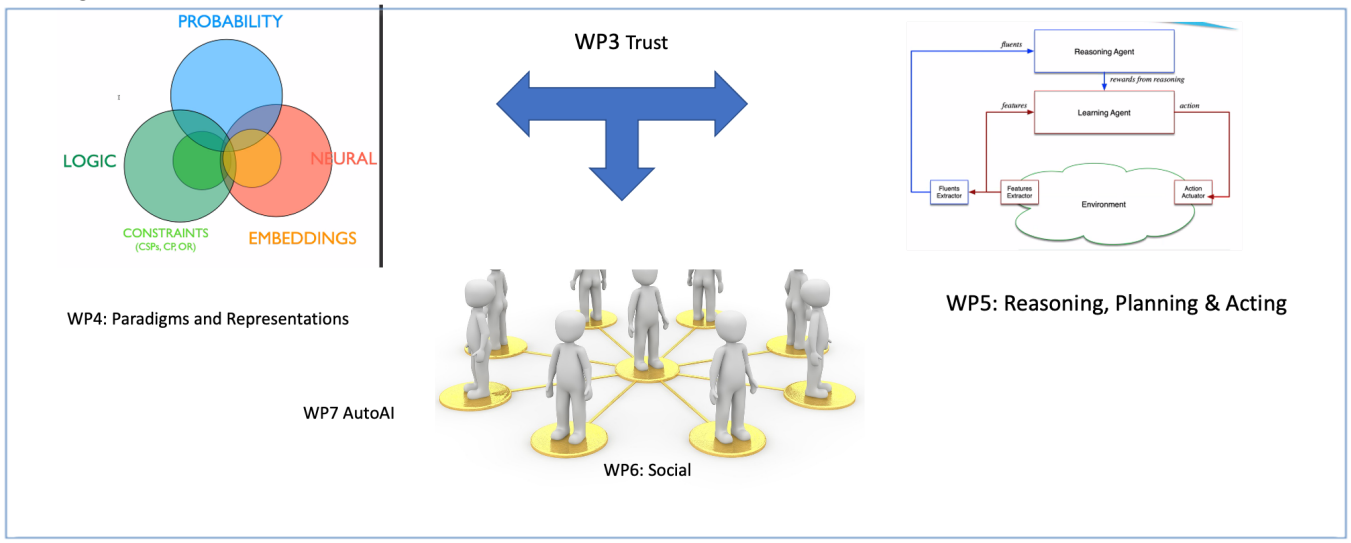


Figure 3.1- Interconnection between WP6 and the other workpackages.

Link between WP3 and WP6:

1. partners in WP6 are very much involved in WP3 as all the issues of trust, privacy, transparency and essential for models where humans are agents are present.
2. Link between WP4 and WP6: WP4 is a foundational WP that provides a space for researching paradigms and representations. In WP6 such paradigms and representations constitute the basis for the creation of models for social interactions between agents.
3. Link between WP4 and WP6: WP4 is about planning and acting, which is essential in social situations. Several partners are both in WP5 and 6, due to this strong link.
4. Link between WP6 and WP7: as autoAI requires in its development the interaction with humans, we expect that social AI contributions in WP6 will be of relevance for WP7.

3.2 Contribution to the TAILOR Objectives and KPIs

WP6 has contributed to the creation of the capacity and critical mass to develop the scientific foundations for Trustworthy AI (Obj 3), the progress the Scientific State-of-the-Art for the Foundations of Trustworthy AI, and in particular to the following KPIs:

#3.1# researchers in that have published papers acknowledging TAILOR
#3.2 # published papers with authors from at least two TAILOR partners
#3.4 Research visits of at least 5 days within the network
#3.3# published papers with authors from TAILOR and outside Europe
#3.5 Research visits of at least 5 days from outside the network
#4.1 Ranking and # of publications acknowledging TAILOR
#4.3 Number of publications / applications showing an increased Performance or new abilities of integrated learning, reasoning and optimisation approaches

As seen from the list of publications (see Section 5), WP6 has contributed to the number of published works acknowledging TAILOR in this area. Of relevance is the fact that several papers were published in very high impact conferences, namely AAMAS, IJCAI, AAAI. A short number of these include authors from different partners. In terms of visits, there was so far only two major exchanges (IST-VUB), but this reduced number is due to the limitations we have faced because of Covid. We expect this number to increase in the next year.

3.3 Links with other networks

Apart from the participation in the joint activities previously described in the Periodic Report, due to a close link between WP6 and HumaneAI-Net, members that are partners of both networks are in process or organising joint events, fostering the interrelationship between the two networks.

In particular WP6 and HumaneAI-Net have organised a joint seminar held on Wednesday, the 11th of May with the title “Trustworthy Human-AI Partnerships” with an invited talk by Prof. Sarvapali Ramchurn (Gopal) from the University of Southampton and coordinator of the UKRI Trustworthy Autonomous Systems hub.

3.4 Report on Meetings and Discussions

The community working on social aspects of AI is quite dynamic and so far we have been organised around a set of formal and informal events.

A Monthly meeting was set up since the start of the project, where members of the community would discuss issues related to the topics of the area. Between the more informal meetings, some other meetings were carried out with invited speakers. Furthermore,, we also hosted meetings with specific topics covering the 4 scientific tasks of the WP. These four discussions were coordinated by the Tasks leaders: Carles Sierra, Michael Woldridge, Ann Nowe and Vito Trianni.

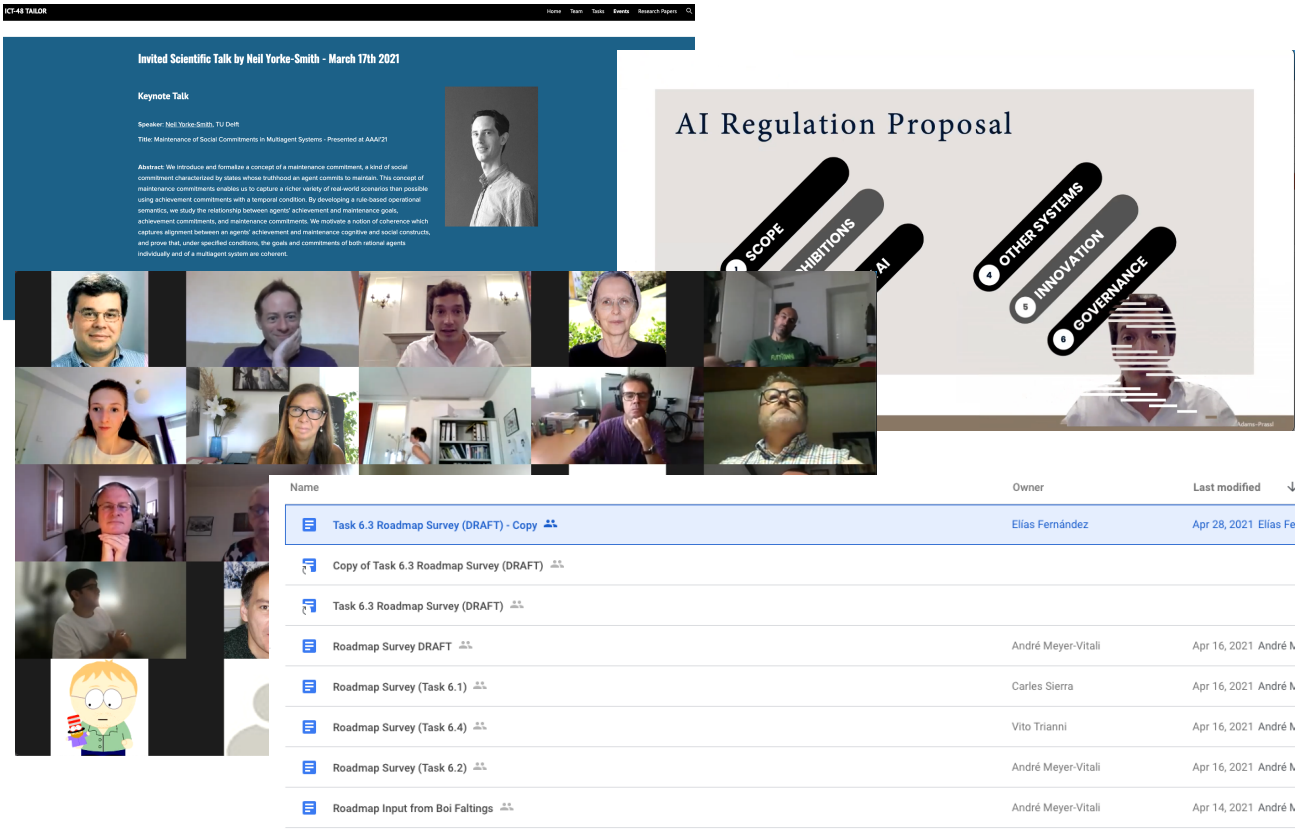


Figure 3.2 Screenshots of examples of WP6 meetings

Furthermore, we also had meetings featuring a wide variety of invited speakers, in particular: the Kick-off meeting (see Annex 1) where we discussed social intelligence and inspiraction from biology. In the kickoff meeting we had as invited speaker Prof. Josef Call, a well known primatologist that gave an inspirational talk about social sognals a collaboration in primates. We also had a meeting on the AI Regulation run by the University of Oxford (see Annex 3); and a joint meeting with the Humane-AI-Net (see Annex 2). Furthermore, we also had invited talks from both within the network (Prof. Sarit Kraus, and Prof Neil Yorke-Smith) and outside of the network (by Jeremy Pitt from the Imperial College in London and Giulia Andrighetto, CNR in Italy).

3.5 List of papers and collaborations from this WP

In Annex 4 we provide a list of papers published by the partners corresponding to the work that has been done over the past year and a half, reflecting some of the research here summarised.

4. Final Conclusions, Reflections and Plan for the next period

We believe that this area is fundamental for the development of trustworthy AI. The work here presented is still preliminary, and reflects the wide variety of challenges and difficulties. In fact, we have identified a set of obstacles for the further development of Trustworthy Social AI, which we will need to address in the near future:

1. Methodological: need for an interdisciplinary vision both in terms of experimental and theoretical approaches. It should combine AI, statistics, psychology, mathematics, population biology, etc.
2. Institutional: There's an enormous leap between AI and social sciences, both in terms of researchers' background, but also in terms of funding institutions, editorial policies, etc..
3. Complexity: Trustworthy AI is an emergent property resulting from many heterogeneous interacting components. Understanding these systems is a difficult task. On top of this, the design of self-regulatory mechanisms and technologies of these systems require a novel dynamical perspective on populations of trustors and trustees, on humans and machines, etc. More so, if embodied AI is considered.

In terms of work, in general, the core members of the WP are very engaged and see TAILOR as a good way to collaborate and discuss ideas around Social AI, aiming to make an impact in the area. The Monthly meetings have constituted a space for discussion and brought in new ideas. Events, such as the kick-off meeting and further invited talks were well attended (average 30 participants).

However, we need to acknowledge that it has been a difficult period for the network due to the remote form of working, increasingly tiredness of online events. Members have reported having difficulty in establishing new links and further cooperation, and PhD students in particular have been quite isolated.

We expect that in the next few Months, with the return to physical events, this lack of cooperation is overcome. PhD students will be able to participate in the Summer School as well as the TAILOR conference being organised in September.

Annex 1. Kick-off Meeting

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

WP6- Social AI: Learning and Reasoning in Social Contexts
Kickoff meeting 20/01/2021- 10am-1pm CET

Registration

https://videoconf-colibri.zoom.us/meeting/register/tZOuf-itqTIrH9wVT5LLEvZwa3CbJTVBxdhD

Schedule

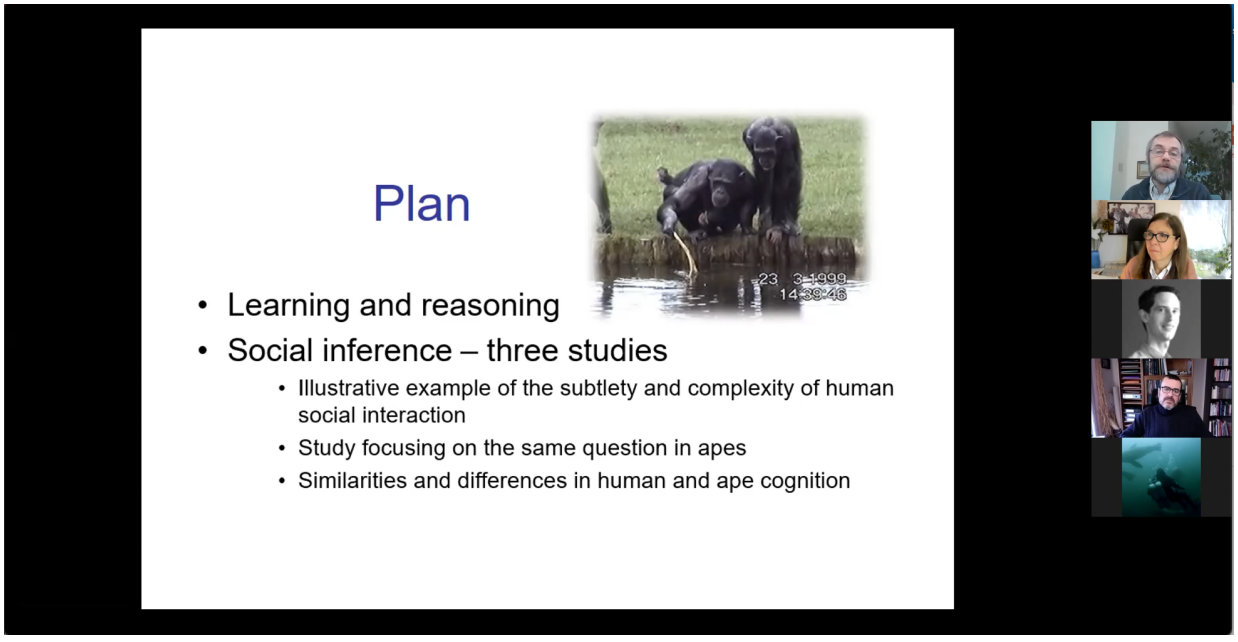
10:00-10:05 - Doors open
10:05-10:15 - Welcome, overview of TAILOR
10:15-10:45 - Keynote Talk- Joseph Call - Suggested Title: What can AI learn from Learning, Reasoning and Social Interactions in Non-human Primates?
10:45-10:50- WP6- The big questions
10:50-11:20 - Tasks

- 6.1 - Modelling social cognition, collaboration and teamwork. Presentation by Carles Sierra -
- 6.2: - Theoretical models for cooperation between agents. Presentation by Michael Woodridge

- 6.3: - Learning from others, Presentation by Ann Nowe
- 6.4 - Emergent Behaviour, agent societies and social networks, Presentation Vito Trianni
- 6.5: - Synergies Industry, Challenges, Roadmap on social AI system- Presentation by André Meyer-Vitali
- 6.6: - Fostering the AI scientific community - Presentation by Ana Paiva,

11:20 - 11:30 - Break
11:30 - 11:35 - Explanation of breakdown rooms- and questions to be addressed
11:35 - 12:15 - Focus groups on tasks 6.1 - 6.4 (what is the problem and how to measure progress?)
12:15 - 12:45 - Presentation and discussion of each focus group on tasks 6.1 - 6.4
12:45 - 13:00 - Open discussion on organization of WP activities (workshops, challenges, site, discussion groups, Tasks 6.5 and 6.6)

Participants: > 45
Keynote Talk: Keynote Talk- Joseph Call



Annex 2. Joint HumaneAI-Net & TAILOR event

TAILOR

Foundation of Trustworthy AI:
Integrating Learning, Optimisation and Reasoning

The TAILOR and HumaneAI-Net Networks will host a joint meeting on Wednesday, the 11th of May, at 10am CET.

Link:

https://videoconf-colibri.zoom.us/j/88275751066?pwd=b3laV1lnRURSN3hPYWUwWEtRdFI2UT09

Schedule

10:00-10:05 - Doors open & Welcome
10:05-10:45 - Invited Talk: Prof. Sarvapali Ramchurn (Gopal)
10:45-11:00- Discussion

Trustworthy Human-AI Partnerships

Abstract: Recent advances in AI, Machine learning and Robotics have significantly enhanced the capabilities of machines. Machine intelligence is now able to support human decision making, augment human capabilities, and, in

some cases, take over control from humans and act fully autonomously. Machines are becoming more tightly embedded into systems alongside humans, interacting and influencing each other in a number of ways. Such human-AI partnerships are a new form of socio-technical system in which the potential synergies between humans and machines are much more fully utilised. Designing, building, and deploying human-AI partnerships present a number of new challenges as we begin to understand their impact on our physical and mental well-being, our personal freedoms, and those of the wider society. In this talk I will focus on the challenges in designing trustworthy human-AI partnerships. I will detail the multiple elements of trust in human-AI partnerships and discuss the associated research challenges. I will also aim to identify the risks associated with human-AI partnerships and therefore determine the associated measures to mitigate these risks. I will conclude by giving a brief overview of the UKRI Trustworthy Autonomous Systems Programme (www.tas.ac.uk), a £33m programme launched in 2020 involving over 20 universities, 100+ industry partners, and over 200 researchers.

Bio: Prof. Sarvapali Ramchurn is a Professor of Artificial Intelligence, Turing Fellow, and Fellow of the Institution of Engineering and Technology. He is the Director of the UKRI Trustworthy Autonomous Systems hub (www.tas.ac.uk) and Co-Director of the Shell-Southampton Centre for Maritime Futures. He is also a Co-CEO of Empati Ltd, an AI startup working on decentralised green hydrogen technologies. His research is about the design of Responsible Artificial Intelligence for socio-technical applications including energy systems and disaster management. He has won multiple best paper awards for his research in multi-agent systems, energy management, and disaster response, and is a winner of the AXA Research Fund Award (2018) for his work on Responsible Artificial Intelligence.

Annex 3. Monthly Meetings: invited presentations

Scientific Talk by Neil Yorke-Smith
March 17th 2021
Speaker: Neil Yorke-Smith, TU Delft
Title: Maintenance of Social Commitments in Multiagent Systems - Presented at AAAI'21
Abstract: We introduce and formalize a concept of a maintenance commitment, a kind of social commitment characterized by states whose truthhood an agent commits to maintain. This concept of maintenance commitments enables us to capture a richer variety of real-world scenarios than possible using achievement commitments with a temporal condition. By developing a rule-based operational semantics, we study the relationship between agents' achievement and maintenance goals, achievement commitments, and maintenance commitments. We motivate a notion of coherence which captures alignment between an agents' achievement and maintenance cognitive and social constructs, and prove that, under specified conditions, the goals and commitments of both rational agents individually and of a multiagent system are coherent.
Bio: Neil Yorke-Smith is an Associate Professor of Socio-Technical Algorithmics in the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS/EWI), Delft University of Technology.

Scientific Talk about the AI-ACT: a legal perspective

Aislinn Kelly-Lyth studied Law at Cambridge and was subsequently awarded a Kennedy Scholarship to attend Harvard Law School. She is now a researcher at an Oxford-based project on algorithmic management, led by Professor Jeremias Adams-Prassl. She is interested in the overlap between technology and law, with a particular focus on data protection and equality rights. Her recent article *Challenging Biased Hiring Algorithms* examined the application of the law to potentially discriminatory automated recruitment practices.

Jeremias Adams-Prassl is Professor of Law at Magdalen College in the University of Oxford. He studied law at Oxford, Paris, and Harvard Law School, and is particularly interested in the future of work and innovation. Jeremias is the author of over 100 articles and books, including most recently *Humans as a Service: the Promise and Perils of Work in the Gig Economy* (OUP 2018) and *Great Debates in EU Law* (MacMillan 2021). His work has been recognised by numerous prizes for teaching, research, and public impact, including the Modern Law Review's Wedderburn Prize, a British Academy Rising Star Engagement Award, and the 2019 St Petersburg Prize. Since April 2021, he has led a five-year research project on Algorithms at Work, funded by the European Research Council and a 2020 Leverhulme Prize. Jeremias tweets at @JeremiasPrassl.

Scientific Talk about the Human-AI collaboration

Speaker: Prof. Sarit Kraus
Abstract: We consider environments where a set of human workers needs to handle a large set of tasks while interacting with human users. We present automated intelligent agents that will work together with the human operators in order to improve the overall performance of such systems and increase both operators' and users' satisfaction. The automated agents could support the operators: the machine learning-based agent follows the operator's work and makes recommendations, helping him interact proficiently with the users. The agents can also learn from the operators and eventually replace the operators in many of their tasks. We will discuss environments where customers seek help regarding tasks they need to perform. We suggest a solution in which customers' calls are attended by machine learning-based agents, and human operators intervene only in cases the virtual agent fails to supply the service. We formally analyze the multiple factors that affect the performance of the suggested system and suggest methods to improve these factors.

Scientific Talk about the Social Norms

Speaker: Giulia Andrighetto
Abstract: Social norms are a crucial part of the solution to our most pressing societal challenges, from mitigating climate change to reducing the spread of infectious diseases. Despite their relevance, how norms shape cooperation among strangers is still insufficiently understood. Influential theories suggest that the level of exogenous threats faced by different societies plays a key role in the strength of the norms that different cultures have evolved. Still

causal evidence of exogenous threats on norms has not been so far collected. Here we deal with this dual challenge using a 30-day collective-risk social dilemma experiment to observe and measure norm change in a controlled setting. We ask whether a looming but uncertain collective catastrophe changes the strength of the social norms of cooperation that may avert it. We find that social norms predict cooperation and causally affect behavior. We also provide the first empirical demonstration that higher risk spontaneously lead to stronger social norms and that, when the risk changes, stronger social norms are more resistant to erosion. Still, the foreseeable loosening of norms in low risk settings has important policy implications. Taken together, our results demonstrate the causal effect of social norms in promoting cooperation and their role in making behavior resilient in the face of exogenous change.

Bio: I am a senior researcher at the Institute of Cognitive Sciences and Technologies of the National Research Council of Italy in Rome, where I am the coordinator of the Laboratory of Agent Based Social Simulation (LABSS). I am also a senior researcher at Mälardalen University, Västerås, Sweden and at the Institute for Future Studies, Stockholm, Sweden. My research focuses on the emergence, enforcement, change and decay of social norms and their effects on cooperation and conflicts. My research topics include cooperation, altruism, honesty, as well as bad norms and misinformation. I use theoretical and computational models, combined with on-line and laboratory experiments, surveys and big data to answer these and related questions about social norms. Although primarily fundamental, my research aims also at informing policy.

Annex 4. Published Papers

	Title	Authors	Venue/Date	DOI/Link
1	*Urbanism and Geographic Crises: A Micro-Simulation Lens on Beirut*	Termos, A.; and Yorke-Smith, N.	*Urban Planning, 7(1), February 2022.*	
2	*Agent-Based Simulation of Short-Term Peer-to-Peer Rentals: Evidence from the Amsterdam Housing Market*	Overwater, A. and Yorke-Smith; N.	*Environment and Planning B: Urban Analytics and City Science, 49(1), January 2022.*	
3	*Agent-Based Simulation of West Asian Urban Dynamics: Impact of Refugees*	Termos, A.; Picascia, S.; and Yorke-Smith, N.	*Journal of Artificial Societies and Social Simulation, 24(1), January 2021.*	
4	*Ultimate Grounding of Abstract Concepts: A Graded Account.*	Reinboth, T. and Farkaš, I.	*Journal of Cognition, 5(1), p.21. 2022.*	DOI: 10.5334/joc.214
5	*A study of fake news reading and annotating in social media context*	Jakub Simko, Patrik Racsko, Matus Tomlein, Martina Hanakova, Robert Moro & Maria Bielikova	*New Review of Hypermedia and Multimedia, 27:1-2, 97-127, 2021.*	DOI: 10.1080/13614568.2021.1889691

6	*Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading*	Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova.	*Adjunct Proceedings of the 29th ACM Conference on User Modelling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, 411–414. 2021.*	DOI: 10.1145/3450614.3463353
7	*Automatic question generation based on sentence structure analysis using machine learning approach*	Miroslav Blšták, and Viera Rozinajová.	*Natural Language Engineering, 1-31. 2021.*	DOI: 10.1017/S1351324921000139
8	*An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes*	Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova.	*Fifteenth ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, 1–11., 2021.*	DOI: 10.1145/3460231.3474241
9	*Exploring Customer Price Preference and Product Profit Role in Recommender Systems*	Michal Kompan, Peter Gaspar, Jakub Macina, Matus Cimerman and Maria Bielikova.	*IEEE Intelligent Systems, 2021.*	doi: 10.1109/MIS.2021.3092768
10	*Multi-Agent Abstract Argumentation Frameworks With Incomplete Knowledge of Attacks*	Andreas Herzig, Antonio Yuste-Ginel.	*Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1922–1928, 2021. ijcai.org	

11	*On the Epistemic Logic of Incomplete Argumentation Frameworks*	Andreas Herzig, Antonio Yuste-Ginel.	*Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021*, pages 681–685, 2021.	
12	*Abstract Argumentation with Qualitative Uncertainty: An Analysis in Dynamic Logic*	Andreas Herzig, Antonio Yuste-Ginel	*Logic and Argumentation - 4th International Conference, CLAR 2021, Hangzhou, China, October 20-22, 2021, Proceedings*, volume 13040, of *Lecture Notes in Computer Science*, pages 190–208, 2021. Springer	
13	*Epistemic Reasoning About Rationality and Bids in Auctions*	Mittelmann, M.; Herzig, A.; and Perrussel, L.	*Logics in Artificial Intelligence - 17th European Conference, JELIA 2021, Virtual Event, May 17-20, 2021, Proceedings*, volume 12678, of *Lecture Notes in Computer Science*, pages 116–130, 2021. Springer	
14	*MARE: an Active Learning Approach for Requirements Classification*	Cláudia Magalhães, Alberto Sardinha, João Araújo.	*RE@Next! track of the 29th IEEE International Requirements Engineering Conference, September 2021.*	https://ieeexplore.ieee.org/document/9714537

15	*Helping People on the Fly: Ad Hoc Teamwork for Human-Robot Teams.*	Ribeiro, J.G., Faria, M., Sardinha, A., Melo, F.S.	*Progress in Artificial Intelligence. EPIA 2021. Lecture Notes in Computer Science (), vol 12981. Springer, Cham. 2021*	https://doi.org/10.1007/978-3-030-86230-5_50
16	*Ad Hoc Teamwork in the Presence of Non-stationary Teammates*	Santos, P.M., Ribeiro, J.G., Sardinha, A., Melo, F.S.	*Progress in Artificial Intelligence. EPIA 2021. Lecture Notes in Computer Science (), vol 12981. Springer, Cham. 2021*	https://doi.org/10.1007/978-3-030-86230-5_51
17	*A Methodology for the Development of RL-Based Adaptive Traffic Signal Controllers*	Guilherme Varela, Pedro Santos, Alberto Sardinha, Francisco Melo	*AAAI Workshop on AI for Urban Mobility (AI4UM), February 2021.*	http://aium2021.felk.cvut.cz/papers/AI4UM_paper_2.pdf
18	*Robust Solutions for Multi-Defender Stackelberg Security Games*	Mutzari Yonatan Aumann and Sarit Kraus	*IJCAI 2022.*	[pdf version]
19	*Toward Policy Explanations for Multi-Agent Reinforcement Learning*	Kayla Boggess, Sarit Kraus and Lu Feng	*IJCAI 2022.*	[pdf version]
20	*Justifying Social-Choice Mechanism Outcome for Improving Participant Satisfaction*	Sharadhi Alape Suryanarayana, D. Sarne, S. Kraus	*AAMAS 2022*	[pdf version]
21	*Online Learning-Based Assignment of Patients to Medical Professionals*	Hanan Rosemarin, Ariel Rosenfeld, Steven Lapp, Sarit Kraus	*Sensors, Special Issue "Human-Computer Interaction in Smart Environments", 2021*	[pdf version]
22	*Information Design in Affiliate Marketing.*	Suryanarayana, Sharadhi Alape, David Sarne, and Sarit Kraus	*Autonomous Agents and Multi-Agent Systems 35,(2):1-28, 2021.*	[pdf version]

23	*Coalition Formation in Multi-defender Security Games*	Dolev Mutzari, Jiarui Gan and Sarit Kraus	*AAAI 2021*	[pdf version]
24	*Manipulation of k-Coalitional Games on Social Networks*	Naftali Waxman, Noam Hazon, Sarit Kraus	*IJCAI 2021.*	[pdf version]
25	*Synthesising Reinforcement Learning Policies Through Set-Valued Inductive Rule Learning*	Coppens, Y., Steckelmacher, D., Jonker, C. M. & Nowe, A	*2021, Trustworthy AI - Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers. Heintz, F., Milano, M. & O'Sullivan, B. (eds.). 1 ed. Cham: Springer International Publishing, p. 163-179 17 p. (Lecture Notes in Computer Science; vol. 12641)*	
26	*Towards a Competence-Based Approach to Allocate Teams to Tasks.*	Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C.	*2021, Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 1504–1506.*	
27	*Building Contrastive Explanations for Multi-agent Team Formation*	Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C.	*Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 516–524. 2022*	

28	*An Anytime Heuristic Algorithm for Allocating Many Teams to Many Tasks*	Georgara, A., Rodríguez-Aguilar, J. A., Sierra, C., Mich, O., Kazhamiakin, R., Palmero-Approsio, A., & Pazzaglia, J.-C.	*Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 1598–1600. 2022*	
29	*Towards a Competence-Based Approach to Allocate Teams to Tasks.*	Georgara, A., Rodríguez-Aguilar, J. A., & Sierra, C. (2021).	*Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 1504–1506. Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). The Motivation to Work. John Wiley & Sons. 2021*	
30	*Enabling Game-Theoretical Analysis of Social Rules*	Montes, N., Osman, N., and Sierra, C.	*(Vol. 339, pp. 90–99). IOS Press.*	
31	*Combining Theory of Mind and Abduction for Cooperation under Imperfect Information*	Montes, N., Osman, N., and Sierra, C.	*(2022) [Under review]*	
32	*Maintenance of Social Commitments in Multiagent Systems.*	Telang, P. R.; Singh, M. P.; and Yorke-Smith, N.	*Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11369–11377, 2021. AAAI Press	

33	*A Lightweight Epistemic Logic and its Application to Planning.*	Cooper, M. C.; Herzig, A.; Maffre, F.; Maris, F.; Perrotin, E.; and Régnier, P.	*Artificial Intelligence*, 298: 103437. 2021.	
34	*A Distributed Differentially Private Algorithm for Resource Allocation in Unboundedly Large Settings.*	Danassis, P.; Triastcyn, A.; and Faltings, B.	*Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems*, of *AAMAS '22*, 2022.	
35	*Value-Guided Synthesis of Parametric Normative Systems.*	Montes, N.; and Sierra, C.	*AAMAS*, pages 907–915, 2021. ACM	
36	*Value-Alignment Equilibrium in Multiagent Systems*	Montes, N.; and Sierra, C.	*TAILOR*, volume 12641, of *Lecture Notes in Computer Science*, pages 189–204, 2020. Springer	
37	*Exploiting environmental signals to enable policy correlation in large-scale decentralized systems.*	Danassis, P.; Erden, Z. D.; and Faltings, B.	*Autonomous Agents and Multi-Agent Systems*, 36(1): 13. 2022.	
38	*Multi-Objective Reinforcement Learning for Designing Ethical Environments*	Rodriguez-Soto, M.; López-Sánchez, M.; and Rodríguez-Aguilar, J. A.	*IJCAI*, pages 545–551, 2021. ijcai.org	
39	*Signalling boosts the evolution of cooperation in repeated group interactions.*	Martinez-Vaquero, L. A.; Santos, F. C.; and Trianni, V.	*Journal of the Royal Society Interface*, 17(172): 20200635. 2021.	