Foundations of Trustworthy AI
Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Foundations, techniques, algorithms and tools to for social AI v.2

Document type (nature)	Report
Deliverable No	D6.2
Work package number(s)	W6
Date	30/6/2024
Responsible Beneficiary	IST
Author(s)	Francisco Melo
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	30/6/24	-	Francisco Melo

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Nic Wilson	#4, UCC	08/08/2024
Annelot Bosman	#7, ULEI	27/08/2024

Table of Contents

Summary of the report

This document details the recent efforts of the TAILOR partners involved in WP6, aiming to build trustworthy social AI by integrating reasoning, learning, and optimization mechanisms into interactions involving agents.

The social and organisational aspects of AI have become significant research areas in our field, encompassing multi-agent systems, human-agent interaction, network systems, game theory, social learning, and more. Social AI focuses on techniques to build AI that is emergent, situated, and capable of performing in social contexts. It ultimately aims to support the creation of hybrid populations that include both AI systems and humans. However, as AI is developed to act in social contexts, be distributed, and integrate within hybrid populations of humans and machines, several challenges arise:

1. How can agents communicate, negotiate and reach agreements in a trustworthy manner?
2. How can agents take into account others (including humans) and establish trustworthy relationships among them?
3. How do networks of agents and humans evolve, and does trust play a role and evolve as well?
4. How are teams created and maintained in hybrid populations of humans and agents?
5. Can we trust systems where AI is distributed?
6.And what foundational methods do we have to guarantee trust in them?

These challenges have been the focus of the collaboration in WP6 of the TAILOR network. During the TAILOR project we have been addressing these issues by combining community efforts to increase knowledge and expertise, promoting and developing “trustworthy social AI.”

This deliverable, the final report from WP6, builds upon the foundations laid out in D6.1, the first WP6 status report of the TAILOR network. It enriches and discusses techniques, algorithms, and tools to build and evaluate trustworthy social AI. Additionally, we report on the networking activities undertaken to achieve these outcomes.

Introduction to the Deliverable

This document represents deliverable D6.2 in TAILOR WP6, titled "Social AI: Learning and Reasoning in Social Contexts." It enriches the overview, stated in D6.1, of the work carried out in WP6. The document is a collaborative effort involving various organisations and researchers across Europe associated with the TAILOR project.

The document is organised as follows:

1. Major Concepts and Definitions Revisited: We begin with a review to the key concepts 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.

2. Challenges and Scientific Tasks: We discuss the challenges related to the six scientific tasks within WP6. This section includes concrete examples of work done by partners contributing to the project, highlighting specific results and advancements.

3. Collaboration and Activities: We provide a brief overview of our collaborative efforts, detailing the various activities organised over the last two years to foster cooperation and knowledge sharing among partners.

4. Future Research Agenda: We outline a future research agenda, paving the way towards a roadmap for trustworthy AI in future endeavours. This section discusses how the community can work together to achieve our collective goals.

5. Deliverables: We also provide a list of recent publications, visiting and presentations made by our partners, showcasing the ongoing research and contributions to the field.

This document serves as an overview of the progress and direction of WP6 within the TAILOR project, illustrating our commitment to advancing trustworthy social AI through collaborative research and development.

Organisation

The following people have been involved in the Deliverable:

Partner	Name	Role
IST-UL	Francisco Melo	W6 lead, and T6.6 lead
IIIA-CSIC	Carles Sierra	T6.1 Lead
UOX	Michael Wooldridge	T6.2 Lead
VUB	Ann Nowe	T6.3. Lead
CNR	Vito Trianni	T6.4. Lead
TNO	Wico Mulder	T6.5 Lead
IST-UL	Isabel Neto	W6 supporting role

1. Major Concepts and Definitions Revisited

In 1994, Prof. Barbara Grosz, proposed a new vision for AI as collaborative[1], at a time when the field was experiencing the second AI winter, characterised by reduced funding, low credibility, and widespread scepticism about AI's potential value. Despite these challenges, Grosz and her colleagues remained confident in AI's potential to make a significant impact. Her inspiring work highlighted the necessity for AI to be situated in its environment, dynamically use data, interact with humans, and most importantly, make decisions collaboratively.

A few years later, the field of multi-agent systems began to flourish. The first edition of the landmark book on multi-agent systems by M. Wooldridge[2] laid out a roadmap for the field, presenting two primary visions: (1) agents as a paradigm for software engineering, and (2) agents as a tool for understanding human societies.

Nearly thirty years after Grosz's vision, AI has become increasingly integrated into daily life, appearing in factories, roads, homes, hospitals, and even schools. Given these new contexts, A. Paiva[3] also recently emphasised that AI-powered machines must now be designed to place humans at the centre and interact with them naturally, marking the rise of social AI.

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

[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).

Now is the time to realise the vision anticipated by Grosz and reiterated by Wooldridge and Paiva: AI situated in social contexts, where agents are AI entities that cooperate and communicate within hybrid populations of humans and machines. The diverse applications of AI are driving significant changes, particularly in our behaviour and interactions with each other and machines. Therefore, it is crucial to reflect on AI's impact on human societies, considering its potential to support increased collaboration, social action, and prosocial (altruistic) behaviour. Integrating machines into social settings necessitates a deeper understanding of their effects on social interactions and their potential to influence human behaviour.

In the last 4 years, The TAILOR network, particularly this work package, brought together scientists dedicated to exploring AI techniques to better understand social interactions in natural and societal contexts. It also focuses on creating AI agents that prioritise social behaviour, fostering cooperation and collective action within human settings.

Based on this goal, we revisit some foundational definitions in this section.

1.1. Social AI definition

Social AI focuses on techniques to build AI agents that model, emerge, are situated, and can perform in social contexts, acting within populations that include both agents and humans.

This entails different dimensions of study:

- **D1. AI for Understanding Social Interactions**: This dimension explores AI's role in understanding social interactions, cooperation, coordination, organisations, and norms.
- **D2. AI with Social Competencies:** This focuses on developing agents/AI that are able to interact in a social manner, including social perception, understanding, and group dynamics.
- **D3. AI for Strategic Decision-Making:** This involves modelling strategic decision-making through game theoretical approaches, both non-cooperative and cooperative.
- **D4. AI Capturing Social Dynamics**: This examines AI's ability to capture the dynamics of social interactions in large simulated and hybrid societies.
- **D5. AI Performing in Social Contexts**: This dimension looks at how AI performs in social contexts and impacts the social environment we live in.

1.2. Social AI applications

Social AI systems are being deployed across various domains, here are some examples:

- **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. See for example the work of Hudson et al. (2023)[4].

- **Agriculture**: In precision agriculture and dairy farming, AI 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. See for example the works of Neethirajan (2024) [5].
- **Transportation**: The traffic and transport sector uses AI-based dialogue mechanisms in traffic management systems. See for example the work of Kuberkar and Singhal (2020) [6].
- **Energy**: AI helps citizens optimise energy consumption and manage resources like car sharing. Future applications include buildings sharing information to collaborate on energy management, contributing to smart city initiatives and efficient building occupancy management. These interconnected socio-technical systems are already being explored through agent-based simulations for urban planning and policy making. See for example the works of González-Méndez et al. (2021) [7].
- **Law Enforcement**: AI is used for federated reasoning to understand debt problems or solve cold cases.
- **Modelling and Simulation**: Social AI models that simulates human or agent behaviour to derive rules and patterns for better understanding. Applications include multi-robot task allocation in search and rescue, traffic management in smart cities, and advanced planning in digital manufacturing. See for example the works of Rocha et al. (2020) [8]
- **Media**: AI impacts the TV entertainment sector by classifying and personalising user interactions based on preferences. Content production becomes a collaborative process between AI systems and humans. See for example the works of Heim, S., & Chan-Olmsted [9].

These applications of social AI across various sectors illustrate its transformative potential. Our scientific tasks T6.5 aimed to further explore and harness these potentials by investigating the synergies between industry challenges and the roadmap for social AI systems (see deliverable D6.4 for detailed insights).

---

[4] Hudson, S., Nishat, F., Stinson, J., Litwin, S., Zeller, F., Wiles, B., ... & Ali, S. (2023). Perspectives of healthcare providers to inform the design of an AI-enhanced social robot in the pediatric emergency department. *Children*, *10*(9), 1511.

[5] Neethirajan, S. (2024). Artificial intelligence and sensor innovations: enhancing livestock welfare with a human-centric approach. *Human-Centric Intelligent Systems*, *4*(1), 77-92.

[6] Kuberkar, S., & Singhal, T. K. (2020). Factors influencing adoption intention of AI powered chatbot for public transport services within a smart city. *International Journal of Emerging Technologies in Learning*, *11*(3), 948-958.

[7] González-Méndez, M., Olaya, C., Fasolino, I., Grimaldi, M., & Obregón, N. (2021). Agent-based modeling for urban development planning based on human needs. Conceptual basis and model formulation. *Land Use Policy*, *101*, 105110.

[8] Rocha Filho, G. P., Meneguette, R. I., Neto, J. R. T., Valejo, A., Weigang, L., Ueyama, J., ... & Villas, L. A. (2020). Enhancing intelligence in traffic management systems to aid in vehicle traffic congestion problems in smart cities. *Ad Hoc Networks*, *107*, 102265.

[9] Heim, S., & Chan-Olmsted, S. (2023). Consumer Trust in AI–Human News Collaborative Continuum: Preferences and Influencing Factors by News Production Phases. *Journalism and Media*, *4*(3), 946-965.

1.3 How to achieve trustworthy Social AI

One of the main outcomes expected from TAILOR is to establish the scientific foundation for Trustworthy AI. While the economic potential of AI and algorithms is immense, it will only be successful if it meets the societal requirements for safety, security, and ethics. The goal is to create systems that are explainable, fair, safe, accountable, private, and sustainable—dimensions that are central to Trustworthy AI and aligned with TAILOR's foundation themes.

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

Trust is a complex, multidimensional concept that encompasses more than just competence but it also captures different phenomena[10][11]. In social contexts, trust describes how humans form relationships, attribute characteristics, and set expectations for others or entities.

Trust emerges from our beliefs about others, the environment, and the objects within it. Trust is inherently associated with risk; without the risk of failure, trust is unnecessary. The trustor must be vulnerable to the trustee's actions for trust to be relevant.

In general, trust can be defined by the assured reliance on the character ability, strength, or truth of someone or something[12]. However, when considering multiple agents with varying competencies and human-system relationships, new challenges arise.

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).

To guarantee trust in these contexts, we can bridge the gap from formal methods, verification, and validation to the engineering, usage, and reinforcement of multi-agent systems. This approach is addressed in our tasks (T6.1-T6.4). Another way is to view AI as a collaborative partner with social competencies, capable of explaining its decision-making process to humans (T 6.1). Additionally, we can implement automated mechanisms to study and understand the behaviour of mixed populations of artificial and human agents, starting from learned subsymbolic representations of behaviour (policy), finding symbolic categorizations, to allow for reasoning, communication, explanation and verification.

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.

[10] 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).
[11] Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In *Trust and deception in virtual societies* (pp. 55-90). Springer, Dordrecht.
[12] https://www.merriam-webster.com/dictionary/trust

We based our work on 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).

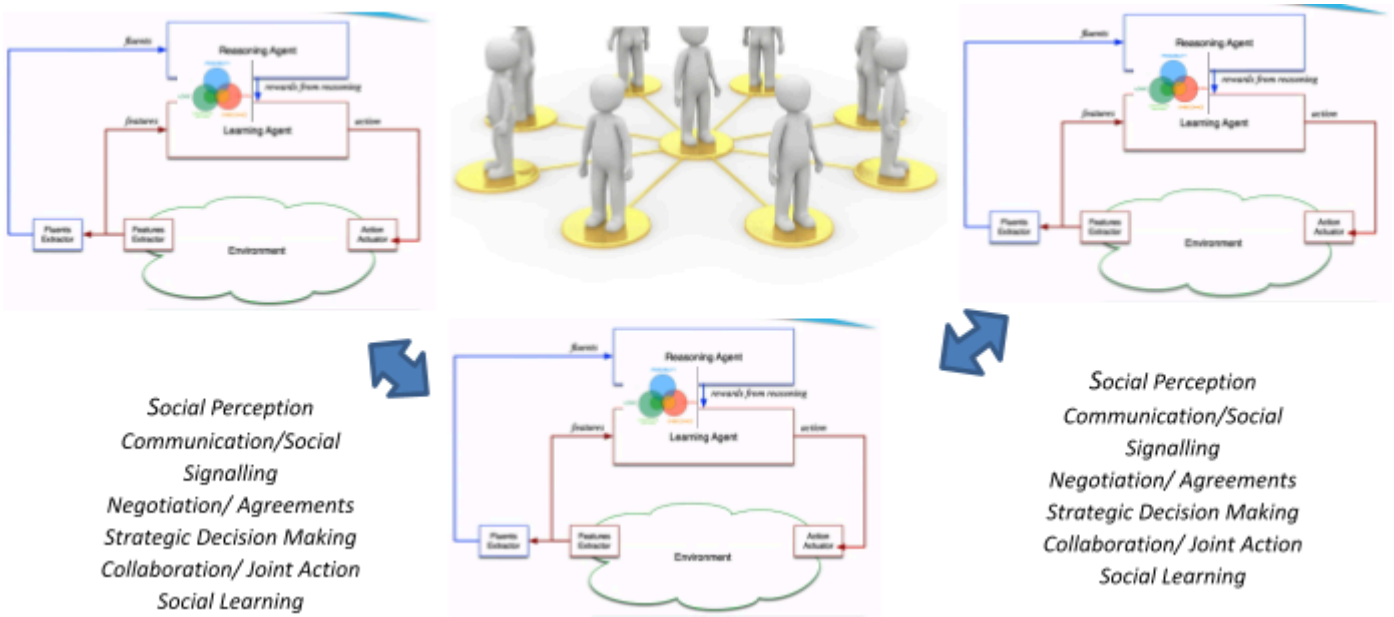

Figure 2.1 – Overview of generic framework for Social AI followed by work package 6. Where: each agent is captured as an entity that is able to perceive the world and act upon it, and its decision making can be a result of different techniques and algorithms. Each agent must perceive others, communicate, negotiate and make strategic decisions.

This work was carried out within 6 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
- Task 6.5 - Synergies Industry, Challenges, Roadmap on social AI system
- Task 6.6 - Fostering the AI scientific community on the theme of social AI

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[13], 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,”[14] . 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[15] , “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”[16] .

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

[13] 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.
[14] 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.
[15] Grosz, B. J. (1996). Collaborative systems (AAAI-94 presidential address). *AI magazine*, *17*(2), 67-67.
[16] 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.

may be very competent in knowing who inhabits each room of the building, and able to move 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).

Some recent work of ours' has tackled these questions of collaboration from three perspectives [17] [18] [19] [20] [21] [22] [23]. 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, 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),[24] a logical language implemented in Prolog that allows us 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

[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] Montes, N., Osman, N., and Sierra, C. (2021). *Enabling Game-Theoretical Analysis of Social Rules* (Vol. 339, pp. 90–99). IOS Press.
[22] Montes, N., Osman, N., and Sierra, C. (2022). *Combining Theory of Mind and Abduction for Cooperation under Imperfect Information European Conference on Multi-Agent Systems, 2022*
[23] Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.
[24] https://www.ai4europe.eu/research/ai-catalog/ngames

game-theoretical solution concepts. This way, a community of agents can draft new rules, 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[25] (Montes, 2021) presents some preliminary results.

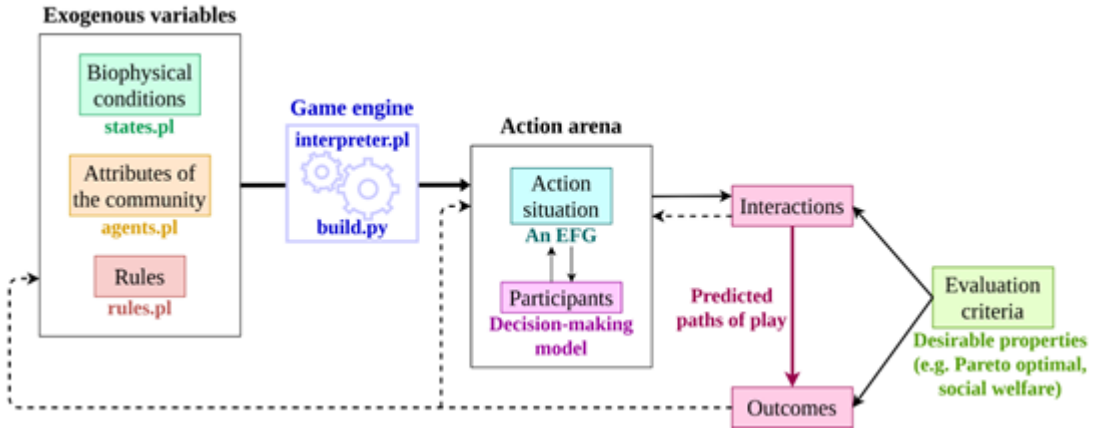

Figure 2.2: Outline of the IAD framework. Adapted from Ostrom 2005.

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[26] . This agent model is based on the combination of Theory of Mind (ToM) and abductive reasoning[27] 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[28] 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 abducted 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)

[25] Montes, N., Osman, N., and Sierra, C. (2021). *Enabling Game-Theoretical Analysis of Social Rules* (Vol. 339, pp. 90–99). IOS Press.
[26] Montes, N., Osman, N., and Sierra, C. (2022). *Combining Theory of Mind and Abduction for Cooperation under Imperfect Information European Conference on Multi-Agent Systems, 2022*
[27] Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, Carles Sierra: Combining theory of mind and abductive reasoning in agent-oriented programming. Auton. Agents Multi Agent Syst. 37(2): 36 (2023)
[28] 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.

architecture. We have tested our implementation in the Hanabi game domain, a cooperative card game that has recently attracted a lot of attention from the AI community[29] , 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; there is need for more than one such team 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 also been 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[30], or grouping students to undertake school projects[31] . These problems have in common that they involve the 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) the team’s collectively acquired competencies / skills / knowledge with respect to the task to be

[29] 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.
[30] 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.
[31] 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.

solved; (ii) the balance of team members' in terms of personality[32] ; (iii) the team's interest (collectively) towards the task to be solved[33] ; and (iv) team's social cohesion[34] . 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[35] [36]) 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.

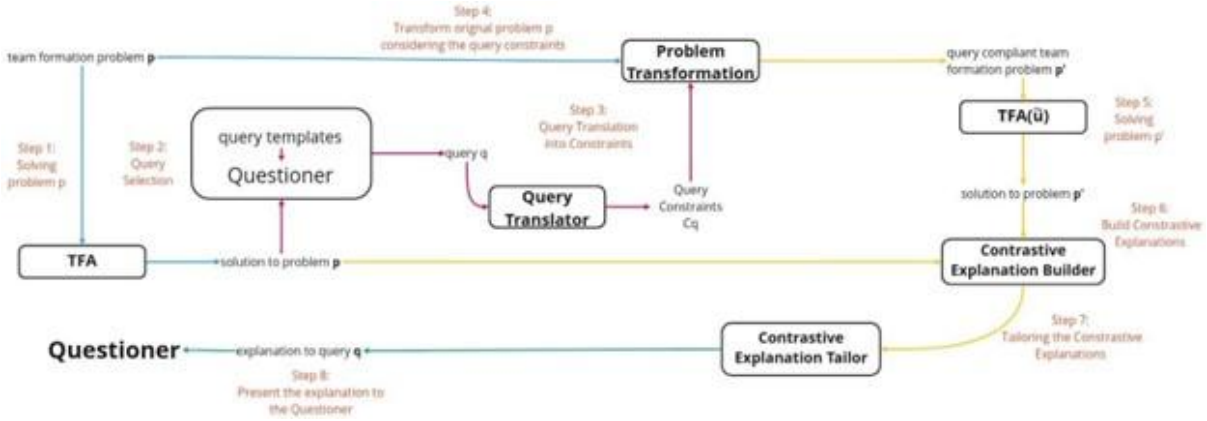

Figure 2.3: General Justification Algorithm for Team Formation.

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.2.3) that wraps existing team formation algorithms and builds contrastive explanations. Such explanations answer the "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.

[32] Belbin, R. (1993). *Team Roles at Work: A Strategy for Human Resource Management*. Butterworth-Heinemann.
[33] Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The Motivation to Work*. John Wiley & Sons.
[34] Randall, L. H., & Kuhnert, K. W. (1993). Using Sociometry to Predict Team Performance in the Work Place. *The Journal of Psychology*, *131*, 21-32.

[35] 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.
[36] 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[37]. 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 is assumed to have a goal (γi), represented as a propositional formula γ_i over some set Φ of Boolean variables intuitively representing 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[38] 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

[37] Kraus, Sarit, and Michael J. Wooldridge. "Delegating Decisions in Strategic Settings." In *ECAI*, vol. 12, pp. 468-473. 2012
[38] 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

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.

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[39], 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[40]. There are two approaches for modelling cooperation in multi-agent problems: non-cooperative setting and cooperative settings. For the non-cooperative settings they proved 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[41], 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.

[39] Dolev Mutzari, Jiarui Gan and Sarit Kraus. Coalition Formation in Multi-defender Security Games, AAAI 2021
[40] Mutzari Yonatan Aumann and Sarit Kraus Robust Solutions for Multi-Defender Stackelberg Security Games, IJCAI 2022.
[41] 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.[42][43] 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 (MARL), Boggess et al.[44] 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 behaviour. 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, VUB and IST jointly create 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[45] and we expect to add these tools to the AI4EU depository.

2.3. Learning in Social Contexts (T6.3)

From over-exploitation of resources to urban pollution, sustaining well-being requires solving social dilemmas of cooperation. In this WP, we studied social dilemmas in populations, in a variety of settings and under different assumptions.

In the first setting, we studied populations of self-interested agents playing a 2-person repeated Prisoner’s Dilemma game, with each player having the option of opting out of the interaction and choosing to be randomly assigned to another partner instead. The partner

[42] Suryanarayana, Sharadhi Alape, David Sarne, and Sarit Kraus. Information Design in Affiliate Marketing. Autonomous Agents and Multi-Agent Systems 35,(2):1-28, 2021.
[43] Sharadhi Alape Suryanarayana, D. Sarne, S. Kraus. Justifying Social-Choice Mechanism Outcome for Improving Participant Satisfaction AAMAS 2022.

[44] Kayla Boggess, Sarit Kraus and Lu Feng. Toward Policy Explanations for Multi-Agent Reinforcement Learning, IJCAI 2022.
[45] A Abels, EF Domingos, A Nowé, T Lenaerts, Mitigating Biases in Collective Decision-Making: Enhancing Performance in the Face of Fake News, arXiv preprint arXiv:2403.08829

selection component makes these games akin to random matching, where defection is known to take over the entire population. Results in the literature have shown that, when forcing agents to obey a set partner selection rule known as Out-for-Tat, where defectors are systematically being broken ties with, cooperation can be sustained in the long run. In this work, we remove this assumption and study agents that learn both action- and partner-selection strategies. Through multi-agent reinforcement learning, we show that cooperation can be sustained without forcing agents to play predetermined strategies. Our simulations show that agents are capable of learning in-game strategies by themselves, such as Tit-for-Tat. What is more, they are also able to simultaneously discover cooperation-sustaining partner selection rules, notably Out-for-Tat, as well as other new rules that make cooperation prevail.

Starting from a baseline model that has demonstrated the potential of rewiring for cooperation, we provide answers to this question over the full spectrum of social dilemmas. Multi-agent Q-learning with Boltzmann exploration is used to learn when to sever or maintain an association. In both the Prisoner's Dilemma and the Stag Hunt games, we observe that the Out-for-Tat rewiring rule, breaking ties with other agents choosing socially undesirable actions, becomes dominant, confirming at the same time that cooperation flourishes when rewiring is fast enough relative to imitation. Nonetheless, in the transitory region before full cooperation, a Stay strategy, keeping a connection at all costs, remains present, which shows that loyalty needs to be overcome for full cooperation to emerge.

In conclusion, individuals learn cooperation-promoting rewiring rules but need to overcome a kind of loyalty to achieve full cooperation in the full spectrum of social dilemmas.

In a second setting, we focus on mobile agents. Here we investigate how mobility costs impact cooperation dynamics. To this end, we study cooperation dilemmas where individuals are located in a two-dimensional space and can be of two types: *cooperators or cleaners,* who pay an individual cost to have a positive impact on their *neighbours and defectors or polluters*, free-riding on others' effort to sustain a clean environment. Importantly, agents can pay a cost to move to a cleaner site. Both analytically and through agent-based simulations we find that, in general, introducing mobility costs increases pollution felt in the limit of fast movement (equivalently slow strategy revision). The effect on cooperation of increasing mobility costs is non-monotonic when mobility co-occurs with strategy revision. In such scenarios, low (yet non-zero) mobility costs minimise cooperation in low density environments; whereas high costs can promote cooperation even when a minority of agents initially defect. Finally, we find that heterogeneity in mobility cost affects the final distribution of strategies, leading to differences in who supports the burden of having a clean environment.

Thirdly, we studied the setting of a campaigner who wants to learn the structure of a social network by observing the underlying diffusion process and intervening on it. Using synchronous majoritarian updates on binary opinions as the underlying dynamics, we offer upper bounds on the campaigner's budget for learning any network with certainty, considering both observation and intervention resources, and further improving them for the case of clique networks. Additionally, we investigate the learning progress of the campaigner when her budget falls below these upper bounds. For such cases, we design a greedy

campaigning strategy aimed at optimising the campaigner's information gain at each opinion diffusion step.

Finally, individual and social biases undermine the effectiveness of human advisers by inducing judgement errors which can disadvantage protected groups. The influence these biases can have in the pervasive problem of fake news by evaluating human participants' capacity to identify false headlines. By focusing on headlines involving sensitive characteristics, we gather a comprehensive dataset to explore how human responses are shaped by their biases. Our analysis reveals recurring individual biases and their permeation into collective decisions. We show that demographic factors, headline categories, and the manner in which information is presented significantly influence errors in human judgement. We then use our collected data as a benchmark problem on which we evaluate the efficacy of adaptive aggregation algorithms. In addition to their improved accuracy, our results highlight the interactions between the emergence of collective intelligence and the mitigation of participant biases.

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

There are very different ways in which AI can be integrated into social systems. The studies conducted in this respect span from mostly theoretical exercises trying to determine in which way it is possible to harness the power of hybrid human-AI collective intelligence, to very applicative studies in which concrete application domains have been used as the playground for self-organising AI systems, as well as for interacting groups of humans and machines.

Theoretical studies have accounted for improving collective decision making processes by addressing the limitations and biases in expert knowledge. Abels et al. address the issue of expert bias in collective decision making using a contextual multi-armed bandit (CMAB) framework.[46] This algorithm identifies and mitigates biased expertise, particularly in groups with homogeneous, heterogeneous, and polarised expert opinions. The CMAB-inspired approach not only counters bias effectively, but also converges more rapidly and achieves higher performance compared to existing methods. In a follow up study,[47] a novel algorithm based on expertise trees models varying depths and breadths of expertise among decision-makers by partitioning problem spaces into regions of differing expertise. This approach is shown to outperform traditional nearest neighbour queries by allowing the selection of more appropriate models based on the problem instance, thus enhancing decision accuracy. These studies contribute to the field by providing robust algorithms that enhance the reliability and efficiency of decisions made through collective expert judgments.

Self-organisation in multi-agent systems and robotics societies requires studies to explore advanced decision making and task allocation methods. Oddi et al. (2022)[48] tackle the

[46] Abels, A., Lenaerts, T., Trianni, V. & Nowé, A. Dealing with expert bias in collective decision-making. Artif. Intell. 320, 103921 (2023).
[47] Abels, A., Lenaerts, T., Trianni, V. & Nowé, A. Expertise Trees Resolve Knowledge Limitations in Collective Decision-Making. in Proceedings of the 40th International Conference on Machine Learning vol. 202 79–90 (2023).
[48] Oddi, F., Cristofaro, A., Trianni, V. (2022). Best-of-N Collective Decisions on a Hierarchy. In: Dorigo, M., et al. Swarm Intelligence. ANTS 2022. Lecture Notes in Computer Science, vol 13491. Springer, Cham.

best-of-N problem in collective decision making by transforming it into a hierarchy of simpler decisions, using an m-ary tree structure to improve speed and accuracy through multi-agent simulations. They also propose adaptive parameter tuning for better performance. In a related study, Oddi et al. (2024)[49] introduce minimalist protocols for quorum sensing in robot swarms, evaluating their efficiency and accuracy in varying swarm densities and environments, providing insights into the trade-offs between computational demands and performance. Albani et al. (2021)[50] address the complexities of task assignment and pathfinding in robot swarms under limited communication. They propose a decentralised approach that combines bio-inspired collective decision making for task allocation and search-based path planning, demonstrating its effectiveness and robustness in various complex environments. Miletitch et al. (2022)[51] investigate the emergence of naming conventions within foraging robot swarms, finding that useful linguistic conventions require a correlation between interaction networks and foraging dynamics, leading to a decentralised algorithm for effective collective categorization. Together, these studies contribute to enhancing the autonomy, efficiency, and flexibility of robot swarms in complex environments.

Various methodological aspects need to be addressed to apply AI and agent-based techniques to complex societal problems in diverse domains. Bbeanu et al. (2023)[52] introduce a protocol for adaptive parallelization of multi-agent simulations, enhancing computational efficiency by leveraging localised dynamics and shared-memory parallel execution. Dyer et al. (2023)[53] propose a framework for learning interventionally consistent surrogate models, using causal abstractions to ensure that surrogates reliably replicate agent-based simulator behaviours under policy interventions, thus facilitating rapid experimentation. Chopra et al. (2024)[54] address privacy concerns in ABMs by employing secure multi-party computation techniques, allowing decentralised computation without centralising sensitive data. Another study by Dyer et al. (2024)[55] presents a scenario generation framework for synthesising populations in ABMs, crucial for planning under uncertainty by generating synthetic populations that match specified target scenarios. Finally, Mulder and Meyer-Vitali (2023) propose a maturity model for collaborative agents in human-AI ecosystems, balancing autonomy and collaboration levels to enhance teamwork efficiency, validated through urban energy efficiency use-cases. These methodological innovations collectively enhance the scalability, privacy, reliability, and practical utility of

[49] Oddi F., Reina A., Trianni V., Minimalist Protocols for Quorum Sensing in Robot Swarms, In: Dorigo, M., et al. Swarm Intelligence. ANTS 2024. To appear.
[50] Albani, D.; Hönig, W.; Nardi, D.; Ayanian, N.; Trianni, V. Hierarchical Task Assignment and Path Finding with Limited Communication for Robot Swarms. Appl. Sci. 2021, 11, 3115.
[51] Miletitch, R., Reina, A., Dorigo, M. et al. Emergent naming conventions in a foraging robot swarm. Swarm Intell 16, 211–232 (2022)
[52] Bbeanu, A-I.; Filatova, T.; Kwakkel, J. K.; and Yorke-Smith, N. Adaptive Parallelization of Multi-Agent Simulations with Localized Dynamics. arXiv preprint 2304.01724. Apr. 2023.
[53] Dyer, Joel, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. "Interventionally Consistent Surrogates for Agent-based Simulators." arXiv preprint arXiv:2312.11158 (2023).
[54] Chopra, Ayush, Arnau Quera-Bofarull, Nurullah Giray-Kuru, Michael Wooldridge, and Ramesh Raskar. "Private Agent-Based Modeling." In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.
[55] Dyer, Joel, Arnau Quera-Bofarull, Nicholas Bishop, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. "Population synthesis as scenario generation for simulation-based planning under uncertainty."In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.

ABMs in complex system simulations. Additionally, Meyer-Vitali and Mulder (2024)[56] argue for the application of traditional engineering methods to enhance the reliability and trustworthiness of AI systems, emphasising the need for stringent design, validation, and certification processes, especially as AI systems become more complex and impactful, as mandated by frameworks like the European AI Act.

Emergent coordination among autonomous agents can be leveraged also for concrete applications. For instance, the research conducted by D'Amato et al. investigates innovative decentralised approaches to railway traffic management. In their 2024 paper,[57] They introduce the concept of self-organising railway traffic management, where trains operate as intelligent agents capable of redefining routes and schedules in response to disruptions. This approach aims to enhance scalability and resilience by eliminating the need for a central authority. They detail the principles and interactions of the decentralised sub-processes and provide a proof of concept using a realistic French railway control area, suggesting the viability of this method. In a follow-up study,[58] they further develop and test the coordination algorithm on a synthetic dataset, revealing that this decentralised approach often converges to optimal solutions comparable to those obtained by centralised methods, demonstrating its potential effectiveness. Together, these studies contribute to the advancement of autonomous, efficient, and resilient railway traffic management systems.

AI techniques can also support deeper understanding and policy decision making for urban dynamics and housing market phenomena. Termos et al. (2021)[59] utilise an ABM based on rent-gap theory to study the impact of refugee migration on West Asian urban environments, focusing on housing market changes and policy efficacy. Termos and Yorke-Smith (2022)[60] apply ABM to simulate Beirut's housing market resilience to historical and speculative crises, aiming to determine optimal capital for urban regeneration. Overwater and Yorke-Smith (2022)[61] investigate the effects of short-term peer-to-peer rentals, like Airbnb, on Amsterdam's housing market through an ABM that models residential migration and economic impacts, revealing the consequences of different regulatory policies. Wiegel and Yorke-Smith (2024)[62] explore the Dutch housing market's response to regulatory changes using an ABM to assess internal demand and the effects of policy measures on a market characterised by a significant social housing sector and supply shortages. Together, these studies demonstrate the application of ABM in examining urban dynamics, providing insights into policy impacts and housing market behaviour.

---

[56] Meijer-Vitali, A., and Mulder, W., (2024). Engineering Principles for Building Trusted Human-AI Systems, Intellisys 2024, Amsterdam
[57] D'Amato, L., Naldini, F., Tibaldo, V., Trianni, V. & Pellegrini, P. Towards self-organizing railway traffic management: concept and framework. J. Rail Transp. Plan. Manag. 29, 100427 (2024).
[58] D'Amato L., Pellegrini P. and Trianni V. A coordination algorithm for decentralised railway traffic management. Submitted to ECAI 2024.
[59] Termos, A.; Picascia, S.; and Yorke-Smith, N. Agent-Based Simulation of West Asian Urban Dynamics: Impact of Refugees. Journal of Artificial Societies and Social Simulation 24(1), 2:1–2:25. Jan. 2021.
[60] Termos, A. and Yorke-Smith, N. Market-Led Urbanism and Geographic Crises: A Micro-Simulation Lens on Beirut. Urban Planning, 7(1), 87–100. Feb. 2022.
[61] Overwater, A. and Yorke-Smith, N. Agent-Based Simulation of Short-Term Peer-to-Peer Rentals: Evidence from the Amsterdam Housing Market. Environment and Planning B: Urban Analytics and City Science, 49(1), 223–240. Jan. 2022.
[62] Wiegel, E. and Yorke-Smith, N. An Agent-Based Market Analysis of Urban Housing Balance in the Netherlands. Real Estate. 1, 80–135. Apr. 2024.

AI-based methods can be applied across various domains using innovative approaches. Wickramasooriya et al. (2024)[63] employ an agent-based model (ABM) to simulate the deployment and dynamics of genetically engineered mosquitoes with gene drive technology, aiming to control malaria vectors. Zhou et al. (2024)[64] investigate financial credit networks through a strategic analysis of prepayments, utilising empirical game-theoretic analysis (EGTA) to identify Nash equilibria and analyse the strategic behaviour of firms. Serramia et al. (2023)[65] develop a method to compute value-aligned norms by encoding ethics into normative systems, framing it as an optimization problem and solving it with standard tools. Bootsma and Mulder (2023)[66] propose extending the Smart Connected Supplier Network (SCSN) vocabulary to improve the resilience of supply chain networks, particularly under uncertain market conditions, by introducing new message types to mitigate risks and enhance trust. These studies demonstrate the versatility of AI methods in addressing complex problems across diverse fields.

2.5. Applications and Impact (T6.5)

The field of social AI has focused our researchers into the realm of AI Assistants.
The advent of more sophisticated artificial intelligence (AI) assistants signals the onset of a technological paradigm shift. Early assistant technologies like Amazon's Alexa and Apple's Siri utilised narrow AI for functions such as text-to-speech and intent classification. In contrast, the new generation of advanced AI assistants employs general-purpose foundation models, enhancing their versatility, autonomy, and range of applications. These advanced assistants offer innovative services to users, including summarization, ideation, planning, and tool use—capabilities expected to evolve as the technology advances. Consequently, advanced AI assistants have the potential to become deeply integrated into our economic, social, and personal lives, transforming how humans interact with and perceive AI.

AI assistants are becoming integrated into nearly every facet of our lives. They have the potential to act as interactive partners, tutors, creative collaborators, research assistants, counsellors, companions, friends, and resources for making long-term plans or life goals. As such, AI assistants could profoundly transform work, education, and creative pursuits, as well as how we communicate, coordinate, and negotiate with each other, ultimately shaping who we aspire to be and become.

[63] Wickramasooriya, S., I. Mahmood, A. Calinescu, M. Wooldridge, and G. Lanzaro. "Exploring the dynamics of gene drive mosquitoes within wild populations using an agent-based simulation." (2024). The Annual Modeling and Simulation Conference (ANNSIM '24), 20th-23rd May 2024, Washington, D.C.
[64] Hao Zhou, Yongzhao Wang (University of Michigan), KONSTANTINOS VARSOS, Nicholas Bishop (University of Oxford), Rahul Savani, Anisoara Calinescu, Michael Wooldridge. "A Strategic Analysis of Prepayments in Financial Credit Networks". Accepted at The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24).
[65] Serramia, Marc, Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. "Encoding Ethics to Compute Value-Aligned Norms." Minds and Machines (2023): 1-30.
[66] Bootsma, J., Mulder, W., (2023). "If only I knew: Extending the SCSN vocabulary to Improve the Resilience of Supply Chain Networks". PRO-VE 2023.

We continue with our Talking Buildings project[67], have new ideas for a project in the field of cancer-treatment, and are active on the crossing border of techniques and thought in the domain of mental healthcare.

2.6 Fostering the AI scientific community on the theme of social AI (T6.6)

This task focused on fostering activities such as bilateral and multilateral meetings among scientists, facilitating student visits, organising workshops on Social AI to promote the field, coordinating summer schools, the TAILOR conference, and other collaborative initiatives related to Social AI.

As evidenced by the list of publications (refer to Section 3.3), WP6 has significantly contributed to the academic discourse, with numerous papers acknowledging TAILOR published in prestigious AI and interaction conferences such as AAMAS, IJCAI, AAAI, HRI, and CHI, totaling approximately 43 publications and 6 posters. Many of these publications feature collaborations with authors from outside the TAILOR network. Moreover, some publications were presented in non-traditional AI conferences, underscoring the multidisciplinary nature of AI in fields like human factors and human-robot interactions.[68] [69].[70] [71] [72] [73].

In terms of academic exchanges, (refer to Section 3.3), there were several visits during this period between institutions within the TAILOR network (e.g., IST-VUB, IST-UvA, IST-Bielefeld University, TNO-University of Groningen, VBU-Warwick University, SCIS-Universita' della Campania, SCIS-Technical University of Crete, Oxford- Bar Ilan University). Additionally, over 15 presentations were organised between universities and companies under the scope of WP6.

Additionally, six TAILOR in-person meetings were organised for knowledge share and synergies with Industry partners. Five successful summer schools were organised in Barcelona (2022,2023,2024), Ljubljana (2023), and Athens (2024), and engaged more than 100 students in total.

[67] https://www.tno.nl/en/newsroom/insights/2022/07-0/talking-buildings-pleasing-partnership/
[68] Conveying Emotions through Shape-changing to Children with and without Visual Impairment. I Neto, Y Hu, F Correia, F Rocha, G Hoffman, H Nicolau, A Paiva, (CHI 24)
[69] "I'm Not Touching You. It's The Robot!"": Inclusion Through A Touch-Based Robot Among Mixed-Visual Ability Children. I Neto, Y Hu, F Correia, F Rocha, J Nogueira, K Buckmayer, G Hoffman, H Nicolau, A Paiva, (HRI 24)
[70] The Effects of Observing Robotic Ostracism on Children's Prosociality and Basic Needs. F. Correia, I Neto, S. Paulo, P. Piedade, H Erel. A. Paiva H Nicolau, (HRI 24)
[71] The Robot Made Us Hear Each Other: Fostering Inclusive Conversations among Mixed-Visual Ability Children. I Neto, F. Correia, F. Rocha, P. Piedade,. A. Paiva H Nicolau, (HRI 23)
[72] The Expression of Emotions in Cooperators and Defectors under Indirect Reciprocity (2024), Henrique Fonseca - under revision
[73] The Evolution of Cooperation under Indirect Reciprocity in the presence of Strangers (2024), Henrique Fonseca - under revision

3. Overview of Activities

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

A Monthly meeting was set up since the beginning 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 (with more than 15 presentations made).

3.1 WP6 and other work packages relation

In a network as extensive as ours, the work undertaken in WP6 must be viewed in conjunction with other work packages, particularly concerning scientific challenges. The primary objective of Trustworthy AI is to establish methods, processes, and algorithms to develop artefacts capable of autonomously acting in our world or making decisions that both companies and humans trust as intelligent. WP6 focuses specifically on the social aspects of constructing such artefacts, but it maintains strong interconnections with all other work packages, as illustrated in Figure 3.1.

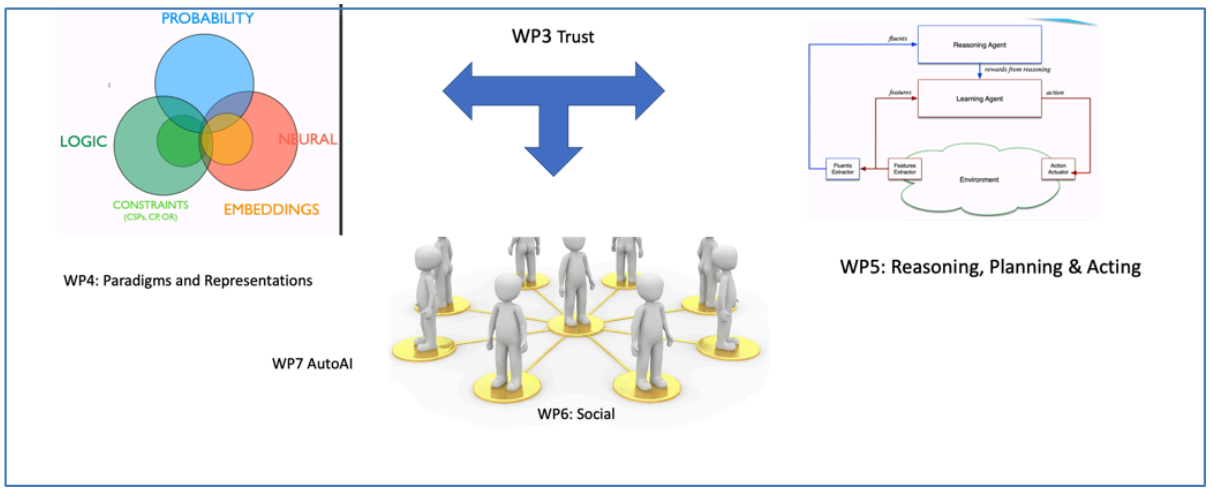

Figure 3.1 - Interconnection between WP6 and other work packages

1. **Link between WP3 and WP6** : Partners in WP6 play integral roles in WP3, particularly in addressing issues of trust, privacy, and transparency essential for models involving humans as agents.
2. **Link between WP4 and WP6:** WP4 serves as a foundational platform for researching paradigms and representations. In WP6, these paradigms and representations form the basis for the creation of models for social interactions between agents.
3. **Link between WP5 and WP6:** WP5 focuses on planning and action, crucial elements in social scenarios. Many partners were involved in both WP5 and WP6 due to their intertwined nature.
4. **Link between WP6 and WP7:** Given that autoAI development necessitates interaction with humans, contributions from social AI in WP6 were highly relevant for WP7.

These interconnections underscore how WP6 collaborates closely with other work packages to advance the broader goals of Trustworthy AI within the TAILOR network.

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), thereby advancing the Scientific State-of-the-Art for these foundations. This contribution is evident in achieving the following key performance indicators (KPIs):

#3.4 Research visits of at least 5 days within the network

#3.5 Research visits of at least 5 days from outside the network

#4.1 Ranking and number 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 can be seen from the list of publications (see Section 3.3), WP6 has contributed to more than 40 of published works acknowledging TAILOR in this area, including several papers published in high impact conferences, namely AAMAS, IJCAI, AAAI. A short number of these include authors from different partners. In terms of visits, there were eleven major exchanges. In total we had 43 papers, 6 posters, 16 presentations, 11 visits (to and from TAILOR partners), 5 summer schools and 6 in-person meetings.

3.3 List of papers and collaborations from this WP

In this section we provide a list of papers and collaborations 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:

Papers list :

- IST-UL
    - Conveying Emotions through Shape-changing to Children with and without Visual Impairment. I Neto, Y Hu, F Correia, F Rocha, G Hoffman, H Nicolau, A Paiva, (CHI 24) https://dl.acm.org/doi/10.1145/3613904.3642525
    - "I'm Not Touching You. It's The Robot!"": Inclusion Through A Touch-Based Robot Among Mixed-Visual Ability Children. I Neto, Y Hu, F Correia, F Rocha, J Nogueira, K Buckmayer, G Hoffman, H Nicolau, A Paiva, (HRI 24) https://dl.acm.org/doi/10.1145/3610977.3634992 (Honorable Mention)
    - The Effects of Observing Robotic Ostracism on Children's Prosociality and Basic Needs. F. Correia, I Neto, S. Paulo, P. Piedade, H Erel. A. Paiva H

Nicolau, (HRI 24) https://dl.acm.org/doi/10.1145/3610977.3634997 (Honorable Mention)
    - The Robot Made Us Hear Each Other: Fostering Inclusive Conversations among Mixed-Visual Ability Children. I Neto, F. Correia, F. Rocha, P. Piedade,. A. Paiva H Nicolau, (HRI 23) https://dl.acm.org/doi/10.1145/3568162.3576997 (Honorable Mention)
    - The Expression of Emotions in Cooperators and Defectors under Indirect Reciprocity (2024), Henrique Fonseca - under revision
    - The Evolution of Cooperation under Indirect Reciprocity in the presence of Strangers (2024), Henrique Fonseca - under revision
- CNR
    - D'Amato, L., Naldini, F., Tibaldo, V., Trianni, V. & Pellegrini, P. Towards self-organizing railway traffic management: concept and framework. J. Rail Transp. Plan. Manag. 29, 100427 (2024).
    - Abels, A., Lenaerts, T., Trianni, V. & Nowé, A. Expertise Trees Resolve Knowledge Limitations in Collective Decision-Making. in Proceedings of the 40th International Conference on Machine Learning vol. 202 79–90 (2023).
    - Abels, A., Lenaerts, T., Trianni, V. & Nowé, A. Dealing with expert bias in collective decision-making. Artif. Intell. 320, 103921 (2023).
    - Oddi, F., Cristofaro, A., Trianni, V. (2022). Best-of-N Collective Decisions on a Hierarchy. In: Dorigo, M., et al. Swarm Intelligence. ANTS 2022. Lecture Notes in Computer Science, vol 13491. Springer, Cham. https://doi.org/10.1007/978-3-031-20176-9_6
    - Miletitch, R., Reina, A., Dorigo, M. et al. Emergent naming conventions in a foraging robot swarm. Swarm Intell 16, 211–232 (2022). https://doi.org/10.1007/s11721-022-00212-1
    - Albani, D.; Hönig, W.; Nardi, D.; Ayanian, N.; Trianni, V. Hierarchical Task Assignment and Path Finding with Limited Communication for Robot Swarms. Appl. Sci. 2021, 11, 3115. https://doi.org/10.3390/app11073115
- University of Oxford
    - Stoian, M.C., Tatomir, A., Lukasiewicz, T. and Giunchiglia, E., 2024. PiShield: A NeSy Framework for Learning with Requirements. arXiv preprint arXiv:2402.18285.
    - Stoian, M.C., Dyrmishi, S., Cordy, M., Lukasiewicz, T. and Giunchiglia, E., 2024. How Realistic Is Your Synthetic Data? Constraining Deep Generative Models for Tabular Data. arXiv preprint arXiv:2402.04823.
    - Stoian, Mihaela Cătălina, Eleonora Giunchiglia, and Thomas Lukasiewicz. "Exploiting t-norms for deep learning in autonomous driving." arXiv preprint arXiv:2402.11362 (2024). NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, July 03–05, 2023, Certosa di Pontignano, Siena, Italy.
    - Chopra, Ayush, Arnau Quera-Bofarull, Nurullah Giray-Kuru, Michael Wooldridge, and Ramesh Raskar. "Private Agent-Based Modeling." In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.

    - Zennaro, Fabio Massimo, Nicholas George Bishop, Joel Dyer, Yorgos Felekis, Ani Calinescu, Michael J. Wooldridge, and Theodoros Damoulas. "Causally Abstracted Multi-armed Bandits." In The 40th Conference on Uncertainty in Artificial Intelligence.
    - Wickramasooriya, S., I. Mahmood, A. Calinescu, M. Wooldridge, and G. Lanzaro. "Exploring the dynamics of gene drive mosquitoes within wild populations using an agent-based simulation." (2024). The Annual Modeling and Simulation Conference (ANNSIM '24), 20th-23rd May 2024, Washington, D.C. Best paper award
    - Dyer, Joel, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. "Interventionally Consistent Surrogates for Agent-based Simulators." arXiv preprint arXiv:2312.11158 (2023).
    - Serramia, Marc, Manel Rodriguez-Soto, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Filippo Bistaffa, Paula Boddington, Michael Wooldridge, and Carlos Ansotegui. "Encoding Ethics to Compute Value-Aligned Norms." Minds and Machines (2023): 1-30.
    - Hao Zhou, Yongzhao Wang (University of Michigan), KONSTANTINOS VARSOS, Nicholas Bishop (University of Oxford), Rahul Savani, Anisoara Calinescu, Michael Wooldridge. "A Strategic Analysis of Prepayments in Financial Credit Networks". Accepted at The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24).
    - Dyer, Joel, Arnau Quera-Bofarull, Nicholas Bishop, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. "Population synthesis as scenario generation for simulation-based planning under uncertainty."In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024. Poster & Presentation
- TU Delft
    - Wiegel, E. and Yorke-Smith, N. An Agent-Based Market Analysis of Urban Housing Balance in the Netherlands. Real Estate. 1, 80–135. Apr. 2024.
    - Bbeanu, A-I.; Filatova, T.; Kwakkel, J. K.; and Yorke-Smith, N. Adaptive Parallelization of Multi-Agent Simulations with Localized Dynamics. arXiv preprint 2304.01724. Apr. 2023.
    - Termos, A. and Yorke-Smith, N. Market-Led Urbanism and Geographic Crises: A Micro-Simulation Lens on Beirut. Urban Planning, 7(1), 87–100. Feb. 2022.
    - Overwater, A. and Yorke-Smith, N. Agent-Based Simulation of Short-Term Peer-to-Peer Rentals: Evidence from the Amsterdam Housing Market. Environment and Planning B: Urban Analytics and City Science, 49(1), 223–240. Jan. 2022.
    - Termos, A.; Picascia, S.; and Yorke-Smith, N. Agent-Based Simulation of West Asian Urban Dynamics: Impact of Refugees. Journal of Artificial Societies and Social Simulation 24(1), 2:1–2:25. Jan. 2021.
- TNO
    - Meijer, A., Mulder, W., (2024) Engineering Principles for Building Trusted Human-AI Systems, Intellisys 2024, Amsterdam

    - Mulder, W., Meyer-Vitali, A., (2023). "A Maturity Model for Collaborative Agents in Human-AI Ecosystems". PRO-VE 2023.
    - Bootsma, J., Mulder, W., (2023). "If only I knew: Extending the SCSN vocabulary to Improve the Resilience of Supply Chain Networks". PRO-VE 2023.
- VUB (including visiting researcher Paolo Turinni)
    - Chin-wing Leung, Tom Lenaerts and Paolo Turrini, To promote full cooperation in social dilemmas, agents need to unlearn loyalty, accepted at IJCAI'24
    - Bara, J., Santos, F.P. & Turrini, P. The impact of mobility costs on cooperation and welfare in spatial social dilemmas. *Sci Rep* **14**, 10572 (2024). https://doi.org/10.1038/s41598-024-60806-z
    - Dmitry Chistikov, Luisa Estrada, Mike Paterson and Paolo Turrini, Learning a Social Network by Influencing Opinions, AAMAS '24: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, May 2024, Pages 363 - 371
    - Chin-wing Leung and Paolo Turrini, Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out ,AAMAS '24: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, May 2024, Pages 1110 - 1118
    - A Abels, EF Domingos, A Nowé, T Lenaerts, Mitigating Biases in Collective Decision-Making: Enhancing Performance in the Face of Fake News, arXiv preprint arXiv:2403.08829
- CSIC
    - Alba Aguilera, Nieves Montes, Georgina Curto, Carles Sierra, Nardine Osman:Can Poverty Be Reduced by Acting on Discrimination? An Agent-based Model for Policy Making. AAMAS 2024: 22-30
    - Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, Carles Sierra:Combining Theory of Mind and Abductive Reasoning in Agent-Oriented Programming. AAMAS 2024: 2839-2841
    - Alba Aguilera, Nieves Montes, Georgina Curto, Carles Sierra, Nardine Osman:Can Poverty Be Reduced by Acting on Discrimination? An Agent-based Model for Policy Making. CoRR abs/2403.01600 (2024)
    - Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, Carles Sierra:Combining theory of mind and abductive reasoning in agent-oriented programming. Auton. Agents Multi Agent Syst. 37(2): 36 (2023)
    - Athina Georgara, Raman Kazhamiakin, Ornella Mich, Alessio Palmero Aprosio, Jean-Christophe R. Pazzaglia, Juan Antonio Rodríguez-Aguilar, Carles Sierra:The AI4Citizen pilot: Pipelining AI-based technologies to support school-work alternation programmes. Appl. Intell. 53(20): 24157-24186 (2023)
    - Nieves Montes, Nardine Osman, Carles Sierra:A Computational Model of Ostrom's Institutional Analysis and Development Framework (Extended Abstract). IJCAI 2023: 6937-6941
    - Mohamed Chetouani, Virginia Dignum, Paul Lukowicz, Carles Sierra:Human-Centered Artificial Intelligence - Advanced Lectures, 18th

European Advanced Course on AI, ACAI 2021, Berlin, Germany, October 11-15, 2021, extended and improved lecture notes. Lecture Notes in Computer Science 13500, Springer 2023, ISBN 978-3-031-24348-6 [contents]
    - Nieves Montes, Nardine Osman, Carles Sierra, Marija Slavkovik:Value Engineering for Autonomous Agents. CoRR abs/2302.08759 (2023)

Posters list per University:

- IST-UL
    - Recognition and Prediction Using Dynamic Movement Primitives (2023), Ali Kordia
    - Optimizing and Coordinating Multiple DMPs Under Constraints to Achieve Collaborative Manipulation Tasks (2024), Ali Kordia
- University of Oxford
    - ICLR'24 Poster: Stoian, M.C., Dyrmishi, S., Cordy, M., Lukasiewicz, T. and Giunchiglia, E., 2024. How Realistic Is Your Synthetic Data? Constraining Deep Generative Models for Tabular Data. arXiv preprint arXiv:2402.04823.
    - Chopra, Ayush, Arnau Quera-Bofarull, Nurullah Giray-Kuru, Michael Wooldridge, and Ramesh Raskar. "Private Agent-Based Modeling." In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.
    - Dyer, Joel, Arnau Quera-Bofarull, Nicholas Bishop, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. "Population synthesis as scenario generation for simulation-based planning under uncertainty."In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.
- TNO
    - Enhancing Collaborative Human-AI Decision-Making in Healthcare : Integrating Theory of Mind for Efficient and Trustworthy Collaborative Decision-Making (2024), Andra Minculescu

Presentations list

- CNR
    - A Consensus Algorithm for Decentralised Real-Time Railway Traffic Management, the 4th International Workshop on Artificial Intelligence for RAILwayS (AI4RAILS). International Conference on Optimization and Decision Science, Ischia, Italy, 2023
- IST-UL
    - Navigating the vast and dynamic digital world poses significant challenges in accessing accurate information, often complicated by manipulative emotional strategies.Dr. Sergio Muñoz, an Assistant Professor at the Polytechnic University of Madrid. Gaips Talks, June 2024

- University of Oxford
  - Chopra, Ayush, Arnau Quera-Bofarull, Nurullah Giray-Kuru, Michael Wooldridge, and Ramesh Raskar. "Private Agent-Based Modeling." In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.
  - Zennaro, Fabio Massimo, Nicholas George Bishop, Joel Dyer, Yorgos Felekis, Ani Calinescu, Michael J. Wooldridge, and Theodoros Damoulas. "Causally Abstracted Multi-armed Bandits." In The 40th Conference on Uncertainty in Artificial Intelligence.
  - Wickramasooriya, S., I. Mahmood, A. Calinescu, M. Wooldridge, and G. Lanzaro. "Exploring the dynamics of gene drive mosquitoes within wild populations using an agent-based simulation." (2024). The Annual Modeling and Simulation Conference (ANNSIM ’24), 20th-23rd May 2024, Washington, D.C. Presentation. Best paper award
  - Hao Zhou, Yongzhao Wang (University of Michigan), KONSTANTINOS VARSOS, Nicholas Bishop (University of Oxford), Rahul Savani, Anisoara Calinescu, Michael Wooldridge. "A Strategic Analysis of Prepayments in Financial Credit Networks". Accepted at The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24). Presentation
  - Dyer, Joel, Arnau Quera-Bofarull, Nicholas Bishop, J. Doyne Farmer, Anisoara Calinescu, and Michael Wooldridge. "Population synthesis as scenario generation for simulation-based planning under uncertainty."In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 381-390. 2024.
- TU Delft
  - Babeanu, A-I.; Filatova, T.; Kwakkel, J. K.; and Yorke-Smith, N. Adaptive Parallelization of Multi-Agent Simulations with Localized Dynamics. In: Proc. of AAMAS’23 Workshop on Multi-Agent-Based Simulation. London, UK. May 2023.
  - Wiegel,~E. and Yorke-Smith, N. No Hope for First-Time Buyers? Towards Agent-Based Market Analysis of Urban Housing Balance. In: Proc. of AAMAS'22 Workshop on Agent-Based Modelling of Urban Systems, 32--35. Auckland, New Zealand (virtual). May 2022.
  - Toman, M. and Yorke-Smith, N. Localised Reputation in the Prisoner’s Dilemma. In: Proc. of 33rd Benelux Conf. on Artificial Intelligence (BNAIC’21), 761–763. Esch-sur-Alzette, Luxembourg. Nov. 2021.
  - Yorke-Smith, N. Beirut: Social Simulation for Urban Dynamics. Dutch Benelux Simulation Society Symposium 2021, virtual event, Oct. 2021.
  - Gevers, L. and Yorke-Smith, N. Cooperation in Harsh Environments: The Effects of Noise in Iterated Prisoner’s Dilemma. In: Proc. of 32th Benelux Conf. on Artificial Intelligence (BNAIC’20), 414–415. Leiden, The Netherlands (virtual). Nov. 2020.
- CSIC
  - Carles Sierra, El papel de la Inteligencia Artificial en la evaluaci´on de la investigaci´on VI Jornadas de An´alisis de la Red de Bilbiotecas CSIC:

conocimiento y avance cient´ıfico, salto al futuro con Red, CSIC, Madrid, Spain, June 2024.
- Carles Sierra, On the Engineering of Social Values 4th TAILOR conference Trustworthy AI from Lab to market, TAILOR, Lisbon, Portugal, 4–5 June 2024.
- Carles Sierra, Can generative AI be made Trustworthy? 4th TAILOR conference Trustworthy AI from Lab to market,TAILOR, Lisbon, Portugal, 4–5 June 2024.
- Carles Sierra, Què pot fer la intel·ligència artificial per l'educació? Obertura del curs 2023/2024, Servei Educatiu Baix Llobregat VIII, Esplugues de Llobregat, Spain, September 2023.

Visiting list - number of visitors working on W6 per university

- Visiting to a Tailor partner
  - TNO:
    - Harmen de Weerd (University of Groningen)
  - Oxford
    - University of Oxford: Michael Wooldridge visited CSIC/IIIA in June 2022 to teach a one-week course on Computational Game Theory
  - VBU:
    - Paolo Turinni, (Warwick University visited the VUB. September 1st-30th & November 1st - December 15 2024
  - CSIC:
    - Beniamino di Martino from Universita' della Campania visits IIIA 2 weeks (April and July 2'24)
    - Gennaro Junior Pezzullo and Alessia Sabia, PhD students from Universita' della Campania visit IIIA for three months each in 2024.
    - Georgios Chalkiadakis from the Technical University of Crete visits IIIA for three months in 2024.

- Visiting to a non-TAILOR partner
  - IST-UL :
    - Isabel Neto (University of Amsterdam) , 1 month on June/July 2024
    - Filipa Correia (Bielefeld University), 1 month on July 2024
  - Oxford
    - University of Oxford: Michael Wooldridge visited Bar Ilan University in February 2023 to teach a one-week course on computational game theory
  - CSIC:
    - Carles Sierra visits Universita' della Campania 1 week in March 2024.
    - Beniamino di Martino from Universita' della Campania visits IIIA 2 weeks (April and July 2'24)

Meetings list

- TAILOR Conference 2023 (Vito Trianni and Andrea Orlandini, CNR) (Ana Paiva, IST)
- TAILOR meeting in Vaals (Wico Mulder, TNO) (Leander Schietgat, VUB) (Ana Paiva, IST)
- TAILOR Conference 2024 (Hao Zhu and Anisoara Calinescu, Oxford; Wico Mulder, TNO) (Tom Lenaerts VUB) (Fransciso Melo, Carla Pacheco, Isabel Neto, Ali Kordia , Henrique Fonseca and students volunteers, IST)
- ICAIF 2023, NY (Anisoara Calinescu, Oxford)
- IJCAI 2024 (forthcoming) (Hao Zhu and Michael Wooldridge, Oxford)
- ICLR 2024, Vienna (Thomas Lukasiewicz, Oxford)

Summer school list :

- TAILOR Schools 2022, 2023 and 2024 (more than 100 participants)
- AIHUB Summer School 2023, Barcelona. https://aihub.csic.es/escuela-verano-2023/
- AIHUB Summer School 2024, València https://aihub.csic.es/escuela-de-verano-2024/

4. Final Conclusions, and Reflections

We consider this area crucial for advancing trustworthy AI. The work presented here is preliminary and highlights a broad array of challenges and complexities. We have identified several obstacles that must be addressed for the further development of Trustworthy Social AI:

- Methodological: There is a pressing need for an interdisciplinary approach encompassing AI, statistics, psychology, mathematics, population biology, and more. Integrating these diverse fields is essential for both experimental and theoretical advancements.

- Institutional: Bridging the gap between AI and social sciences involves overcoming significant differences in researchers' backgrounds, funding mechanisms, editorial policies, and institutional support.

- Complexity: Trustworthy AI emerges from interactions among heterogeneous components. Understanding these complex systems requires novel perspectives on designing self-regulatory mechanisms and technologies, especially in contexts involving human-machine interactions and embodied AI.

In terms of our work, the core members of the work package were deeply engaged, and proudly collaborating with the view of TAILOR as a valuable platform for collaborative efforts in Social AI, aiming to drive impactful innovations in the field. Monthly meetings were proven instrumental for discussion and idea exchange. Events like the annual meeting and invited

talks have been well-attended, with +100 participants. Additionally, the several exchange visits constituted a unique opportunity to boost the academic career of the researchers and broaden their collaboration network. TAILOR project was a fantastic opportunity to meet, collaborate and find long-lasting synergies with European Universities in a highly important topic of Trustworthy AI and its impact on Social Dynamics.