Towards Sustainable Human-Agent Teams: A Framework for Understanding Human-Agent Team Dynamics

Blue Sky Ideas Track

Rui Prada
INESC-ID and Instituto Superior
Técnico, Universidade de Lisboa
Lisbon, Portugal
rui.prada@tecnico.ulisboa.pt

Astrid C. Homan
University of Amsterdam
Amsterdam, The Netherlands
A.C.Homan@uva.nl

Gerben A. van Kleef
University of Amsterdam
Amsterdam, The Netherlands
G.A.vanKleef@uva.nl

ABSTRACT

Human-agent teamwork is a promising research stream with great potential to impact society. Research on collaborative AI and human-agent interaction has tackled the problem from several perspectives, but we argue that a focus on teams as a unit and a model for human-agent team dynamics is missing. Such a focus is particularly relevant if we aim at involving agents as active team members and at building sustainable teams over time. A team perspective on human-agent collaboration requires new models that pose challenges for AI and humans alike. AI needs new models to build an understanding of team level variables, such as team structure and cohesion, to be able to monitor the team and act on the team beyond performing the task. Humans, in turn, need to be able to incorporate agents as team members in their mental models of teamwork and integrate them into team processes. Such human-agent team dynamics models should be built taking into account four different levels: individual, interpersonal, team, and organisational. We believe that to fulfill this vision we need to bring together the different fields of AI and social sciences.

KEYWORDS

Human-agent teams; collaborative AI; human-agent interaction; teamwork; team dynamics

ACM Reference Format:
Rui Prada, Astrid C. Homan, and Gerben A. van Kleef. 2024. Towards Sustainable Human-Agent Teams: A Framework for Understanding Human-Agent Team Dynamics: Blue Sky Ideas Track. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 5 pages.

1 INTRODUCTION

With the increasing inclusion of AI in the tools and processes used in workplaces, the collaboration between people and AI systems is both inevitable and desirable [11]. In particular, AI is gaining agency in the interaction with people which raises the need for new approaches to human-AI interaction and opportunities for collaboration. As this collaboration increases in complexity, number of participants, and time scale, the stress on the agents' capabilities increases. In this paper, we focus on collaborative tasks that involve teams (i.e. more than 2 members). We want to stress that moving towards teamwork with several members shifts the interaction dynamics from interpersonal to team-level dynamics and that both AI agents and humans need models to cope with human-agent team dynamics.

We aim to create autonomous agents who are embedded in the organisational structure. In this respect, teams are the cornerstone of organisational work [9]. Even though attention regarding the integration of automation in teams is surging [20], we still lack a clear understanding of how to develop effective, sustainable, and viable human-agent teams. We argue that we need a shift in our thinking about the use of agents in teams, not only by seeing agents as actual team members rather than just as tools [14].

New insights into agents and humans are needed. These can be built on top of AI teams and human-human teams research. However, this research is mostly developed in separate silos. We need a broad, integrative understanding of the characteristics of both fields. In particular, we need to shift our outlook on teamwork beyond human-agent interactions and consider the uniquely team-related aspects that are critical for understanding how to build successful human-agent teams.

2 TEAMING UP HUMANS WITH AGENTS

Work in the AI community has shown that agents are able to team up with other agents [6] [25], but the methods used do not work well when teaming such agents with people. For example, Carroll et al. [1] show that agents that learn how to play a collaborative game through self-play with each other do not perform well when playing together with a human. In the development of agents, the focus has been on developing task work skills rather than on skills that are needed for teamwork [14][20]. This focus makes it unlikely to create successful human-agent teams that persist over time.

Agents for human-agent teams need to take a human-centred approach, incorporating human factors in the supported interaction dynamics. Research on human-agent interaction has acknowledged this need. Agents built to interact with people have included computational models of emotional behaviour and social skills [19] that enable them to engage in social interaction and establish rapport, for example. However, to establish collaboration, the agents need to go beyond being able to engage in social interactions, as collaboration, although grounded on social interactions, requires a deeper engagement involving, for example, establishing common ground and shared mental models [21].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Moreover, most studies on human-agent collaboration have studied the interaction processes at the interpersonal level, even if the agents and humans are grouped in teams. For example, a recurrent concern in human-agent interactions is the study of trust [16][8], which is crucial for collaboration and teamwork. However, typically the research conducted models or evaluates if humans trust the agent(s) partner(s) but not if they trust the partnership. One can argue that in dyadic interactions trust in the partner is very much aligned with trust in the partnership, but they are different constructs. In the case of teams, this difference becomes quite relevant. Evidence shows that the collective intelligence of teams is more dependent on a team's emergent social processes than on the intelligence of the individuals that constitute the team [26]. Such social processes include interpersonal dynamics, but more prominently include team-level processes and organisational-level factors. Hence trusting a team is more than trusting its members individually. The development of trust in a team depends on team-level variables and dynamics, such as diversity of the members, task allocation, collective decision-making strategies, cohesion, team identity, and leadership. In other words, trust in a team should depend more on the team processes and attitudes of the members toward the team (i.e., a team-level approach) than on the characteristics of the individuals and their attitudes towards the other members (i.e., an interpersonal or dyadic approach).

Some studies of human-agent teams have reported the effects of team-level variables on the perception of humans and the performance of the team. For example, human attitudes in teams vary according to the size of the team and human/agent composition of the team [2], task allocation can influence human-agent interactions in teams [22], and agents displaying emotions at the team level positively affect team identification and trust [3]. In turn, agents can also affect team processes. For example, a robot moderator can influence team cohesion [24], support the resolution of conflicts [12], and a social robot leading a team can affect the team's performance by using different leadership styles [17].

However, in general, team variables and processes are measured in human-agent team studies but they are not modelled in the agents. We argue that agents need models of teamwork to be able to understand it and to be able to act on it. Agents need skills in how to perform on a team, independently of the tasks at hand. This is crucial if we aim at engaging the agents with teams that are persistent over time and that can perform different tasks, and is even more important for agents that deal with multiple team compositions and multiple teams, which is a common practice in organisations.

In turn, humans tend to have high task-work expectations of agents, expecting them to be perfect in their task role within the team, and tend not to expect teamwork behaviors from such agents [15]. Humans do not expect agents to focus on team processes [18]. In other words, the teamwork skills of agents that may be needed to make human-agent teams work better, paradoxically appear not to be expected or appreciated by humans.

The current mental model of humans is not yet aligned with a reality that incorporates teamwork with agents in organisational settings. The mental model of humans regarding agents is still more aligned with automation than with agency. The balance of responsibility and trust in the interaction with agents is currently a big challenge. The exaggerated expectations and unclear models of the capabilities of the agents may lead to overtrust, which hinders the success of the collaboration and may lead to the dismissal of responsibility by the human. Conversely, mistrust may lead to underperformance of the team as humans do not fully rely on the agents' capabilities.

The current state of affairs is not conducive to the successful incorporation of agents in organisational teams. Not only do we need more research, but we need to develop a new model of shared team cognition in which the models humans and agents have of human-agent teamwork are aligned to guide that research. Although this new model will be informed by models based on AI-AI teamwork and human-human teamwork, it will be different given the new nature of the members. For example, although human-like qualities of agents are important for effective interaction with agents, agents should not be regarded as humans. Agents need to be accepted as a new type of partner with specific qualities that may have superhuman performance in certain tasks and may have distributed embodiment that interacts with and perceives many different aspects of the environment at the same time. Furthermore, we should not disregard the ability of humans to adapt to the agents as well, as long as they are able to grasp an understanding of the agents and the teamwork dynamics.

Furthermore, we need a systemic view of human-agent teamwork to support designers and managers of teams. It is important for people in charge of setting up and managing human-agent teams to have control over important variables that shape and regulate the team interaction dynamics and drive its performance and maintenance processes. Human-agent teams need to fit the organisational needs and accommodate the individual characteristics and needs of their members (both human and agent).

3 A FRAMEWORK FOR UNDERSTANDING HUMAN-AGENT TEAM DYNAMICS

Given the multilevel structure in which organisational teams are embedded, we argue that to understand and develop human-agent teams, both humans and agents should be aware of the elements that affect the interaction. From the extensive research on human teams, we know that teams are complex and dynamic, and change and adapt over time [10]. Teams are composed of individual members, who bring their skills and needs to the team. Individual members have dyadic interpersonal interactions with other members which shape their relationships and their standing. When bringing individual members together, a certain team composition is created, which is characterized by processes and dynamics, which result in certain outcomes. Of course, on a higher level, teams are a part of departments and organisations, which set the context of teams by shaping the culture and structure of the organisation [7].

We propose a framework based on the input-process-output (IPO) model of team dynamics [10] to structure and guide the research on human-agent teams (see Figure 1) that incorporates the most common research questions on human-human teams. The framework identifies input factors that influence and restrict the team's dynamics. Such dynamics are defined by a set of processes that determine the team members' behaviour and support the team's performance and development, which are expressed by a

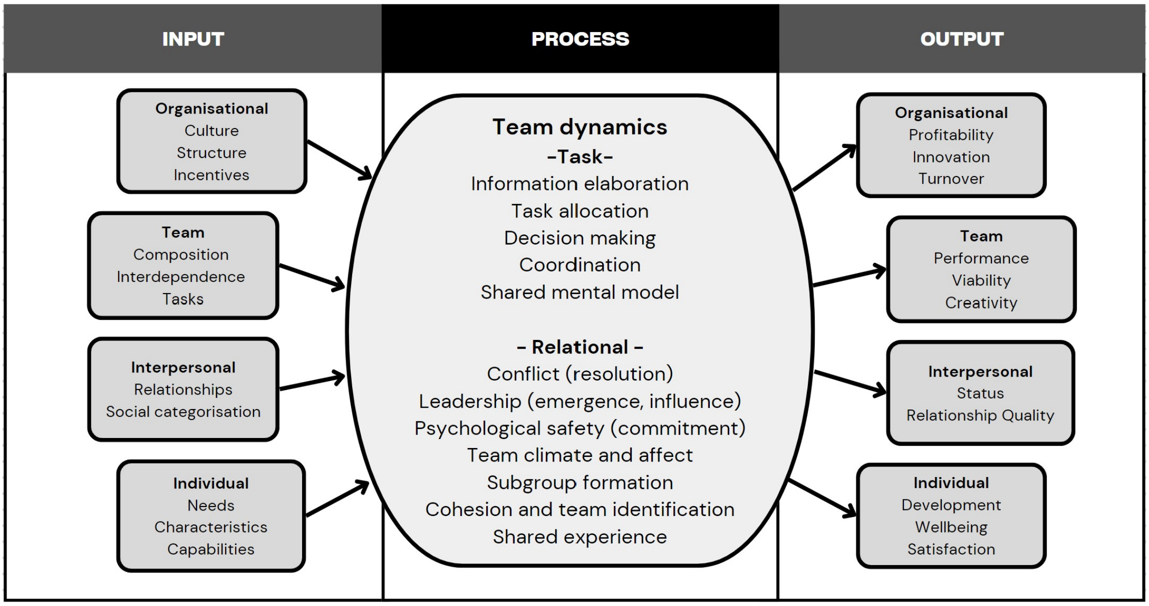

Figure 1: An Input-Process-Output framework for Human-Agent Teams research.

set of output variables that determine the team's outcomes and their effects. Note that, given the cyclical nature of team interactions [10] a feedback loop from outputs to input is expected.

It is important to consider the input and output variables in the multilevel structure of teamwork. First of all, we should study the influence of the *individual* characteristics, capabilities, and needs. This is particularly relevant, as agents and humans are quite different at this level. In turn, the processes of teamwork have an impact and shape the individuals as well. Performing in the team grants the opportunity for individuals to develop their knowledge and skills, and will affect their job satisfaction and well-being.

Teamwork is a social activity built on *interpersonal* interactions between team members. A priori social relations and social categorisation shape how team dynamics may develop. For example, it is easier to work together when relations of trust are established and tensions due to social comparison are low, i.e. individual differences are not a cause of problems. Distrust of AI technology and lack of acceptance of the different social nature of agents are challenges to address. Teamwork is, in fact, a way to develop the interpersonal level, as familiarity grows and relations can strengthen (or deteriorate) during team processes. Impact on relationship quality and social status is expected for both humans and agents.

At the *team* level, the input factors that are relevant are the composition of the team (number of members, type of members), the degree of interdependence of the members in reaching their goals or tasks, and the structure of the task and how it affects the team in terms of roles and hierarchies. These input factors shape the processes and outcomes of teams, by feeding into affective, behavioural, and cognitive concepts such as information exchange, motivation, process losses, emotions, psychological safety, and conflict behaviours. These team dynamics in turn feed into team outcomes such as performance, creativity, and viability. Of course, these relationships are not unidirectional and there are feedback loops between the different phases (e.g., when certain team members are not satisfied they might leave the team, which changes the composition of the team).

Finally, relevant input variables at the *organisational* level are the structure of the organisation that determines, for example, the access to resources and legitimate roles of social power (e.g., leadership), the culture of the organisation that establishes the social values that teams should follow, and incentives that define the expected rewards and penalties applied to the teams' performance. Output relevant variables are the profitability of the organisation, determined by the performance of its teams, the levels of innovation, and turnover, which are important to maintain the levels of the organisation's competitiveness and social impact.

At the core, team dynamics are driven by social processes that can be related to the execution of the task or deal with more relational aspects of the team's interactions. Notable task-related processes are information elaboration, task allocation, collaborative decision-making, coordination, and the development of shared mental models. In turn, common relational team dynamics processes deal with conflict resolution, leadership and social influence, psychological safety and participants' commitment, team climate and affect, subgroup formation, cohesion and team identification, and the development of shared experience.

4 DISCUSSION

The problem of creating human-agent teams needs to be addressed from the perspective of the AI and the perspective of the humans. The proposed framework provides insights into which aspects of teamwork are important to incorporate into these perspectives. It can be seen as a tool for mapping the landscape of research on human-agent teams, helping to position current research and identifying uncovered ground for future research. The framework identifies items that need to be considered while designing human-agent team scenarios, items that agents and humans need to address, and items that should be controlled and measured in studies.

We aim to raise attention to the importance of modelling the processes that compose team dynamics, in particular, the ones that are not directly concerned with task performance. We also highlight the importance of addressing teamwork factors at different levels (i.e. individual, interpersonal, team, and organisational).

We argue that to make a real impact, achieving sustainable human-agent teams that perform over time and in different contexts should be a core goal. To ensure this sustainability the four levels must be considered when developing agents for human-agent teams. Each level raises different challenges.

For example, agents need to be able to calibrate their behaviour to sustain human needs [5] and support wellbeing, since efforts to increase the level of autonomy in agents may backfire as it might harm humans' autonomy needs, while efforts to increase the level of task performance in the agents may backfire as this might harm humans' competence needs [13].

Additionally, agents need to be able to cope with the relational dimension of the interactions. For example, agents need to support the building of trust and support the humans' social relatedness needs taking into account the diversity of the members of the team and the subjectivity of the human participants. The introduction of the agents cannot harm the quality of the relationships that humans build and their social status on the team. The social status that agents take, depending on the role that they perform, must be accepted by the team members.

At the team level, the composition, including role assignment, and the task definition are crucial and should be defined taking interdependence into account, for example, when the team is less goal interdependent, cohesion is less likely to develop[4]. Agents, and team managers, need to observe and attend to issues in team dynamics that may harm team viability and performance, such as role uncertainty and feelings of being dispensable [23].

Furthermore, for the team to be sustainable at the organisational level, it needs to adhere to the organisation structure and culture. Proper incentives need to be studied and defined including those involving the agents, in the case that the agents are shared resources among different teams in the organisation, for example.

Finally, we believe that from a management point of view, setting up, deploying, monitoring, and managing human-agent teams is uncharted territory that needs further research to ensure that the organsational needs are met. Otherwise, human-agent teams are unlikely to succeed in the real world. This last point includes taking into consideration legal and ethical aspects as well.

The proposed framework can guide research toward achieving sustainable human-agent teams. From a methodological point of view, we propose research in the following items to achieve such a goal:

- As a first step, we need to gather knowledge and data from different research communities that are studying human-agent teams. In particular, we need to bring computer science and social sciences together, breaking the silos, to build a shared understanding of knowledge, data, and challenges in the research field. This may require establishing human-agent teams as a research field of its own.
- Building multi-agent simulations, based on the common knowledge and data, of teams with machine-like and human-like agents to experiment with different factors that impact team dynamics and the sustainability of teams.
- Performing experimental studies with humans and agents in team settings to understand the conditions and situations where good teamwork emerges. These studies need to address teams with several members (more than 2) and use agents with high levels of agency, which is currently not common practice.
- Identifying gaps in research models and data and conducting efforts to collect and curate datasets and build knowledge to close such gaps. This item is both a result and a requirement for the previous three items. Related to this is the definition of benchmarks that define human-agent team scenarios to support the research. The scenarios should include variables identified in the framework presented in this paper.
- Deploy human-agent teams in "natural" scenarios in organisational settings to increase the ecological validity of the results. This can make use of virtual scenarios (e.g. games) but should address the needs of real organisations and use people from such organisations as participants.

5 CONCLUSIONS

We argue that to create successful and sustainable human-agent teams, research should zoom in on team dynamics. These task and relational processes guide and emerge from the behaviour of team members, and are shaped by input factors on different levels. How to make these processes effective in human-agent teams is still an open question. We argue that we need to integrate insights from the perspective of AI and human studies to successfully answer this question. We need to study how agents are developed considering the different team processes and how humans integrate the agents in such processes. These can be built from the insights of novel human-agent team experiments, the systematic collection of data of human-agent teams' interactions, and by running social simulations with teams and their dynamics at the core.

It is crucial to model the team explicitly as a unit in the models of the agents to make it a central concept. And to study how a model of human-agent teams can be developed and installed in the mental models of humans and integrated into organisational processes. This research should address the different levels of analysis that influence team dynamics: individual, interpersonal, team, and organisational.

When developing agents for teams, we should aim at building foundational AI models of teamwork-specific skills that are independent of the task at hand and are able to work with diverse teams and contexts.

ACKNOWLEDGMENTS

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) grant UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020) and the HumanE-AI-Net project funded by the European Union Horizon 2020 research and innovation grant 952026.

REFERENCES

[1] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. In *NeurIPS'2019 - Proceedings of the 33rd Conference on Neural Information Processing Systems (Advances in Neural Information Processing Systems, Vol. 32)*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 5174–5185.

[2] Wan-Ling Chang, Jeremy P. White, Joohyun Park, Anna Holm, and Selma Šabanović. 2012. The effect of group size on people's attitudes and cooperative behaviors toward robots in interactive gameplay. In *Proceedings of RO-MAN'12 - the 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 845–850. https://doi.org/doi:https://doi.org/10.1109/ROMAN.2012.6343857

[3] Filipa Correia, Samuel Mascarenhas, Rui Prada, Francisco S. Melo, and Ana Paiva. 2018. Group-based emotions in teams of humans and robots. In *Proceedings of HRI 2018 - the 13th International Conference on Human-Robot Interaction*. ACM/IEEE, 261–269.

[4] Stephen H. Courtright, Gary R. Thurgood, Greg L. Stewart, and Abigail J. Pierotti. 2015. Structural interdependence in teams: An integrative framework and meta-analysis. *Journal of Applied Psychology* 100, 6 (2015), 1825–1846. https://doi.org/10.1037/apl0000027

[5] Edward L. Deci and Richard M. Ryan. 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry* 11, 4 (2000), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

[6] Barbara J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86, 2 (1996), 269–357.

[7] J. Richard Hackman and Charles G. Morris. 1975. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in Experimental Social Psychology* 8 (1975), 45–99.

[8] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527. https://doi.org/10.1177/0018720811417254

[9] Daniel R. Ilgen. 1999. Teams embedded in organizations: Some implications. *American Psychologist* 54, 2 (1999), 129–139.

[10] Daniel R. Ilgen, John R. Hollenbeck, Michael Johnson, and Dustin Jundt. 2005. Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology* 56 (2005), 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070250

[11] Matthew Johnson and Alonso Vera. 2019. No AI is an Island: The Case for Teaming Intelligence. *AI Magazine* 40, 1 (2019), 16–28.

[12] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of HRI'2015 - The 10th International Conference on Human-Robot Interaction*. ACM/IEEE, 229–236. https://doi.org/10.1145/2696454.2696460

[13] Sophie Kerstan and Jan B. Schmutz. 2022. How simple assumptions about AI change knowledge sharing in Human-AI team decision-making. In *Academy of Management 82nd Annual Meeting*. AOM.

[14] Lindsay Larson and Leslie A. DeChurch. 2020. Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly* 31, 1 (2020), 101377.

[15] Lindsay Elizabeth Larson and Aaron Schecter. 2022. AI teammate function and TMS in Human-AI teams. In *Academy of Management 82nd Annual Meeting*. AOM.

[16] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104–123. https://doi.org/doi/10.1037/1076-898X.6.2.104

[17] Sara L. Lopes, José Bernardo Rocha, Aristides I. Ferreira, and Rui Prada. 2021. Social robots as leaders: leadership styles in human-robot teams. In *Proceedings of RO-MAN'21 - the 30th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 258–263. https://doi.org/10.1109/RO-MAN50785.2021.9515464

[18] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of Management Review* 26, 3 (2001), 356–376. https://doi.org/10.5465/amr.2001.4845785

[19] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, and Ana Paiva. 2022. FAtiMA Toolkit: Toward an Accessible Tool for the Development of Socio-emotional Agents. *Transactions on Interactive Intelligent Systems* 12, 1 (2022), 1–30.

[20] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors* 64, 2 (2022), 904–938. https://doi.org/10.1177/0018720820960865

[21] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interactions* 6, GROUP (2022), 13:1–29. https://doi.org/10.1145/3492832

[22] Sarah Strohkorb Sebo, Ling Liang Dong, Nicholas Chang, and Brian Scassellati. 2020. Strategies for the inclusion of human members within human-robot teams. In *Proceedings of HRI'2020 - The 15th International Conference on Human-Robot Interaction*. ACM/IEEE, 309–317. https://doi.org/10.1145/3319502.3374808

[23] James A. Shepperd. 1993. Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin* 113, 1 (1993), 67–81. https://doi.org/10.1037/0033-2909.113.1.67

[24] Elaine Short and Maja J. Matarić. 2017. Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *Proceedings of RO-MAN'17 - the 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 385–390. https://doi.org/10.1109/ROMAN.2017.8172331

[25] Milind Tambe and Weixiong Zhang. 2000. Towards Flexible Teamwork in Persistent Teams: Extended Report. *Autonomous Agents and Multi-Agent Systems* 3 (2000), 159–183. https://doi.org/10.1023/A:1010026728246

[26] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (2010), 686–688.