Quantifying cooperation between artificial agents using information theory

Patricia Wollstadt, Matti Krüger

2022

Preprint:

This is an accepted article published in HHAI2022: Augmenting Human Intellect. The final authenticated version is available online at:
https://doi.org/10.3233/FAIA220224

June 2022

Poster: Quantifying cooperation between artificial agents using information theory

Patricia WOLLSTADT [a,1] and Matti KRÜGER [a]
[a] *Honda Research Institute Europe, Offenbach am Main, Germany*

Introduction When designing human-machine interaction (HMI) systems, it is often assumed beneficial if systems behave *cooperatively* towards a human operator [23,13, 2,5]. Cooperative machine behavior is assumed to lead to, for example, higher trust, acceptance, and usability, [8,23,5], while, on the other hand, pure automation has been criticized for leading to a lack of engagement, loss of expertise, or reduced trust on user side [9]. It is thus hypothesized that making HMI more cooperative leads to more satisfying and effective exchanges between machines and human users. The design of cooperative HMI systems requires a definition of cooperative interactions. Moreover, to be able to control, optimize, or evaluate system behavior and its effects on human users, it is necessary to quantitatively describe cooperative interactions [17,11]. Even though the interest in cooperative HMI [12,5,6] has significantly increased in recent years and other disciplines such as psychology [18], biology [21,20], or game theory [19,14] have a long tradition in researching cooperative behavior, the concept of cooperation in HMI and its quantification stays elusive [2,22,11]. In the present work, we therefore develop a novel definition of cooperative behavior in HMI contexts and present an approach to quantify cooperative behavior based on this definition, using recent methods from information theory. As a first demonstration, we successfully apply our approach to a model system from reinforcement learning.

Methods As a first step, we propose a novel definition of cooperative behavior based on prior work in HMI and related disciplines. As a prerequisite, our definition assumes two or more agents with joint or individual goals, where a) a goal is reachable via subtasks that are *interdependent* such that the agents have to coordinate their actions, and b) agents commit to working jointly in a coordinated fashion [2,12,4,5,10]. Then, we define cooperation as a joint, coordinated activity towards solving interdependent subtasks, which leads to a *mutual facilitation* of individual agents' actions with respect to the current goal [12,10]. To enable cooperative actions, agents have to be equipped with suitable sensors and effectors for manipulating the environment and information sharing. Furthermore, agents must be able to generate relevant internal representations of the environment and other agents, from which future actions can be planned an controlled.

It is central to definitions found in literature that the joint activity strives to facilitate individual agents' actions towards their (sub)goals. In other words, cooperation should lead to a synergistic effect of the joint effort towards the goal. To evaluate whether co-

[1] Corresponding Author: Patricia Wollstadt, Honda Research Institute Europe, Carl-Legien-Str. 30, 63073 Offenbach am Main, Germany; E-mail: patricia.wollstadt@honda-ri.de.

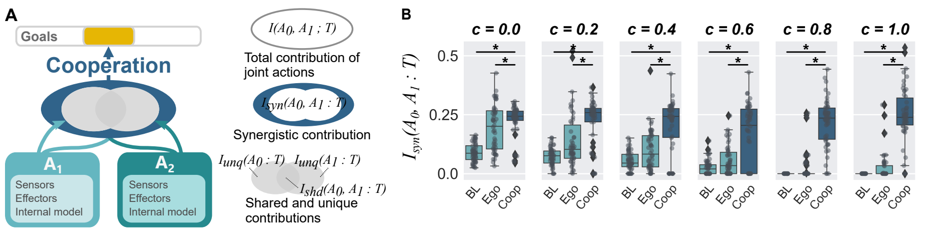

Figure 1. A) Illustration of PID framework. B) Synergy estimated from model system.

operation can indeed be identified such a synergistic effect, we propose to apply the recently introduced information-theoretic framework of partial information decomposition (PID) [24,25,3,7,15]. PID describes how two or more input variables and their interaction contribute to the outcome of a target variable. The contribution can either be provided *uniquely* by one of the inputs, it can be redundantly *shared* by both inputs, or it can be provided *synergistically* by both inputs, describing a contribution that is exclusively provided by both inputs together and that can not be obtained from one input alone (Figure 1A). We propose to use the synergy measure [3,16] to quantify the cooperative contribution of two or more agents' actions towards a common goal.

Experiments and Results As a first evaluation, we apply our definition and the proposed measure to the level-based foraging environment [1], a 2D grid-world, in which multiple agents and food items with different levels are placed. The agents' goal is to collect as many food items as possible, while an item can only be collected if an agent's level or the sum of the agents' levels simultaneously collecting the item is equal or larger than the item's level. We set up the environment to require different levels of cooperation between agents by setting the number of food items that could only be collected collectively, c, to 0, 20, 40, 60, 80, or 100 %, respectively. We further modified the heuristics for selecting an agent's next action proposed by [1] to obtain agents capable of various degrees of cooperative and non-cooperative behavior: i) a baseline heuristic (`BL`) that selects the next action at random, ii) a non-cooperative heuristic (`Ego`) that always goes to the closest visible food and tries to collect it irrespective of its level, iii) a cooperative heuristic (`Coop`) implementing our cooperation definition that targets the food that is closest to the center of all agents and that is compatible with the agents' summed level. To quantify the degree of cooperation, we estimated the synergy between the two agents' actions as input variables, and the current sum of collected food items as target variable. We use the measure proposed in [3] and implemented in [16]. As hypothesized, cooperative behavior was reflected by a high synergistic contribution of agents'actions towards the target. We found that synergy was significantly higher for the `Coop` than for both baseline heuristics, and that the difference was more pronounced in cooperative environments (Figure 1B).

Conclusion We introduced a novel framework for quantifying cooperative behavior using recent methods from information theory. Our approach is scenario agnostic, and does not require a-priori knowledge of possible agent strategies or behaviors. The approach makes only mild assumptions about data observable from the interaction such that we believe it to be applicable to a wide range of scenarios. We successfully demonstrate a first application in a model system were we find a clear distinction between cooperative and non-cooperative agent behaviors by the proposed measure. Evaluations in more complex scenarios, ideally involving human agents, will be subject to future work.

References

[1] Stefano V. Albrecht and Subramanian Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. *12th International Conference on Autonomous Agents and Multiagent Systems 2013, AAMAS 2013*, 2:1155–1156, 2013.

[2] Klaus Bengler, Markus Zimmermann, Dino Bortot, Martin Kienle, and Daniel Damböck. Interaction Principles for Cooperative Human-Machine Systems. *Information Technology*, 54(4):157–164, 2012.

[3] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[4] Michael E. Bratman. Shared cooperative activity. *The Philosophical Review*, 101(2):327–341, 1992.

[5] Judith Bütepage and Danica Kragic. Human-robot collaboration: from psychology to social robotics. *arXiv preprint arXiv:1705.10146*, 2017.

[6] Riccardo Gervasi, Luca Mastrogiacomo, and Fiorenzo Franceschini. A conceptual framework to evaluate human-robot collaboration. *The International Journal of Advanced Manufacturing Technology*, 108:841–865, 2020.

[7] Aaron J. Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, 477(2251):20210110, 2021.

[8] Marc Hassenzahl and Holger Klapperich. Convenint, clean, and efficient? The experiential costs of everyday automation. In *Proceedings of the 8th nordic conference on human-computer interaction: Fun, fast, foundational*, pages 21–30, 2014.

[9] Jean-Michel Hoc. From human–machine interaction to human–machine cooperation. *Ergonomics*, 43(7):833–843, 2000.

[10] Jean Michel Hoc. Towards a cognitive approach to human-machine cooperation in dynamic situations. *International Journal of Human Computer Studies*, 54(4):509–540, 2001.

[11] Nathanael Jarrasse, Vittorio Sanguineti, and Etienne Burdet. Slaves no longer: review on role assignment for human-robot joint motor action. *Adaptive Behavior*, 22(1):70–82, 2013.

[12] Gary Klein, Paul J. Feltovich, Jeffrey M. Bradshaw, and David D. Woods. Common ground and coordination in joint activity. *Organizational Simulation*, 53:139–184, 2005.

[13] Matti Krüger, Christiane B. Wiebel, and Heiko Wersing. From tools towards cooperative assistants. *Proceedings of the 5th International Conference on Human Agent Interaction - HAI '17*, pages 287–294, 2017.

[14] Xueqin Liang and Zheng Yan. A survey on game theoretical methods in human–machine networks. *Future Generation Computer Systems*, 92:674–693, 2019.

[15] Abdullah Makkeh, Aaron J Gutknecht, and Michael Wibral. Introducing a differentiable measure of pointwise shared information. *Physical Review E*, 103(3):032149, 2021.

[16] Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. Broja-2pid: A robust estimator for bivariate partial information decomposition. *Entropy*, 20(4):271, 2018.

[17] Eric Meisner, Selma Šabanović, Volkan Isler, Linnda R. Caporael, and Jeff Trinkle. Shadowplay: A generative model for nonverbal human-robot interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction, HRI'09*, pages 117–124, 2009.

[18] Alicia P. Melis and Dirk Semmann. How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2663–2674, 2010.

[19] Stefanos Nikolaidis, Jodi Forlizzi, David Hsu, Julie Shah, and Siddhartha Srinivasa. Mathematical models of adaptation in human-robot collaboration. *arXiv preprint arXiv:1707.02586*, 2017.

[20] Martin A. Nowak. Evolving cooperation. *Journal of Theoretical Biology*, 299(0):1–8, 2012.

[21] Joel L. Sachs, Ulrich G. Mueller, Thomas P. Wilcox, and James J. Bull. The evolution of cooperation. *The Quarterly Review of Biology*, 79(2):135–160, 2004.

[22] Alessandra Sciutti, Ambra Bisio, Francesco Nori, Giorgio Metta, Luciano Fadiga, Thierry Pozzo, and Giulio Sandini. Measuring Human-Robot Interaction Through Motor Resonance. *International Journal of Social Robotics*, 4(3):223–234, 2012.

[23] Bernhard Sendhoff and Heiko Wersing. Cooperative intelligence-a humane perspective. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020.

[24] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[25] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv Preprint arXiv:1004.2515 [cs.IT]*, pages 1–14, 2010.