A Maturity Model for Collaborative Agents in Human-AI Ecosystems

Wico Mulder[1], André Meyer-Vitali[2]

[1] TNO, Netherlands
wico.mulder@tno.nl
[2] DFKI, Germany
andre.meyer-vitali@dfki.de

Abstract. AI entities lean on the aspects of their autonomy to carry out their tasks and perform intelligently. But when these entities collaborate in human-AI teams, their levels of autonomy and collaboration have to be balanced out. We present a maturity model for agents regarding this aspect of balancing. Whereas simple AI systems use pre-designed mechanisms, more advanced systems are able to learn this from experience. The maturity model is a two-dimensional matrix in which the degree of agency forms the horizontal axis, and the level of interaction the vertical axis. We validate the use of this maturity model with use-cases in the field of urban energy efficiency.

Keywords: agency · collaborative networks · human-AI teaming

1 Introduction

When humans and AI entities in the form of agents collaborate, the AI entities are often characterised by a high degree of autonomy [1]. This autonomy is required for delegating tasks, intelligent behaviour and making decisions, but also leads to a dilemma: in order to act at a group level and participate in collaborative decision-making, AI Entities have to deliberate and possibly adapt their behaviour to influences of others in the group [17, 12]. While sharing information with other team members, humans as well as AI entities, they have to balance their autonomy on the cost of adaptivity to decisions of others [13].

The evolution of AI entities regarding their interaction with humans can be seen from two perspectives; on the one hand we see a technological evolution resulting in more advanced individual AI entities, on the other hand we see an evolution from isolated tool-based expert systems towards interactive and collaborative systems. The latter has also led to systems in which AI entities interact with each other as well as with humans. The role of such agents is evolving from being merely task-oriented and assistive to being collaborative companions that care for each other and in some situations also for their the surrounding environment [5].

We introduce the concept of *Human-AI ecosystems*, which refers to a mixed group of humans and AI entities that interact with each other in order to solve

tasks but also share the responsibility of preserving the environment that allows them to carry out those tasks. In an ecosystem each and every member is valued for their strengths and can be supported by the rest of the group when there is a need to. It is thereby essential to be aware of the diverse mutual dependencies within such an ecosystem. In Human AI ecosystems, humans are not subservient to AI entities and AI entities are also not always subservient to humans. Both are mutually enforcing each other while meeting their individual as well as their team goals. The different strengths and weaknesses of the members in the ecosystem complement each other to survive, individually as well as at group level.

In this paper we propose a maturity model that can be used to identify the maturity of collaboration of AI entities in such human-AI ecosystems. The model can be used to reflect on expected capacities, the role and the responsibilities of AI entities with respect to the team. It can be used in the process of planning and engineering AI entities to act in a human AI ecosystem as well as in the process of road mapping [11].

The rest of the paper is organised as follows: Section 2 introduces the structure of the maturity model. Section 3 validates the use of the model in an urban energy management case. Section 4 discusses on possible other cases and future work. Section 5 concludes the paper.

2 Collaborative Agent Maturity Model

We present our *collaborative agent maturity model (CAMM)*. The model takes into account the levels of agency and interaction of AI entities and can be used to identify the maturity of collaboration of these entities in human-AI ecosystems. Note that the term 'model' is an ambiguous term in the field of AI, as it may also relate to a computational model, theoretical model or architectural model. Here we use the term model to refer to a framework that describes the capabilities of an AI entity in terms of processes, expectations and practices while it acts in a human-AI ecosystem. An extensive overview of AI maturity models is written by Sadiq et al. (Todo : reference opnemen).

The model manifests itself in the form of a two dimensional matrix (see figure 1). Horizontally we differentiate stages of agency ranging in level from human controlled to fully autonomous. Vertically we identify stages of interaction, varying from merely sharing information about alignment of tasks, to forms of interaction in which an agent takes into account the mental state of others. The arrows represent the actual and designed future states of the AI entities in the use-cases (see section 3).

2.1 Levels of agency

Artificial Intelligence is based on the principles of autonomy and agency. Autonomy, the quality or state of being self-governing, is required to avoid purely

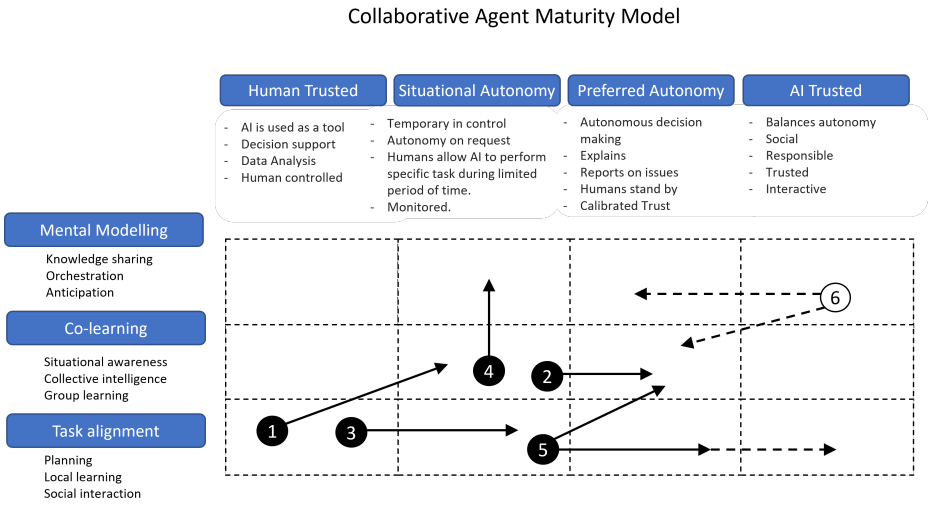

Fig. 1. the Collaborative Agent Maturity Model. It expresses the maturity in collaboration of an AI entity (agent) in a human-AI ecosystem in terms of its level of agency (horizontal) and form of interaction (vertical). The starting point of the arrow reflects the current maturiry level of the AI entity, whilst the head points towards its desired or planned maturity level.

predictable and reactive behaviour[18]. Whenever an AI entity commits on contributing to a team intention, it has to balance its level of autonomy with required levels of interaction.

Agency refers to the capacity of acting as a result of making decisions. An AI entity that possesses agency shows the ability to make choices and take actions based on its internal decision-making processes, rather than simply following a set of pre-determined rules or being controlled by an external operator [14].
In the model we distinguish four levels of agency with respect to an AI entity acting in a team:

1. **Human trusted**: AI entities (agents) provide operational assistance or facilitate in decision making. Humans are using the AI entity in the form of a tool at their own responsibility. One can think of agents or algorithms that help with design of manufactured goods, help in medical diagnostics, or classify customers buying items.
2. **Situational autonomy**: The AI entity carries out dedicated tasks and acts autonomously during limited periods of time, often on request by a human. Humans maintain a central role. They control the process and have responsibility as they are the ones who oversee situations and can quickly and creatively handle or decide on next steps. The AI is still assisting, but instead of acting as a tool, it takes the role of a companion. One may think of AI in self driving cars and algorithms that trade on a market. Examples can

be found in the field of advanced monitoring in networked systems [9], and in the field of human-centric working [10].

3. **Preferred autonomy**: AI entities preferably take decisions by themselves while humans stand by. They take part in deliberations and collective decision-making processes and share their findings when appropriate, i.e. when overruling decisions or when something goes wrong. In the team the AI entities and the humans lean on their mutual dependencies and collaborate in their production tasks. B humans are still taking final responsibility. Transparency and explainability allow the AI entity to be understood and trusted. Examples can be found in the early domain of explainable AI systems [19, 8, 2].
4. **AI trusted**: An AI trusted agent is able to learn and adapt to changes in their environment without human intervention. It can automatically adjust its balance between its autonomy with necessary interaction in the team. It learns from earlier interactions with others in the team both agents as well as humans. As a result it is prepared to deal with disruptive surprises and as a result it can contribute to the resilience of the human-AI ecosystem.

2.2 Levels of interaction

Interaction is a prerequisite for deliberation, delegating tasks and sharing knowledge within an ecosystem. We distinguish levels of interaction varying from simple sharing information about coordination of tasks, to higher orders of interaction that include the exchange of information about the learning process and each others mental states.
The maturity model distinguishes three levels of interaction:

1. **Task alignment**: Sharing information about tasks and planning is crucial for beneficial coordination in a team. In recent work [7] we explained the use of interaction design patterns. They facilitate the understanding of team processes and make them more transparent in terms of their internal task-handling processes. The modularity of those patterns facilitates the design and operation of teams that consist of humans and agents.
2. **Co-learning**: The AI entities not only exchange information about their individual tasks, but also about their learning experience, the learning process and models that they have learned. Co-learning allows human-AI teams to be adaptive and better deal with changing environments [15] . In literature also the term *Federated learning* is used when multiple actors build a common model without sharing data.[21]. Critical issues such as data privacy and access to heterogeneous data can be addressed properly. The term 'federated' is used to stress the strength of the learning approach in terms of an organisation, a federation of entities.
3. **Mental modelling**: In order to take into account the behaviour of others and anticipate them, AI entities share information about their mental state. This can be done either in an explicit way, i.e. by means of communicating state information or share concise descriptions of their knowledge, or implicitly by means of observing and reasoning. When humans and software

agents share goals they typically share knowledge on their individual beliefs and intentions [16, 6].
Rather than reasoning only with one's own beliefs, desires, intentions, emotions, and thoughts, a person or agent with the awareness of others' states of mind can consider different and possibly more mindful acts. Examples are AI systems that use the concept of Theory of Mind which allows AI entities to more easily understand, predict, and even manipulate the behaviour of others and thereby respond adequately [20].

3 Using the maturity model

We validate the use of the maturity model and indicate the stages of maturity of the AI entities in practical use-cases.

The "Talking Buildings project" is an applied research project in the field of urban energy management. AI entities that represent buildings are teaming up with humans to minimize their energy consumption while maintaining the required levels of comfort. Inside a building humans and AI entities strive to energy efficiency. They are part of a human-AI ecosystem. Similarly, when zooming out and regarding buildings on a campus or city level we see a human-AI ecosystem involving a group of buildings that deliberate on energy consumption in order to reduce peak loads, while taking into account their individual energy needs. The humans in this group can be building owners, policy makers or maintenance engineers.

The project adheres to the general shift of focus from individual energy efficiency systems containing isolated smartness towards interactive and collaborative energy management systems. The evolution of the notion of human-AI ecosystems goes hand in hand with contemporary challenges and developments in the field of resilient power grid infrastructures and networks of renewable energy sources.

We use the maturity matrix to identify and classify the maturity of the AI entities in the following five use-cases:

1. **Zone comfort**: Inside buildings AI algorithms at zone-level collaborate in the process of operational climate control. The AI entities represent various zones in a building. Their task is to minimize energy consumption while keeping the temperatures in the zones within human defined comfort levels. Sensors provide data on occupancy, climate and user preferences. In figure 1 we identify these agents with arrow number 1. Many AI systems today can be categorized at this maturity level.
2. **Co-learning buildings**: A campus setting involves multiple buildings. AI entities representing those buildings negotiate on time and power level for the heating and cooling systems. They optimize on their own energy consumption and take into account the energy needs of others as well. They also interact with human grid operators and take into account the preferences of the building owners. Together, they learn so-called flexibility profiles in a

collaborative co-learning setting. This type of human-AI ecosystem, e.g. a campus or a part of the city, avoids heating the buildings up all at once in the morning or, when its getting hot outside, avoids turning on their cooling systems all at once. The case also involves the act of learning and clustering buildings based on their energy consumption behaviour. In the figure the maturity of the AI entities is denoted with arrow number 2. The evolution towards the maturity level of co-learning is a common step in distributed systems. Examples can be also be found in other industrial domains, e.g. in distributed planning in the field of smart manufacturing [4].

3. **Completion of missing parts**: This case is about the completion of missing information. It sometimes happens that sensor-overviews, floorplans and even designs of climate systems are not up-to-date. It is the collaborative task of the humans and the AI entities in the ecosystem to complete the missing links. Each member of the human-AI ecosystem may have a certain notion on how the sensor-data and the building topology relate with each other. In the figure the AI entities involved are denoted with arrow number 3.
4. **Deliberation**: This use-case focuses on the interaction in the human-AI ecosystems. The interactions are about decision-making in which the team members take each other's mental state into account as they deliberate and decide how to fulfil their individual energy needs. While each actor has its own local desires, the group goal is to reduce energy peaks on the grid. Current research in the project focuses on using the Theory of Mind. In the figure this is indicated with arrow 4.
5. **Dialogue-based support**: Use-case number five is a service management case. Interaction goes via a dialogue based app. Information provisioning to enlighten and possibly solve a particular situation involves processes of causal learning. The AI entity is regarded as a companion of a human support- and maintenance engineer. In the figure this is indicated with arrow 5.

4 Discussion & future work

We added arrow nr. 6 in the matrix to reflect actual discussions in public media about AI overtaking jobs or growing beyond human control, e.g. in the domain of Large Language Models [3]. Although these public opinions must be taken seriously, one might rather consider them in terms of blind usage by the end-users and intensify research on data sovereignty and trustworthiness. In a human-AI ecosystem AI entities are not there to overtake jobs, but rather to help. The need for AI in energy management, like in all other industries, is clear; In order to meet the energy transition goals humanity needs to team up with AI in order to deal with the complexity of future energy systems and the expected shortage in personnel for commissioning and maintaining those systems.

In future work we address the topic of trustworthy collaboration in human-AI ecosystems. An essential aspect is to allow AI entities to make mutual assump-

[3] futureoflife.org/open-letter/pause-giant-ai-experiments/

tions,intentions and expectations explicit such that they can be used in deliberation and communication to achieve shared goals and to resolve conflicts of interest. They can are either be 1) shared explicitly, 2) be based on expectations of average behaviour patterns or 3) observed, learned and anticipated from others' behaviour. We will study mechanisms that support anticipation and team orchestration. We also want to extend our work on co-learning and self organization in order to address the challenges on trust and resilience as a result from interaction within the network itself.

Although interactions can be explored in simulated environments where agents are represented as avatars, i.e. active digital twins of real objects, and humans participate either interactively or by modelling their (social) behaviour, critical real-life applications are the ultimate goal and proof of value. Therefore, we will continue our applied work in the field of energy management and study how new concepts can be of use in a real-life context, e.g. in the context of Urban Energy Regulation.

In early work of Camarinha et al. [3], a collaborative network was defined as a network of enterprises (or individuals) which are supported by a computer network. With the rise of AI, those two types of networks are becoming more and more intertwined. Today, collaborative networks have evolved from a research discipline to practical applications across various fields. It is now time for the next step in the evolution of collaborative networks and move towards a symbiotic systems in which AI entities become part of the collaborative networks. We consider Human-AI ecosystems to be one of such a new type of collaborative networks.

5 Conclusion

The evolution of AI is not merely an evolution of algorithmic and technological power. It is also characterized by the increase of collaboration and mutual care. In this paper we presented a maturity model for agents that are part of human-AI ecosystems. The model allows one to classify AI entities as they act as companions and learn in a social context. We gave some examples and plotted various use-cases of the Talking Buildings project in this model. In parallel we emphasized the concept of human-AI ecosystems, where agents and humans take each other into account the intentions of each other and care for the environment that allows them to breathe.

Acknowledgements

This project is supported by the European research and innovation program TAILOR. (*www.tailor-network.eu*) under grant agreement nr. 952215.

References

[1] Akata, Z. et al. "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence". In: *Computer* 53.8 (Aug. 2020), pp. 18–28. ISSN: 0018-9162. DOI: 10.1109/MC.2020.2996587. URL: https://www.computer.org/csdl/magazine/co/2020/08/09153877/1lUB5gL2CnS (visited on 02/02/2022).

[2] Barredo Arrieta, A. et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103 (visited on 12/08/2022).

[3] Camarinha-Matos, L. M. "Collaborative networked organizations: Status and trends in manufacturing". In: *Annual Reviews in Control* 33.2 (2009), pp. 199–208. ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2009.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1367578809000558.

[4] Didden, J. B. H. C., Dang, Q.-V., and Adan, I. J. B. F. "Decentralized learning multi-agent system for online machine shop scheduling problem". In: *Journal of Manufacturing Systems* 67 (2023), pp. 338–360. ISSN: 0278-6125. DOI: https://doi.org/10.1016/j.jmsy.2023.02.004. URL: https://www.sciencedirect.com/science/article/pii/S0278612523000286.

[5] Diggelen, J. van, Jorritsma, W., and Vecht, B. van der. "Teaming up with information agents". In: *arXiv:2101.06133 [cs]* (Jan. 2021). arXiv: 2101.06133. URL: http://arxiv.org/abs/2101.06133 (visited on 03/28/2022).

[6] Dunin-Keplicz, B. M. and Verbrugge, R. *Teamwork in Multi-Agent Systems: A Formal Approach*. 1st. Wiley Publishing, 2010. ISBN: 978-0-470-69988-1.

[7] Meyer-Vitali, A., Mulder, W., and Boer, M. H. T. de. "Modular Design Patterns for Hybrid Actors". In: *Cooperative AI Workshop*. Vol. 2021. NeurIPS. arXiv: 2109.09331. Dec. 2021. URL: http://arxiv.org/abs/2109.09331 (visited on 11/17/2021).

[8] Miller, T. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *arXiv:1706.07269 [cs]* (Aug. 2018). arXiv: 1706.07269. URL: http://arxiv.org/abs/1706.07269 (visited on 11/27/2021).

[9] Mulder, W., Meijer, G. R., and Adriaans, P. W. "Collaborative Learning Agents Supporting Service Network Management". In: *Service-Oriented Computing: Agents, Semantics, and Engineering*. Ed. by Kowalczyk, R. et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 83–92. ISBN: 978-3-540-79968-9.

[10] Peeters, M. M. et al. "Hybrid collective intelligence in a human–AI society". In: *AI & SOCIETY* 36 (Mar. 2021). DOI: 10.1007/s00146-020-01005-y.

[11] Phaal, R. et al. "A framework for mapping industrial emergence". In: *Technological Forecasting and Social Change* 78.2 (2011), pp. 217–230. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2010.06.018. URL: https://www.sciencedirect.com/science/article/pii/S0040162510001393.

[12] Ramchurn, S. D., Stein, S., and Jennings, N. R. "Trustworthy human-AI partnerships". In: *iScience* 24.8 (Aug. 2021), p. 102891. ISSN: 2589-0042. DOI: 10.1016/j.isci.2021.102891. URL: https://www.sciencedirect.com/science/article/pii/S2589004221008592 (visited on 06/21/2022).

[13] Reyes, D., Dinh, J., and Salas, E. "What Makes a Good Team Leader?" In: *The Journal of Character & Leadership Development* 6 (July 2019), pp. 88–100.

[14] Ross, S. "The Economic Theory of Agency: The Principal's Problem". In: *American Economic Review* 63 (Feb. 1973), pp. 134–39.

[15] Schoonderwoerd, T. A. J. et al. "Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task". In: *International Journal of Human-Computer Studies* 164 (Aug. 2022), p. 102831. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2022.102831. URL: https://www.sciencedirect.com/science/article/pii/S107158192200060X (visited on 11/24/2022).

[16] Stijn, J. J. van et al. "Team design patterns for moral decisions in hybrid intelligent systems: 2021 AAAI Spring Symposium on Combining Machine Learning and Knowledge Engineering, AAAI-MAKE 2021". In: *AAAI-MAKE 2021 Combining Machine Learning and Knowledge Engineering*. CEUR Workshop Proceedings (Apr. 2021). Ed. by Martin, A. et al., pp. 1–12. URL: http://www.scopus.com/inward/record.url?scp=85104628466&partnerID=8YFLogxK (visited on 02/07/2023).

[17] Thiebes, S., Lins, S., and Sunyaev, A. "Trustworthy artificial intelligence". In: *Electronic Markets* 31.2 (June 2021), pp. 447–464. ISSN: 1422-8890. DOI: 10.1007/s12525-020-00441-4. URL: https://doi.org/10.1007/s12525-020-00441-4 (visited on 09/28/2022).

[18] Vecht, B. van der et al. "Influence-Based Autonomy Levels in Agent Decision-Making". In: *Coordination, Organizations, Institutions, and Norms in Agent Systems II*. Ed. by Noriega, P. et al. Lecture Notes in Computer Science. event-place: Berlin, Heidelberg. Springer, 2007, pp. 322–337. ISBN: 978-3-540-74459-7. DOI: 10.1007/978-3-540-74459-7_21.

[19] Waa, J. et al. "Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations". In: *Frontiers in Robotics and AI* 8 (May 2021). DOI: 10.3389/frobt.2021.640647.

[20] Weerd, H. de, Verbrugge, R., and Verheij, B. "Higher-order theory of mind is especially useful in unpredictable negotiations". In: *Autonomous Agents and Multi-Agent Systems* 36.2 (May 2022), p. 30. ISSN: 1573-7454. DOI: 10.1007/s10458-022-09558-6. URL: https://doi.org/10.1007/s10458-022-09558-6 (visited on 02/14/2023).

[21] Zhang, H., Bosch, J., and Olsson, H. H. "Real-time End-to-End Federated Learning: An Automotive Case Study". In: *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. July 2021, pp. 459–468. DOI: 10.1109/COMPSAC51774.2021.00070.