Causing Intended Effects in Collaborative Decision-Making

André Meyer-Vitali[1,*,†], Wico Mulder[2,†]

[1]DFKI, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2]TNO, Zernikelaan 14, NL-9747 AA Groningen, The Netherlands

Abstract

When humans and software agents collaborate on taking decisions together in hybrid teams, they typically share knowledge and goals based on their individual intentions. Goals can be modelled as the effects that are caused by events or actions taken. In order to decide and plan which actions to take, it is necessary to understand which actions or events cause the intended effects. In other words, we consider causal inferencing in a reverse way: instead of asking whether certain actions or events indeed cause corresponding effects, we consider establishing and using a causal model for determining the appropriate cause or causes, such that the causal chain results in the desired and intended outcomes. For example, your goal can be to arrive at a destination at a given time. By reasoning back which actions are required to get you there, piece by piece, a causal path can be constructed to determine the departure time and modes of traffic along the route. Due to shared intentions and causal models, humans and agents can mutually trust each other regarding their actions and outcomes.

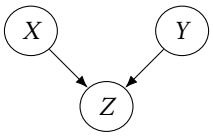
Keywords

causality, trust, human-centric AI, collaborative decision making, hybrid teams, agents, human-agent collaboration, theory of mind, search, planning, urban AI, sustainability

1. Introduction

The aim to empower humans by using systems with artificial intelligence, where the AI systems can be trusted [1, 2], leads to a dilemma: intelligent systems are characterised by a high degree of autonomy, which is required for delegating tasks to intelligently behaving AI systems (agency). However, we also want to control and understand autonomous agents in order to trust them. Similarly, for human-agent collaboration it is necessary that the parties understand each other. We want to use the complementary capabilities of individual humans and agents to improve hybrid decision-making. Therefore, we propose to use shared epistemic and causal models for achieving shared goals with a Theory of Mind (ToM) to resolve conflicts of interest. As a result,

HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26—27, 2023, Munich, Germany
*Corresponding author.
†These authors contributed equally.
andre.meyer-vitali@dfki.de (A. Meyer-Vitali); wico.mulder@tno.nl (W. Mulder)
https://www.dfki.de/en/web/about-us/employee/person/anme08 (A. Meyer-Vitali);
https://www.tno.nl/en/about-tno/our-people/wico-mulder/ (W. Mulder)
0000-0002-5242-1443 (A. Meyer-Vitali); 0000-0002-8607-0055 (W. Mulder)



(a) Structural Causal Model (SCM)

$$U = \{X, Y\}$$
$$V = \{Z\}$$
$$F = \{f_Z : Z = 2X + 3Y\}$$

(b) Structural Equation Model (SEM)

Figure 1: Causal Models

humans and agents can trust each other, even when they disagree in their goals and preferred solutions.

2. Motivation

When humans and agents share goals to collaborate in hybrid teams [3, 4, 5], they typically share knowledge based on their individual beliefs and intentions [6, 7]. Each human and agent has its own points of view (POV), reflected by the beliefs (individual "knowledge" and experience) and intentions that guide its behaviour. By negotiating about the individual intentions, they can formulate a shared goal that they want to achieve together by making use of each others' complementary capabilities and knowledge. A shared goal can be seen and modelled as a potential outcome, i.e. an effect, that is caused by one or more interventions (actions or events). Consequently, in order to decide and plan which actions to take, it is necessary to understand which actions or events cause the intended effects, related to the shared goal.

Structural Causal Models (SCM) or Structural Equation Models (SEM) [8] represent such understanding of casual relationships as graphs (SCM) and sets of equations (SEM), as shown in figure 1.

Both representations are useful for different purposes. SCMs support the human understanding and explanations, while SEMs are more suitable for representing causality in combination with logical expressions. The latter enables the possibility to communicate about causal models and intentions formally and concisely among agents.

Shared causal models increase trust among team members, because they help to explain to each other *why* certain actions are to be taken. They can explain the relationships between causes and effects. Delegation without reason or motivation is not trustworthy (unless the authority or reputation of the delegator is very high). This enables users to better understand the rationale and have greater confidence in others making a fair and unbiased decision.

There are several important aspects by which causality can improve the trustworthiness of AI systems (Causality for Trust, C4T). Besides precision and accuracy, which are fundamental to trustworthiness in AI, they are [9, 10, 11, 12]:

- Transparency & Interpretability. The reasoning behind decisions is explainable and easily understood by humans. Causal models provide the reasons for predictions and causal explanations help to build a correct mental model of the problem.

- Reproducibility. The ability to repeat experiments and get the same results increases the trustworthiness and accuracy of scientific output.
- Fairness. Causal AI can remove bias because it understands how variables are interconnected and dependent on each other. Understanding causal relationships between sensitive input variables (such as gender or race) and predicted outcomes is important for assessing biased behaviour. Counterfactual fairness is achieved when the output is identical for each sensitive input variable.
- Robustness. Causal models can avoid the brittleness of most machine learning systems, due to spurious correlations. They can handle data that is not independent and identically distributed (IID) or out of distribution (OOD), because they can discern between relevant and irrelevant data and variables [13, 14].
- Privacy. The robustness of causal models helps in preventing privacy attacks, because weaknesses of trained models cannot easily be exploited, for example in federated learning.
- Safety & Accountability (Auditing). Regulations for safe-guarding AI systems for use in critical applications and domains demand impact assessment (IA) to prevent from algorithmic and data-driven harm by finding potential negative effects before (large-scale) deployment. Causal models that represent dependencies between system design and impact can be used to assess and mitigate corresponding risks by identifying which system elements are responsible for undesired effects.

3. Concepts

Causal inference [15, 16] is typically concerned with the resulting effect when a corresponding event (cause) occurs, according to a given causal model, such that the respective dependency can be verified. Causal inference asks whether an event indeed causes a certain effect by determining the likelihood that one event was the cause of another. In contrast to statistical correlations, causal relationships are asymmetrical [17, 18, 19], i.e. that there is a directed relationship from a cause to an effect, rather than a spurious co-occurrence of events.

Causal discovery, on the other hand, is concerned with determining whether a change in one variable (representing a state, action or event) indeed causes a change in another, in order to distinguish between correlated and causal relationships. Approaches to make the distinction are interventions, random control trials and counterfactual reasoning.

Counterfactuals refer to alternative choices that could have been made ***in the past*** and the corresponding effects that they might have caused.

On the other hand, we want to find the causes (events or actions) that achieve a given effect. Therefore, we are concerned with *counterfactual exploration* ***for the future***. Starting from intended effects (such as individual or collaborative goals) we search for appropriate causes from which these effects will follow. If we know the causes for our intended effects, we can plan the actions that will lead to them. Thus, we are interested in establishing the causal chain that results in the desired and intended outcomes. Therefore, we are concerned with counterfactual exploration to answer the following questions. Which effects will result from different alternative choices for actions that we are going to take now? And which of those effects match with our goals and intentions?

4. Methodology

For collaborative decision-making (CDM), it is essential that each human and agent is aware of each others' points of view and understands that others possess mental states that might differ from one's own - which is known as a Theory of Mind (ToM). ToM is defined as the human cognitive ability to perceive and interpret others in terms of their mental states, such as beliefs, desires, goals, intentions and emotions, and it is considered an indispensable requirement of human social life [20, 21, 22, 23, 24].

We distinguish three different ways in which shared causal models involve a Theory of Mind. The intentions and causal models are either (1) shared explicitly, (2) observed and anticipated from others' behaviours or (3) based on expectations of average behaviour patterns.

Using option (1), each agent knows several causal models and other facts about its context (beliefs). These models can be shared among agents by negotiating about group goals and individual intentions [25, 26, 27, 5, 28, 29, 30, 31]. Then, the causal models can be used in decision-making according to a Theory of Mind. Consequently, groups of agents share a common goal to collaborate and learn causal models from each other and cause appropriate interventions to achieve their goals.

In causal inferencing, deliberate and controlled actions are called interventions. Interventional distributions are typically written as probabilities $P(Y|do(X = x_0))$, where Y is the desired effect and the *do*-operator represents the intervention of deliberately adjusting X to the value of x_0. Consequently, we need to find x_0 in a process that we call *counterfactual exploration*.

Counterfactual exploration searches backwards along the paths in a causal model to find the most probable outcome, as close as possible to Y, by setting X to likely values and estimating the resulting effects along the causal chain. Efficient causal paths can be found by searching to satisfy anticipated intentions and group goals. This process makes use of the backdoor criteria, as described in [32]. Hence, counterfactual exploration is a complementary method to either causal inference or causal discovery.

By using counterfactual exploration, the desired effects that were established in a hybrid human-agent team can be achieved with high probability and reliability. The use of a causal model guarantees that the outcomes are indeed causally related to the interventions and interpretable. Therefore, humans and agents can trust each other regarding the right actions to take.

5. Experiment: Talking Buildings

The above-mentioned concepts are explored in a scenario of optimising the energy consumption of users of a multi-tenant building. Besides finding sustainable ways of producing more energy (out of scope of this project), we must reduce our energy consumption and use the available energy as efficiently as possible. Therefore, it is crucial to create awareness and means of control at the points of storage and consumption.

The *Talking Buildings* project is an applied AI project that involves hybrid interaction in a social urban context. The project aims to find new forms of living, working and urban rhythms related to energy sustainability and flexibility [33, 34, 35, 36].

Our activities partly focus on operational climate management using real-life data from actual buildings. This involves activities related to situations inside the building, such as learning to control the climate system in an efficient way, while maintaining the level of comfort perceived by its users.

The other part of our work is the study of energy consumption patterns and interactions among buildings. This part involves the development of a computational model of interactions in urban energy regulation based on epistemic logic and Theory of Mind [27, 37, 38]. The setting requires agents to communicate not purely with facts, but also based on their beliefs and knowledge. The goal is to create an agent-based computational model of efficient knowledge transfer in a human-AI ecosystem.

6. Conclusions

As outlined above, we use our counterfactual exploration process to find efficient causal paths that achieve the shared intentions in hybrid teams. The causal relationships are explainable and transparent to each of the actors involved (both humans and agents), which leads to the emergence of mutual trust.

Acknowledgement

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence, Electronic Markets 31 (2021) 447–464. URL: https://doi.org/10.1007/s12525-020-00441-4. doi:10.1007/s12525-020-00441-4.

[2] S. D. Ramchurn, S. Stein, N. R. Jennings, Trustworthy human-AI partnerships, iScience 24 (2021) 102891. URL: https://www.sciencedirect.com/science/article/pii/S2589004221008592. doi:10.1016/j.isci.2021.102891.

[3] J. J. van Stijn, M. A. Neerincx, A. ten Teije, S. Vethman, Team design patterns for moral decisions in hybrid intelligent systems: 2021 AAAI Spring Symposium on Combining Machine Learning and Knowledge Engineering, AAAI-MAKE 2021, AAAI-MAKE 2021 Combining Machine Learning and Knowledge Engineering (2021) 1–12. URL: http://www.scopus.com/inward/record.url?scp=85104628466&partnerID=8YFLogxK, publisher: CEUR-WS.

[4] B. M. Dunin-Keplicz, R. Verbrugge, Teamwork in Multi-Agent Systems: A Formal Approach, 1st ed., Wiley Publishing, 2010. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470665237.

[5] A. Meyer-Vitali, W. Mulder, M. H. T. de Boer, Modular Design Patterns for Hybrid Actors, in: Cooperative AI Workshop, volume 2021 of *NeurIPS*, 2021. URL: http://arxiv.org/abs/2109.09331, arXiv: 2109.09331.

[6] A. S. Rao, M. P. George, BDI agents: From theory to practice, in: Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95), 1995, pp. 312–319. URL: http://www.agent.ai/doc/upload/200302/rao95.pdf.
[7] D. C. Dennett, The intentional stance, The intentional stance, The MIT Press, Cambridge, MA, US, 1987. Pages: xi, 388.
[8] J. Pearl, An Introduction to Causal Inference, The International Journal of Biostatistics 6 (2010). URL: https://www.degruyter.com/document/doi/10.2202/1557-4679.1203/html. doi:10.2202/1557-4679.1203, publisher: De Gruyter.
[9] B. Greifeneder, Three Ways A Causal Approach Can Improve Trust In AI, 2021. URL: https://www.forbes.com/sites/forbestechcouncil/2021/11/01/three-ways-a-causal-approach-can-improve-trust-in-ai/, section: Innovation.
[10] N. Ganguly, D. Fazlija, M. Badar, M. Fisichella, S. Sikdar, J. Schrader, J. Wallat, K. Rudra, M. Koubarakis, G. K. Patro, W. Z. E. Amri, W. Nejdl, A Review of the Role of Causality in Developing Trustworthy AI Systems, 2023. URL: http://arxiv.org/abs/2302.06975. doi:10.48550/arXiv.2302.06975, arXiv:2302.06975 [cs].
[11] B. Bartling, E. Fehr, D. Huffman, N. Netzer, The Causal Effect of Trust, 2018. URL: https://papers.ssrn.com/abstract=3286177. doi:10.2139/ssrn.3286177.
[12] J. Y. Yap, A. Tomlinson, A Causality-Based Model for Describing the Trustworthiness of a Computing Device, in: M. Yung, J. Zhang, Z. Yang (Eds.), Trusted Systems, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016, pp. 130–149. doi:10.1007/978-3-319-31550-8_9.
[13] E. Sherman, I. Shpitser, Identification and Estimation of Causal Effects from Dependent Data, 2019. URL: http://arxiv.org/abs/1902.01443. doi:10.48550/arXiv.1902.01443, arXiv:1902.01443 [stat].
[14] C. Zhang, K. Mohan, J. Pearl, Causal Inference under Interference and Model Uncertainty, 2023. URL: https://openreview.net/forum?id=TYKk9SWhke0.
[15] J. Pearl, Causality: Models, Reasoning and Inference, 2nd edition ed., Cambridge University Press, Cambridge, U.K. ; New York, 2009.
[16] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, 1st edition ed., Basic Books, New York, 2018.
[17] H. Price, Agency and Causal Asymmetry, Mind 101 (1992) 501–520. URL: https://www.jstor.org/stable/2253900, publisher: [Oxford University Press, Mind Association].
[18] D. Kutach, Causal Asymmetry, in: D. Kutach (Ed.), Causation and its Basis in Fundamental Physics, Oxford University Press, 2013, p. 0. URL: https://doi.org/10.1093/acprof:oso/9780199936205.003.0007. doi:10.1093/acprof:oso/9780199936205.003.0007.
[19] J. Ismael, Reflections on the asymmetry of causation, Interface Focus 13 (2023) 20220081. URL: https://royalsocietypublishing.org/doi/10.1098/rsfs.2022.0081. doi:10.1098/rsfs.2022.0081, publisher: Royal Society.
[20] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and Brain Sciences 1 (1978) 515–526. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0. doi:10.1017/S0140525X00076512, publisher: Cambridge University Press.
[21] S. Baron-Cohen, A. M. Leslie, U. Frith, Does the autistic child have a “theory of mind”

?, Cognition 21 (1985) 37–46. URL: https://www.sciencedirect.com/science/article/pii/0010027785900228. doi:10.1016/0010-0277(85)90022-8.
[22] C. Frith, U. Frith, Theory of mind, Current Biology 15 (2005) R644–R645. URL: https://www.cell.com/current-biology/abstract/S0960-9822(05)00960-7. doi:10.1016/j.cub.2005.08.041, publisher: Elsevier.
[23] L. Byom, B. Mutlu, Theory of mind: mechanisms, methods, and new directions, Frontiers in Human Neuroscience 7 (2013). URL: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00413.
[24] M. C. Buehler, T. H. Weisswange, Theory of Mind based Communication for Human Agent Cooperation, in: 2020 IEEE International Conference on Human-Machine Systems (ICHMS), 2020, pp. 1–6. doi:10.1109/ICHMS49158.2020.9209472.
[25] H. de Weerd, R. Verbrugge, B. Verheij, Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information, Autonomous Agents and Multi-Agent Systems 31 (2017) 250–287. URL: https://doi.org/10.1007/s10458-015-9317-1. doi:10.1007/s10458-015-9317-1.
[26] D. Bang, C. D. Frith, Making better decisions in groups, Royal Society Open Science 4 (2017) 170193. URL: https://royalsocietypublishing.org/doi/10.1098/rsos.170193. doi:10.1098/rsos.170193, publisher: Royal Society.
[27] L. Dissing, T. Bolander, Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic, volume 2, 2020, pp. 1615–1621. URL: https://www.ijcai.org/proceedings/2020/224. doi:10.24963/ijcai.2020/224, iSSN: 1045-0823.
[28] H. de Weerd, R. Verbrugge, B. Verheij, Higher-order theory of mind is especially useful in unpredictable negotiations, Autonomous Agents and Multi-Agent Systems 36 (2022) 30. URL: https://doi.org/10.1007/s10458-022-09558-6. doi:10.1007/s10458-022-09558-6.
[29] E. Erdogan, F. Dignum, R. Verbrugge, P. Yolum, Abstracting Minds: Computational Theory of Mind for Human-Agent Collaboration, in: HHAI2022: Augmenting Human Intellect, IOS Press, 2022, pp. 199–211. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA220199. doi:10.3233/FAIA220199.
[30] C. Kennington, Understanding Intention for Machine Theory of Mind: a Position Paper, in: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 450–453. doi:10.1109/RO-MAN53752.2022.9900783, iSSN: 1944-9437.
[31] M. K. Ho, R. Saxe, F. Cushman, Planning with Theory of Mind, Trends in Cognitive Sciences 26 (2022) 959–971. URL: https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(22)00185-1. doi:10.1016/j.tics.2022.08.003, publisher: Elsevier.
[32] J. Pearl, The do-calculus revisited, in: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI'12, AUAI Press, Arlington, Virginia, USA, 2012, pp. 3–11.
[33] A. Nikitas, K. Michalakopoulou, E. T. Njoya, D. Karampatzakis, Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era, Sustainability 12 (2020) 2789. URL: https://www.mdpi.com/2071-1050/12/7/2789. doi:10.3390/su12072789, number: 7 Publisher: Multidisciplinary Digital Publishing Institute.
[34] F. Cugurullo, Urban Artificial Intelligence: From Automation to Autonomy in the Smart

City, Frontiers in Sustainable Cities 2 (2020). URL: https://www.frontiersin.org/articles/10.3389/frsc.2020.00038.

[35] A. Luusua, J. Ylipulli, M. Foth, A. Aurigi, Urban AI: understanding the emerging role of artificial intelligence in smart cities, AI & SOCIETY (2022). URL: https://doi.org/10.1007/s00146-022-01537-5. doi:10.1007/s00146-022-01537-5.

[36] C. Nevejan, P. Sefkatli, S. Cunningham, City Rhythm, 2018.

[37] R. Verbrugge, Testing and Training Theory of Mind for Hybrid Human-agent Environments, in: A. P. Rocha, L. Steels, H. J. v. d. Herik (Eds.), Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020, SCITEPRESS, 2020, p. 11. URL: https://vimeo.com/396473042.

[38] I. Tsoukalas, Theory of Mind: Towards an Evolutionary Theory, Evolutionary Psychological Science 4 (2018) 38–66. URL: https://doi.org/10.1007/s40806-017-0112-x. doi:10.1007/s40806-017-0112-x.