Using Perceptual and Cognitive Explanations for Enhanced Human-Agent Team Performance

Mark A. Neerincx[1,2(✉)], Jasper van der Waa[1], Frank Kaptein[2], and Jurriaan van Diggelen[1]

[1] TNO, Kampweg 55, 3769 DE Soesterberg, Netherlands
{mark.neerincx,jasper.vanderwaa,jurriaan.vandiggelen}@tno.nl
[2] Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, Netherlands
f.c.a.kaptein@tudelft.nl

Abstract. Most explainable AI (XAI) research projects focus on well-delineated topics, such as interpretability of machine learning outcomes, knowledge sharing in a multi-agent system or human trust in agent's performance. For the development of explanations in human-agent teams, a more integrative approach is needed. This paper proposes a perceptual-cognitive explanation (PeCoX) framework for the development of explanations that address both the perceptual and cognitive foundations of an agent's behavior, distinguishing between explanation generation, communication and reception. It is a generic framework (i.e., the core is domain-agnostic and the perceptual layer is model-agnostic), and being developed and tested in the domains of transport, health-care and defense. The perceptual level entails the provision of an Intuitive Confidence Measure and the identification of the "foil" in a contrastive explanation. The cognitive level entails the selection of the beliefs, goals and emotions for explanations. Ontology Design Patterns are being constructed for the reasoning and communication, whereas Interaction Design Patterns are being constructed for the shaping of the multimodal communication. First results show (1) positive effects on human's understanding of the perceptual and cognitive foundation of agent's behavior, and (2) the need for harmonizing the explanations to the context and human's information processing capabilities.

Keywords: Explainable AI · Human-agent teamwork · Cognitive engineering
Ontologies · Design patterns

1 Introduction

Advances in Artificial Intelligence (AI) and Information & Communication Technology (ICT) have been manifested in various automated systems, such as sensing technology, machine learning modules, Internet of Things, conversational agents and cognitive robotics. The embodiment in artificial, virtual or physical, agents enables automation to evolve as a member of mixed human-agent teams. A major challenge is to combine, automate and embody the information processes in such a way that agents really become full-fledged team-members, complementing and collaborating with the human team-members.

The coordination and collaboration in human-agent teamwork requires intelligent reactive and anticipatory behaviors of the agents. More specifically, they require a shared

D. Harris (Ed.): EPCE 2018, LNAI 10906, pp. 204–214, 2018.
https://doi.org/10.1007/978-3-319-91122-9_18

knowledge representation, methods to comply with policies and agreements for responsible teamwork, and the learning and effectuation of successful patterns of joint activities (e.g., [1, 11, 18, 25]). In addition, according to the fifth challenge of Klein et al. [17], the agent should be able to make pertinent aspects of their status and intentions obvious to their teammates. In our view, this means that there is a need for mutual human-agent exchange of the reasons and foundations of actions, and that the agent should provide explanations for adequate human understanding and appreciation (incl. trust) of its performances.

Symbolic AI, such as BDI-agents with built-in Beliefs, Desires and Intentions based on folk psychology, provides explicit opportunities for the generation of explanations that are understandable and useful for the human team-member. For example, there have been developed explanation methods for fire-fighting teams [10], and tactical air combat teams [12, 28]. However, intelligent agents will be embodying more-and-more sub-symbolic machine learning methods for which it is far from clear how to derive an explanation logic for the human-agent collaboration.

So, for the envisioned human-agent teamwork, we need to develop methods for explainable AI (XAI) for adequate human understanding and appreciation (including trust) of symbolic *and* sub-symbolic agent performance. Such explanation should support human-in-the-loop (co-)learning of the human-agent teams.

To meet this challenge, our research focuses on the development of complementary explanation methods for agents using a holistic approach which covers all three phases of an explanation (see Fig. 1).

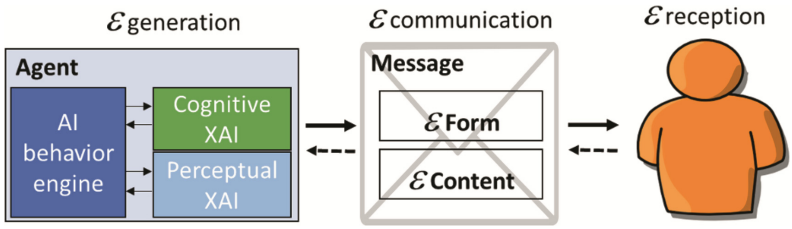

Fig. 1. Different phases of an explanation (ε refers to "the concept of explanation", cf. [29])

The first phase concerns *ε-generation*. In this phase, we distinguish between Perceptual XAI and Cognitive XAI. Perceptual XAI aims to explain the perceptual foundation of the agent behavior, and is usually connected to the sub-symbolic reasoning parts of an AI architecture. This paper discusses two types of perceptual XAI: The construction of an Intuitive Confidence Measure and the classification of facts and foils for contrastive explanations [30–32]. The second type is cognitive XAI, which can explain why a certain action was chosen (e.g., by relating them to goals or beliefs; [13–15]). This part interacts with the sub-symbolic part of the AI behavior engine to ground its belief base. The construction of goal-based deliberative explanations is an example of cognitive XAI.

The second phase in the explanation process is *ε-communication*. This process is characterized by the form of explanation (e.g. textual, or using dedicated images as proposed by Beller et al. [2]), and the content of the explanation (i.e. what is

communicated). We will propose the use of ontologies to standardize the language which is used to provide explanations [27, 29]. For the design of the form of the explanations (which need to be adaptive and interactive), we are developing interaction design patterns (cf., [23, 33, 34]).

The third phase of explanation, *ε-reception*, concerns how well the human understands the explanation. With respect to XAI reception, some user studies (e.g., [22]) have been conducted, but there is a significant lack of empirical research with actual human task performers who need explanations in realistic human-agent settings (e.g., [20, 21]). This paper summarizes some first results of our evaluations.

We are developing a general, perceptual-cognitive, explanation framework that can be applied and refined in different projects, crossing domains. In this way, we can provide technological progress and empirical grounding, building general models and methods for explanation that can be instantiated in different domains. First prototypes have been developed and tested with end-users in transportation, healthcare and defense domains. Results show that human needs and preferences for explanations depend on their individual characteristics (e.g., age and experience) and on the operational context. The paper will present an overview of the explanation framework and the first prototype designs and evaluation outcomes.

2 Perceptual XAI

We propose two methods for perceptual XAI: (1) The construction of an Intuitive Confidence Measure (ICM) and (2) the identification of the counterfactual reference (i.e., the foil that is set against the fact in a contrastive explanation).

2.1 Intuitive Confidence Measure

Waa et al. [30, 31] developed a generic confidence or certainty measure for agent's machine learning (ML) that can be understood by the human. It is model-agnostic, i.e., the measure can be used for any machine learning model as it depends solely on the input and output of a trained model and future feedback about that output. The confidence (or uncertainty) reflects machine learning model's expected performance on a single decision or classification. Our Intuitive Confidence Measure (ICM) should be easily understood by humans without ML-knowledge, and behave in a predictable way. We designed ICM to be intuitive by basing it on the notion of similarity and previous experiences: Previous experiences with the ML model's performance directly influence the confidence of a new output, and this influence is based on how similar those past data points are to the new data point (similarity can be represented as a distance in an n-dimensional space). The ICM is applicable to any (semi-)supervised machine learning model, that is either trained online or offline, by treating it as a black box. One of the use cases is an agent that predicts if its remote teammate is required to be at location in the near future (e.g. at a work-desk in a dynamic positioning ship). The agent can explain how likely it is that this single prediction (output) is correct [31].

We have applied the ICM explanation method to the domain of monitoring dynamic positioning systems, i.e. highly automatic systems which aim to maintain a vessel's position and heading using dedicated propellers and thrusters [6]. Occasionally, the system requires human intervention, for example to warn the user about potential problems that are predicted by a machine learning model. Because these predictions may be wrong, the user must have an appropriate level of trust in this type of advice. We used the intuitive confidence measure to provide this kind of advice. The message below is an example of this type of *ε-communication*, where the ICM measure in the advice is communicated to the user. Furthermore, the particular design pattern behind the message also allows extra information to be provided on request of the user (in the example on the type of *changes,* and the type of *conditions*) (Fig. 2).

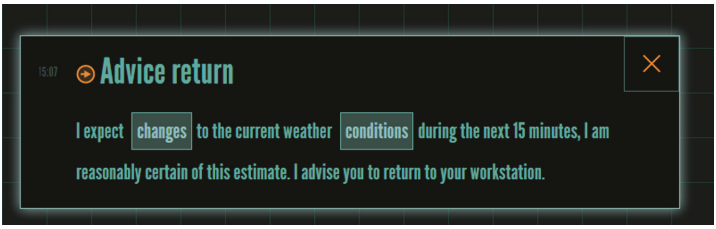

Fig. 2. Advice accompanied by the ICM XAI method.

To test the *ε-reception* of the ICM XAI method, we performed a user experiment [31], where we compared ICM with two different methods of computing confidence measures. The results confirmed the expectation that the intuitiveness of the XAI method is an important factor to consider when designing explainable smart systems. The user study showed large individual variations concerning user's interest in a confidence value and its utility in different situations. In comparison with other confidence measures, the ICM proves to be relatively easy to understand although not always preferred. The main reason for this was that the users, without ML-knowledge, attributed desired properties to other measures they did not fully understand. Whereas, with the ICM they understood the measure well enough to even identify its weaknesses; the users quickly thought to understand complex methods while, in fact, they did not (see [31], for more details of the user study).

2.2 Counterfactual Reference

For humans, explanations are most often contrastive, i.e., they refer to specific counterfactual cases, called foils [20]. When asking why an agent made a specific prediction or decision, humans would like to be informed about the contrast against the prediction or decision. So, an explanation should compare agent's output to an alternative counterfactual output (i.e., the foil). In other words, a good explanation considers both the fact (output) and the foil (alternative output).

Therefore, the generic PeCoX method for automated explanation extraction *(*i.e., *ε-generation)* includes the identification of the foil for a contrastive explanation. This

method should be able to deal with the context-dependency by learning which foil matches with which input-output pair based on interaction feedback from the user (e.g., through one-shot learning in combination with high generalization). This method is an extension of past work that constructs an explanation with the help of a localized interpretable machine learning model, such as a decision tree, through error weighting based on the distance between the data point of interest and the rest [24]. This way, the explanations can become more focused with less redundant information as the 'why' question is answered in a more precise manner.

More specifically, the PeCoX foil identification method is inspired by the LIME method, which provides an algorithm that learns an interpretable model locally around the prediction and a method to explain models by presenting representative individual predictions [24]. Further, like the Intuitive Confidence Measure of Sect. 2.1, LIME is model-agnostic. In our framework, it is further important to classify the foils for the generation of *adaptive* explanations (i.e., attuned to a specific user group).

The content and format of the explanations should be well-tailored to the user group. Ontology design patterns are constructed (see Sect. 4) that specify the high-level feature set, such that feature-based explanations are easy to comprehend by the user. Furthermore, interaction design patterns are generated that specify the corresponding multimodal dialogues for contrastive explanations.

We are applying the contrastive explanation method to the domain of type1 diabetes mellitus (T1DM). T1DM is a chronic condition where insufficient insulin is produced by the pancreas, affecting blood glucose levels. Daily self-management is needed for a balanced glucose level, harmonizing insulin doses (via pen or pump), food intake (amount of carbohydrates) and activities (e.g. sport). High and low glucose levels may lead to a hypo or hyper, and on the long term to serious health complications. The *Personal Assistant for a healthy Lifestyle* (PAL[1]) project develops a system for children aged 8–14, their parents and health care professionals that advances child's diabetes self-management. The PAL system comprises an *Embodied Conversational Agent*

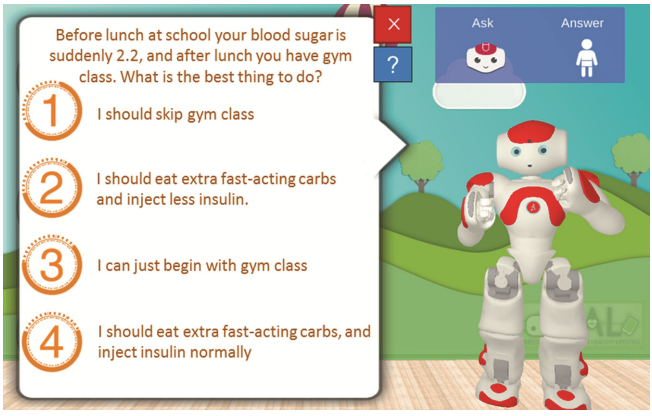

Fig. 3. Example of PAL quiz-question for learning to act on a low blood sugar level.

[1] http://pal4u.eu.

(ECA) robot and (mobile) avatar (Fig. 3), a set of *mobile health* (mHealth) applications (e.g., diabetes diary, educational quizzes), and *dashboards* for the caregivers (i.e., health care professionals and parents). All parts are interconnected with a shared knowledge-base and reasoning mechanism. health care professional, parent or child). Example contrastive explanations concerns PAL predictions that a child will not complete his or her self-management tasks. Such predictions should be explained to the child, parent and healthcare professional in different ways. Section 3 will provide more concrete explanation examples of the PAL system.

3 Cognitive XAI

In this part of the framework we consider explanations from the *intentional stance* [5]. When taking the intentional stance, you assume the action is a consequence of the intentions of the agent performing the action. You then explain the action by giving the reasons for the underlying intention. An explanation like such typically consists of beliefs, goals, and/or emotions [4, 5, 7, 9, 19]. For example, a support agent that tells its user to eat vegetables every day might provide the following explanation: 'I hope (emotion) that you will take my advice to eat vegetables every day because I want (goal) you to adopt a healthy lifestyle, and I think (belief) that you currently do not eat enough vegetables'. A method for explaining the reasons is being developed for a cognitive-affective (BDI-based) agent that updates its beliefs, goals and emotions based on events perceived in the environment [13, 14]. This agent can explain and justify its actions by communicating (a) the beliefs that underpin the actions, (b) the goals that inform the human of the agent's desired state when acting, and (c) the emotions that trigger or shape the actions.

3.1 Goal- and Belief-Based Explanations

The reasoning of a BDI based agent often consists of many beliefs and goals. If we use all of those for the explanation then this might overflow the user with information, which would thus not help us to make the behaviour intelligible [16]. Current work in XAI for artificial agents has thus focused on finding guidelines for which beliefs and goals are most valuable to use in an explanation towards end-users [3, 10]. This work confirms that a good explanation is short and thus contains few beliefs and goals. Similar to the evaluation of the Intuitive Confidence Measure in Sect. 2.1, a user study showed individual differences in explanation preferences. Particularly, adults showed a stronger preference than children for goals over beliefs in the explanations [14]. So, individual characteristics of the user must be taken into account when choosing which beliefs and goals should be selected as content for the explanation (Fig. 4).

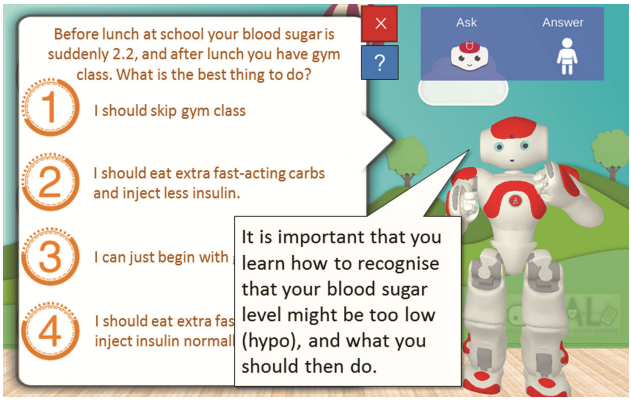

Fig. 4. Example of goal-based explanation in PAL.

3.2 Emotion-Based Explanations

Many artificial agents use computational modelling of emotion (based on cognitive appraisal theory) for their behaviour [15]. Here it is proposed that computational modelling of emotion can enhance XAI in several ways. The previous section mentioned the difficulty of selecting beliefs and goals as content for the explanation. The first use of emotions in XAI is as heuristic to identify the most important beliefs and goals for explanation. For example, the goal with the strongest influence for computing the desirability of an event can be used to explain the action the agent did after perceiving the event. The second use of emotions in XAI addresses that humans often use emotions when explaining behaviour [7]. The simulated emotions can be used to provide or enrich the action explanation (e.g., I was disappointed and therefore stopped my action). The third use addresses that emotions themselves can require to be explained (e.g., I was disappointed, because I had to do the same task over and over again) (Fig. 5).

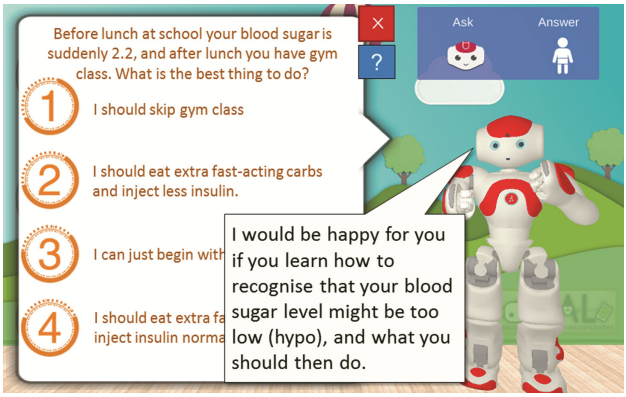

Fig. 5. Example of emotion-based explanation in PAL.

4 Ontology and Interaction Design Patterns

In our research and development projects, human-agent teams are modeled as joint cognitive systems, in which ontologies define the knowledge structures of these systems (i.e., the concepts, with their relationships, which underpin the cognitive processes of the teamwork). The explicit modeling of key concepts in explanations, enhances knowledge sharing (cf., [27]). These ontological models can be specified at different levels of generality [26]. The PeCoX top-level ontologies model general concepts that are domain independent, including explanation and human factors concepts (such as confidence and emotion, respectively). Specific domain and task aspects are captured in corresponding (lower level) ontological sub-models. For the construction of generic adaptive explanations, the Perceptual-Cognitive XAI framework is providing an extendable PeCoX-ontology in the form of *Ontology Design Patterns* (ODP, [8, 29]).

A formal definition of an explanation, like ODP, should specify its components, their roles and their interactions. Two ontologies provide starting points for such a definition. First, Su et al. [27] present an explanation ontology for constructing explanations for two agents that need to come to a partial shared understanding. They distinguish two knowledge types for agents that are engaged into an explanation process: (1) The *domain knowledge* entails representations of the to-be-explained concepts with the explanation parameters, and (2) the *explanation knowledge* entails the concepts that describe the explanation process and the permitted interactions in this process. Second, Tiddi et al. [29] studied approaches of Cognitive Sciences to model explanations with the instantiations in each of the analyzed disciplines. Their ontology intends to provide an abstract description which can be applicable to any context where an agent automatically produces explanations.

The ontologies of Su et al. [27] and Tiddi et al. [29] do insufficiently address (1) the perceptual XAI foundation of *ε-generation*, and (2) the situated human needs for explanation, i.e., the *ε-reception* (see Fig. 1). Section 2 describes the perceptual XAI concepts that are being specified an included in the PeCoX ontology. For the *ε-reception*, the social sciences provide theories and "models" for specific human information processes, for example, on the perception and activation of intentional and affective behaviors [20]. We are focusing and formalizing these "sub-models" into ODPs that support the predictions of *ε-reception* outcomes.

Whereas OPDs are used to support the design and reasoning of the communication processes and content of explanations, we use *Interaction Design Patterns* (IDP) for the specification of the form or shape of this communication [23, 33, 34].

5 Conclusions

This paper presented an integrative development approach for explainable AI in human-agent teams, addressing the perceptual and cognitive foundations of an agent's behavior during the explanation generation, communication and reception. The proposed PeCoX framework is being developed and tested in the domains of transport, health-care and defense. This framework distinguishes between perceptual and cognitive explanations.

On a technological side it provides tools for generating explanations from perceptual components (often implemented using sub-symbolic or connectionist approaches, such as neural nets), and cognitive components (often implemented using symbolic approaches, such as rule-based and ontology-based knowledge systems). On a human-side it provides tools for understanding the different types of explanatory support that a human would want in different contexts on both a cognitive as well as a perceptual level. Ontology Design Patterns are being constructed for the reasoning and communication of explanations, whereas Interaction Design Patterns are being constructed for the shaping of the adaptive, multimodal communications.

First results of the PeCoX framework development were acquired in the domains of transport, health-care and defense. At the *perceptual level*, first, the Intuitive Confidence Measure (ICM) proves to be a good candidate to enhance human's understanding of the perceptual and cognitive foundation of agent's behavior in a dynamic position ship. The ICM-evaluation showed the need to attune the explanations to the context and the biases of human's perception of their own understanding. The participants in the evaluation of different confidence measures quickly thought to understand complex methods while, in fact, they did not. The explanation should be formulated in a "human-aware" way that minimizes the risk for such misbeliefs. Furthermore, the "foil" identification seems to be an effective method to generate desired contrastive explanations and situate them in a relevant context.

At the *cognitive level*, the goal, belief and emotion models provide the analytical knowledge foundation of the explanations. Similarly to the "ICM evaluations" at the perceptual level, the cognitive evaluations showed a need for further adaptation of explanations to the momentary context (e.g., age and role of the user). The user studies provide the empirical foundation of the explanation adaptation (i.e., the harmonization of belief-, goal- and emotion communication to the momentary user state and context).

It should be noted that, for effective and efficient explainable AI, a balance must be found between technological possibilities (which types of explanations can be provided by the underlying control logic of the autonomous system?), and user needs (given the current context, which type of explanations must be given for the human to establish an appropriate level of trust in the system?). This might involve making the technology better explainable, or accepting that the user cannot be explained in every aspect, and trying to mitigate possible negative consequences.

Acknowledgements. This research is supported by the European PAL project (Horizon2020 grant nr. 643783-RIA), and the TNO seed Early Research Program "Applied AI".

References

1. Bradshaw, J.M., et al.: From tools to teammates: joint activity in human-agent-robot teams. In: Kurosu, M. (ed.) 2009 Proceedings of the HCD 2009. LNCS, vol. 5619, pp. 935–944. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02806-9_107
2. Beller, J., Heesen, M., Vollrath, M.: Improving the driver–automation interaction: an approach using automation uncertainty. Hum. Factors **55**(6), 1130–1141 (2013)

3. Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do you get it? User-evaluated explainable BDI agents. In: Dix, J., Witteveen, C. (eds.) MATES 2010. LNCS (LNAI), vol. 6251, pp. 28–39. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16178-0_5
4. Churchland, P.M.: Folk psychology and the explanation of human behavior. In: Greenwood, J. (ed.) The Future of Folk Psychology: Intentionality and Cognitive Science. Cambridge University Press, Cambridge (1991)
5. Dennett, D.C.: Three kinds of intentional psychology. In: Healey, R. (ed.) Reduction, Time and Reality. Cambridge University Press, Cambridge (1981)
6. van Diggelen, J., van den Broek, H., Schraagen, J.M., van der Waa, J.: An intelligent operator support system for dynamic positioning. In: Fechtelkotter, P., Legatt, M. (eds.) AHFE 2017. AISC, vol. 599, pp. 48–59. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60204-2_6
7. Döring, S.A.: Explaining action by emotion. Philos. Q. **53**, 214–230 (2003)
8. Gangemi, A., Presutti, V.: Ontology design patterns. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. IHIS, pp. 221–243. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_10
9. Harbers, M., Broekens, J., van den Bosch, K., Meyer, J.J.: Guidelines for developing explainable cognitive models. In: Proceedings of ICCM, pp. 85–90, January 2010
10. Harbers, M., van den Bosch, K., Meyer, J.J.: Design and evaluation of explainable BDI agents. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 2, pp. 125–132. IEEE (2010)
11. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 303–312. ACM (2017)
12. Haynes, S.R., Cohen, M.A., Ritter, F.E.: Designs for explaining intelligent agents. Int. J. Hum Comput Stud. **67**(1), 90–110 (2009)
13. Kaptein, F., Broekens, J., Hindriks, K.V., Neerincx, M.: CAAF: a cognitive affective agent programming framework. In: Traum, D., Swartout, W., Khooshabeh, P., Kopp, S., Scherer, S., Leuski, A. (eds.) IVA 2016. LNCS (LNAI), vol. 10011, pp. 317–330. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47665-0_28
14. Kaptein, F., Broekens, D.J., Hindriks, K.V., Neerincx, M.A.: Personalised self-explanation by robots: the role of goals versus beliefs in robot-action explanation for children and adults. In: RO-MAN 2017 (2017)
15. Kaptein, F., Broekens, D.J., Hindriks, K.V., Neerincx, M.A.: The role of emotion in self-explanation by cognitive agents. In: DFAI Workshop at ACII 2017 (2017)
16. Keil, F.C.: Explanation and understanding. Annu. Rev. Psychol. **57**, 227–254 (2006)
17. Klein, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intell. Syst. **19**(6), 91–95 (2004)
18. Lohani, M., Stokes, C., Dashan, N., McCoy, M., Bailey, C.A., Rivers, S.E.: A framework for human-agent social systems: the role of non-technical factors in operation success. In: Savage-Knepshield, P., Chen, J. (eds.) Advances in Human Factors in Robots and Unmanned Systems. AISC, vol. 499, pp. 137–148. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-41959-6_12
19. Malle, B.F.: How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. MIT Press, Cambridge (2004)
20. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. arXiv preprint (2017). arXiv:1706.07269

21. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum. In: IJCAI 2017 Workshop on Explainable AI (XAI), p. 36 (2017)
22. Narayanan, M., Chen, E., He, J., Kim, B., Gersham, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv preprint (2018). arXiv:1802.00682v1
23. Neerincx, M.A., van Diggelen, J., van Breda, L.: Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains. In: Harris, D. (ed.) EPCE 2016. LNCS (LNAI), vol. 9736, pp. 211–220. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40030-3_22
24. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM, August 2016
25. Scheutz, M., DeLoach, S.A., Adams, J.A.: A framework for developing and using shared mental models in human-agent teams. J. Cogn. Eng. Decis. Making **11**, 203–224 (2017)
26. Staab, S., Studer, R. (eds.): Handbook on Ontologies. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-540-24750-0
27. Su, X., Matskin, M., Rao, J.: Implementing explanation ontology for agent system. In: 2003 Proceedings IEEE/WIC International Conference on Web Intelligence, WI 2003, pp. 330–336. IEEE (2003)
28. Taylor, G., Knudsen, K., Holt, L.S.: Explaining agent behavior. Ann Arbor **1001**, 48105 (2006)
29. Tiddi, I., d'Aquin, M., Motta, E.: An ontology design pattern to define explanations. In: Proceedings of the 8th International Conference on Knowledge Capture, 8 p. ACM (2015)
30. van der Waa, J., van Diggelen, J., Neerincx, M.A., Raaijmakers, S.: ICM: an intuitive, model independent and accurate certainty measure for machine learning. In: 10th International Conference on Agents and AI (ICAART 2018) (2018)
31. van der Waa, J., van Diggelen, J., Neerincx, M.A.: The design and validation of an intuitive certainty measure. In: Proceedings of IUI 2018 Workshop on Explainable Smart Systems (2018)
32. van der Waa, J., Robeer, M.J., van Diggelen, J., Brinkhuis, M.J.S., Neerincx, M.A.: Contrastive explanation for machine learning in adaptive learning (in preparation)
33. Wang, W.: Self-management support system for renal transplant patients: understanding adherence and acceptance. Ph.D. thesis. Delft University of Technology, The Netherlands (2017)
34. van Welie, M., van der Veer, G.C.: Pattern languages in interaction design: structure and organization. In: Proceedings of Interact 2003, 1–5 September, Zürich, Switzerland, pp. 527–534, IOS Press, Amsterdam (2003)