Designing Artificial Reasoners for Communication

Blue Sky Ideas Track

Emiliano Lorini
IRIT, CNRS, Toulouse University
Toulouse, France
Emiliano.Lorini@irit.fr

ABSTRACT

In order to endow a conversational agent with sophisticated social intelligence, machine learning (which is prominent in LLM-based systems like Chat-GPT) is not enough. Logic-based reasoning and decision-making is needed. We need formal languages as well as reasoning and planning algorithms based on them for modeling and endowing the agent with intentional communication, theory of mind, explanatory capability and norm compliance. We identify some requirements that such languages should satisfy as well as a number of challenges regarding their combination and their integration with machine learning methods.

KEYWORDS

Communication, intention, norms, explanation

ACM Reference Format:
Emiliano Lorini. 2024. Designing Artificial Reasoners for Communication : 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, 6 pages.

1 INTRODUCTION

We are living in an era of profound technological changes in which AI systems are becoming increasingly powerful and pervasive. It is the era of machine learning in which deep neural architectures and reinforcement learning models can be trained on huge amounts of data and, consequently, achieve performances and carry out tasks in a way that was unimaginable until a few years ago. It is the era of conversational agents based on Large Language Models (LLMs): without the need for explicit background knowledge, they are capable of conversation that is highly informative and fully understandable to humans. The ability of such systems to generate satisfactory answers to complex questions, based solely on the correlations learned from a massive dataset, is impressive. However, we are still far from having a statistical machine that learns how to reason generally and accurately so that it can effectively solve problems not encountered during the learning phase.

General problem solving relies, among other things, on an intelligent system's inferential, planning and decisional capabilities that allow it to creatively perform new tasks in a goal-directed manner. These capabilities cannot be fully acquired through statistical learning or reinforcement learning. Rather, they are hardwired in the system itself. This is true for both human and artificial agents. In the case of humans, they are the product of biological evolution which has effectively combined and harmonized them with the learning capabilities. In the case of artificial agents, we have nowadays a rich toolbox of logic-based models and automated reasoning procedures. They are the result of decades of research in the area of so-called symbolic AI. Notable examples are efficient solvers for propositional logic and more complex logics, formal verification methods, planning models and algorithms. It is one of the main challenges to integrate them with machine learning models, both at the theoretical and the algorithmic level.

The importance of this challenge is perfectly exemplified by a dialogue system such as Chat-GPT that, as emphasized by existing regulations of AI systems like the AI Act, is expected to be trustworthy. For such system to interact with a human not only effectively and informatively, but also in a reliable, socially appropriate and normatively irreprochable way, it must be augmented with sophisticated reasoning and decision-making capabilities. The latter include, among other things, the capacity i) to communicate in a goal-directed way; ii) to represent and reason about the interlocutor's cognitive state, also called a theory of the interlocutor's mind, and to tailor communication to it; iii) to provide satisfactory explanations to the interlocutor; iv) to take legal and moral norms into consideration and choose a communicative action or plan according to them. The first two capacities are essential aspects of intentional communication. We think that it is not possible to endow an agent with such capabilities without the help of formal logic. We need expressive languages for representing: the agent's goal in communication; the agent's model of the interlocutor's cognitive state; the explanations that the agent has to provide to the interlocutor; and the norms with which the agent is expected to comply during interaction. Moreover, we have to develop automated reasoning procedures and algorithms based on these languages to be implemented in the system. To make the system more flexible and adaptable to different contexts of interaction is key to integrate such languages and algorithms with machine learning models.

In this paper, we identify some requirements that such languages should satisfy (e.g., which expressiveness they should have, which concepts they should capture) and some solutions to automate them. Furthermore, we single out a number of challenges we deem crucial regarding their combination and their integration with machine learning methods. Our analysis is relevant for any AI application involving an artificial agent that must communicate with the humans and exhibit social intelligence. We use the generic term *conversational agent* to denote this category of agents. It includes not only

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).

text-based dialogue systems but also embodied agents with a multimodal component (e.g., verbal, facial and bodily expressions). We put aside the natural language processing (NLP) aspects involved in communicative interaction. We only focus on its social reasoning aspects. The choice of separating the reasoning aspects from the NLP aspects is perfectly in line with existing computational models of dialogue developed in the HMI domain [24].

The paper is organized in two main sections. Section 2 introduces the concept of intentional communication and clarifies how to formalize and automate it in a conversational agent by means of formal logic. In Section 3 we identify some challenges: how to learn the theory of the interlocutor's mind; how to combine intentional communication with norm compliance, one the hand, and with explanation, on the other hand. In Section 4 we conclude.

2 INTENTIONAL COMMUNICATION

Is a system like Chat-GPT based on a LLM able to communicate with its human interlocutor? If communicating just means exchanging information *meaningfully*, then the answer tends to be affirmative. Indeed, it has been successfully trained through reinforcement learning to perform the right dialogue move depending on the multiple states of conversation it might face. It exhibits good adherence to Grice's four maxims of conversation [35] by being as informative as is required, by providing information which is relevant to context of interaction and by perspicuously engaging in conversation. More generally, it has acquired a remarkable competence to exchange information in a meaningful and comprehensible way. But, if communicating means exchanging information *knowingly* and *purposively*, the answer is negative. A dialogue system like Chat-GPT has no goal guiding its decision-making and behavior. Furthermore, it has no understanding of the intention behind the interlocutor's utterance or expectation about the consequences of its actions on the interlocutor's mind. More generally, it cannot engage in intentional communication.

As designers of a dialogue system we would like it to be customizable: we would like to specify its goals in a top-down manner, depending on the functionalities it is expected to deliver. Moreover, we would like it to be capable to anticipate the potential consequences of its actions on the mind of the interlocutor so that it has control over these effects and can decide to promote them, if it appraises them as positive, or to prevent them, if it appraises them as negative. For example, we would like the system to be able to evaluate the satisfaction, positive emotions as well as the frustration, negative emotions and stress it could induce in the interlocutor and to take these aspects into consideration in its decision-making. These are the reasons why intentional communication is paramount.

2.1 Requirements

For an agent to communicate intentionally two conditions must be satisfied. First of all, it must have one or more goals, — also called perlocutionary goals in speech act theory [18, 68, 72]—, that motivate it to exchange information with the interlocutor. That is, it must decide to and consequently intend to inform the interlocutor about something *in order to* achieve such goals. The agent's intention to inform the interlocutor that a fact φ is true can be seen as the intention to see to it that the interlocutor believes that φ.

Secondly, it must have a representation of the interlocutor's actual cognitive state, including her actual cognitive attitudes (e.g., beliefs, desires, preferences, intentions) and emotions as well as a model of the causal relations between the interlocutor's cognitive attitudes, emotions and behavior. This is commonly called Theory of Mind (ToM) [32] and relies on the so-called intentional stance [23] through which we view and explain the behaviors of others and our own behaviors in terms of mental properties.

The perlocutionary goal of the communicating agent can be of various types. For example, it can be a persuasive goal aimed at inducing the interlocutor to form a certain belief (*persuasion-targeted intentional communication*) or an influencing goal aimed at inducing a certain behaviour in the interlocutor (*influence-targeted intentional communication*). It can be of explanatory type, aimed at helping the interlocutor to understand why a certain event took place or why a certain fact is true (*explanation-targeted intentional communication*). The following example illustrates persuasion-targeted intentional communication.

Example 2.1. Imagine Ann is interacting with her conversational agent Rob. Rob wants to convince Ann to adopt a more environmentally sustainable lifestyle by stopping driving to work. Rob's decision of what information to give Ann depends on its knowledge's of Ann's cognitive state and, in particular, of Ann's preference over the different outcomes (e.g., money, health, environment protection). If Rob thinks that Ann's main concern is money, it will decide to inform Ann that using the car is significantly more expensive than using the bike *in order to* convince her to stop using her car.

Speech act theory (SAT) distinguishes perlocutionary goals/acts from illocutionary ones such as assertions, requests and commands that are achieved/performed *in* saying something. According to SAT, an agent achieves its perlocutionary goal (e.g., to persuade or influence the interlocutor) by the performance of an illocutionary act (e.g., an assertion, a command). The hearer's recognition of the illocutionary force of the speaker's act is based on the hearer's inferential capability and on their sharing of a set of conventional rules [31]. Here we consider a broader notion of intentional communication that does not necessarily require the performance of an illocutionary act by the speaker or the recognition of an illocutionary force by the hearer. Moreover, following [12], we assume it does not necessarily require a language shared by them. It can be purely behavioral. When we base intentional communication on the notion of perlocutionary goal, it also becomes natural to distinguish cooperative communication, in which the participants' goals coincide, from strategic communication, in which they differ.

2.2 Logical Specification

To be able to formalize and automate the reasoning and decision-making of a conversational agent engaged in intentional communication with a human we need a formal language for specifying the agent's perlocutionary goal and theory of the human's mind. As pointed out above, the latter includes the agent's beliefs about the human's current cognitive attitudes and emotions as well as the agent's beliefs about the causal relations between the human's cognitive attitudes, emotions and behavior. Such causal relations can be specified in a top-down manner by grounding them either i) on weak rationality assumptions, or ii) on principles of well-established

psychological theories of human motivation and emotion. Examples of the former are the epistemic rationality assumption that a person will believe something if she believes to be a necessary consequence of what she learns to be true; or the practical rationality assumption that a person will not decide to perform an action if she believes it will produce some undesirable effect without producing any desirable effect. Examples of the latter are self-efficacy theory (SET) [7], regulatory focus theory (RFT) [40], and appraisal theories of emotion (ATEs) [29, 45, 60, 65]. For instance, SET stipulates that a person will not be motivated to engage in an activity unless she believes that she is *capable of* carrying it out. Thus, by exploiting its knowledge of SET, the conversational agent may try to convince its human interlocutor that she is capable of completing a certain task in order to get her to commit to carrying it out. According to ATEs, the emotion of a person is triggered by a specific pattern of cognitive attitudes (appraisal phase) and, after being triggered, it induces a specific behavioral response (action tendency) or cognitive reinterpretation of the situation (coping strategy). For example, a person is fearful to the extent that she believes a undesirable event will likely occur which gives her the urge to flee away from it. The agent could exploit this theory to influence its interlocutor. In particular, it could bypass the interlocutor's rationality by leveraging the causal chain "belief $\wedge$ desire $\overset{\text{cause}}{\Longrightarrow}$ emotion $\overset{\text{cause}}{\Longrightarrow}$ action" through communication to induce her to act in a certain way. For example, it could inform the person (e.g., there is an angry lion behind you!) to trigger a certain emotion (e.g., fear) and, consequently, to induce a certain behavioral response (e.g., escape).

Formal languages particularly suitable for this type of specification are those of logics of cognitive attitudes among which epistemic logics [27, 38, 46], logics of preference [74], their combinations [17, 48, 54] and logics of emotions [1, 21, 71] are some representative examples. Epistemic logic is particularly suited to modeling an agent's higher-order beliefs about another agent's beliefs and the notions of common belief and knowledge [14, 70]. The concept of common belief has been recently applied to deep reinforcement learning models used in games involving a ToM component [28]. Notions of epistemic planning based on epistemic logic [9, 10] or on some interesting fragments of it [20, 57, 58] have been proposed. More recently [22], the notion of epistemic planning was generalized to cognitive planning where the goal of the planning agent is not necessarily a belief state of the target agent but, more generally, a cognitive state. Applications of epistemic and cognitive planning to modeling dialogue can be found in [43, 50]. The idea of using planning in combination with a logic of cognitive attitudes or with a more general notion of information state for modeling dialogue was advocated much earlier in [2, 3, 16, 18, 19, 66, 73]. Recent advances in epistemic and cognitive planning have offered new languages, new algorithms as well as a better understanding of the complexity of plan-based dialogue modeling.

The epistemic/cognitive planning approach is well-suited for handling intentional communication: it can be used by the agent to calculate a communicative plan aimed at achieving its perlocutionary goal(s), given its beliefs about the interlocutor's mind. In [22], it is shown that a NP-complete logic of cognitive attitudes is sufficient to model communication between an artificial planning agent and a human. Moreover, in this simple scenario the cognitive planning problem is Σ_P^2-complete. Consequently, it is possible to successfully implement a planning model of intentional communication between an artificial agent and a human using a SAT solver. But the epistemic/cognitive planning approach also has limitations. First of all, it is difficult to access the interlocutor's cognitive attitudes unless she explicitly reveals them. Secondly, imperfect rationality, cognitive biases and personality traits of humans are hard to specify in a top-down way as they are idiosyncratic. Thus, it is essential to integrate the approach with machine learning to enable the agent to learn part of the theory of the interlocutor's mind, the first challenge we discuss in the next section.

3 CHALLENGES

We discuss three challenges we deem fundamental about the integration of logic-based intentional communication with machine learning, normative and explanatory reasoning.

3.1 Challenge I: Learning ToM

The first challenge is how to learn the theory of the human interlocutor's mind. One possible method is inductive logic programming (ILP) [56]. It has the advantage of being logic-based and therefore can be easily integrated with a logic-based model of intentional communication. By means of ILP the conversational agent could learn part of the human's stable beliefs and preferences, where 'stable' means that they do not change over the course of interaction.

Example 3.1. Rob could use ILP to learn the condition under which Ann prefers using her bike to using her car. It could rely on a series of observations about the past situations in which Ann decided to use the bike (resp. the car) and try to find the best explanation of these observations, assuming that Ann's preferences are stable and she is minimally rational so as to choose what is compatible with her preferences. For instance, Rob could learn that Ann conditionally prefers cycling to work to driving, *if it is not a rainy day and the outside temperature is not too high.*

Another method that could be exploited is reinforcement learning by supposing that i) a state in the Markov Decision Process (MDP) representing the interaction between the agent and the human is identified with a knowledge base of the agent, similarly to the notion of 'belief MDP' [42]; ii) the agent's goal determines its reward, ii) the information in the agent's knowledge base as well as its goal are expressed in a language suitable for representing cognitive attitudes, as detailed in Section 2.2. The agent receives a positive reward at a state of the MDP if from the information in its knowledge base it can deduce that the goal is achieved. A model-free algorithm such as Q-learning can then be exploited by the agent to learn the quality of an action executed at a state.

Example 3.2. Suppose Rob's goal is to motivate Ann to use her bike. Through Q-learning, Rob can learn the quality of an informative action depending on what it knows about Ann's cognitive state. For example, it can learn the quality of *informing Ann that* the outside temperature is not too high, when it only knows that Ann believes that it is not a rainy day.

Other approaches to learning ToM exist: based on neural network and meta-learning [62], on Bayesian ToM [5, 6] and bayesian inference [26], on inverse RL [59]. The problem is that they are not

based on logic and, consequently, their integration with a logic-based model of intentional communication seems very complex.[1]

3.2 Challenge II: Norm Compliance

A normative reasoner must have i) knowledge of legal and moral norms as well as ii) the capacity to take them into consideration in its decision-making and planning process and, consequently, to comply with them. This is fundamental for making the agent trustworthy and for aligning its behaviour with the user's expectations. In the case of a conversational agent such as Chat-GPT, examples of norms with which it is expected to comply with are the prohibition to spread unverified information, the prohibition to lie, deceive or manipulate. To obtain this functionality, we need a formal language to represent basic epistemic concepts (e.g., belief, knowledge) as well as the concept of agency (i.e., the fact that an agent causes a state of affairs to be true). Such expressiveness is necessary to be able to define concepts such as deception, manipulation and lie that are in the content of the norms that the agent is expected to comply with. For example, the prohibition to deceive can be expressed as the prohibition to make someone believe that a certain fact is true when we believe that the fact is false. Good starting points for developing such language are the action-based logic of persuasion presented in [11], the causal analysis of persuasion presented in [77], and the epistemic causal logic presented in [67].

The language used to specify the norm compliance module of the conversational agent should be combined with the language used to specify its planning module of intentional communication, as described in Section 2.2. The language resulting from this combination should be expressive enough to represent, at the same time, cognitive attitudes, the elements of the agent's theory of the interlocutor's mind, and the notion of agency, which is included in the norms' contents. The challenging task is to find the right trade-off between expressiveness and computability. Ideally, the language should remain implementable using a SAT solver or a QBF solver, to make the approach exploitable in practice.

Thanks to this combination, it will be possible to include norms in the conversational agent's planning process. The agent will look for an informative plan aimed at achieving its perlocutionary goal, while maximizing compliance with the norms. This is in line with the idea of logic-based planning and decision-making under ethical values and constraints studied in [33, 34, 47]. Another interesting direction to be explored is the integration of the RL-based solution discussed in Section 3.1 and the norm compliance module. The normative reasoning module can be used by the agent to identify and then discard those actions whose execution would lead to the violation of the normative specifications. This can be done either during the learning process or after learning when an optimal policy has to be computed through the learned Q-function. This is in line with the idea of shielding [4].

3.3 Challenge III: Explanatory Communication

The third challenge we discuss is that of enriching a logic-based model of intentional communication with explanatory reasoning. At a cognitive level, an explanation can be seen as a causal attribution, namely, as a belief (or knowledge) of the explainer about the actual cause of a given fact (i.e., a belief that a certain fact φ_1 is true *because of* another fact φ_2) [37, Chapter 3]. But explanation also has a communicative counterpart which is highlighted in social psychology [41], argumentation [15, 63, 76], dialogue models [13] and explainable AI [44, 55, 69]. As pointed out in Section 2.1, the perlocutionary goal of a conversational agent engaged in intentional communication could be explanatory, namely, the goal of letting the interlocutor know why a certain fact φ is true. We call *explanatory communication* this kind of intentional communication. To be able to model explanatory communication properly, we need to combine the language of cognitive attitudes and the cognitive planning algorithm based on it, as described in Section 2.2, with a language for specifying explanations. A reasonable assumption is that explanations are intrinsically causal. Under this assumption, a good starting point for the development of such language are logics and semantics for causal reasoning based on structural equation models [30, 36, 39] or on causal rules [8, 49]. An important distinction that the formal language should capture is between others- and self-explanation. In fact, ideally a conversational agent interacting with a human should be able to explain the cognitive state and the behavior of its interlocutor as well as its own cognitive state and behavior. The explanation of an agent's action, belief or intention makes usually reference to the *reasons* determining it. (See [51, 75, 78] for more details.) Therefore, while others-explanation relies on a theory of the interlocutor's mind, self-explanation presupposes a form of meta-cognition and introspection by the explainer.

Example 3.3. Suppose Ann asks Rob why it informed her that the outside temperature is not too high. Rob can sincerely report that it did that because *it wants* her to use the bike instead of the car. This is a self-explanation that requires Rob to introspect its perlocutionary goal and, more generally, its cognitive state.

Introspection, typical of humans, is the basis of reflection and conscious control over one's decisions and judgments. As shown by McCarthy [52], it can be captured using logic. But it lacks altogether in existing LLM-based dialogue systems. Endowing such systems with introspection is useful in making them capable of detecting their own hallucinatory states which are difficult to discriminate from veridical ones without the help of inference [25, 61].

4 CONCLUSION

Let's take stock. We have put intentional communication at the center of the design of conversational agents that are supposed to interact with humans. We hope we have convinced the reader of the need to use logic to model this concept appropriately. We have identified three challenges regarding the integration of intentional communication with learning, normative and explanatory reasoning. The ways a logic-based model of intentional communication could be combined, at the architectural level, with an LLM-based dialogue system are manifold. For instance, it could be used to generate prompts for the LLM in order to better control its output, or it could be coupled with the LLM in order to check norm compliance during conversation with the human interlocutor. These are non-trivial engineering problems that are beyond the scope of the present paper.

[1] There are also approaches to goal recognition based on plan observation [53, 64]. They can be expressed in logic but are limited as models for learning ToM. Indeed, apart from goals, they do not consider cognitive attitudes of the observed agent. Moreover, they assume it always executes an optimal plan which is not realistic for humans.

REFERENCES

[1] C. Adam, A. Herzig, and D. Longin. 2009. A logical formalization of the OCC theory of emotions. *Synthese* 168, 2 (2009), 201–248.

[2] J. F. Allen and C. R. Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence* 15, 3 (1980), 143–178.

[3] J. F. Allen, L. K. Schubert, G. Ferguson, P. A. Heeman, C. Hee Hwang, T. Kato, M. Light, N. G. Martin, B. W. Miller, M. Poesio, and D. R. Traum. 1995. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence* 7, 1 (1995), 7–48.

[4] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. AAAI Press, 2669–2678.

[5] C. Baker, R. Saxe, and J. Tenenbaum. 2011. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*. 2469–2474.

[6] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 1019–1038.

[7] A. Bandura. 1997. *Self-Efficacy: The Exercise of Control*. Freeman, New York.

[8] A. Bochman. 2003. A Logic For Causal Reasoning. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*. Morgan Kaufmann, 141–146.

[9] T. Bolander and M. B. Andersen. 2011. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics* 21, 1 (2011), 9–34.

[10] T. Bolander, M. Holm Jensen, and F. Schwarzentruber. 2015. Complexity Results in Epistemic Planning. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*. AAAI Press, 2791–2797.

[11] G. Bonnet, C. Leturc, E. Lorini, and G. Sartor. 2021. Influencing Choices by Changing Beliefs: A Logical Theory of Influence, Persuasion, and Deception. In *Proceedings of the Second International Workshop on Deceptive AI (DeceptAI 2021)*. Communications in Computer and Information Science (CCIS), Vol. 1296. Springer, 302–321.

[12] C. Castelfranchi, G. Pezzulo, and L. Tummolini. 2010. Behavioral Implicit Communication (BIC): Communicating with Smart Environments. *International Journal of Ambient Computing and Intelligence (IJACI)* 2, 1 (2010), 1–12.

[13] A. Cawsey. 1991. Generating interactive explanations. In *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91)*. AAAI Press, 86–91.

[14] H. H. Clark and C. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, A. Joshi, B. Webber, and I. Sag (Eds.). Cambridge University Press, 10–63.

[15] O. Cocarascu, A. Rago, and F. Toni. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*. ACM, 1261–1269.

[16] P. R. Cohen. 2020. Back to the Future for Dialogue Research. In *Proceedings of the The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. AAAI Press, 13514–13519.

[17] P. R. Cohen and H. J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42 (1990), 213–261.

[18] P. R. Cohen and H. J. Levesque. 1990. Rational interaction as the basis for communication. In *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.). MIT Press, 221–256.

[19] P. R. Cohen and C. R. Perrault. 1979. Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science* 3, 3 (1979), 177–212.

[20] M. C. Cooper, A. Herzig, F. Maffre, F. Maris, E. Perrotin, and P. Régnier. 2021. A lightweight epistemic logic and its application to planning. *Artificial Intelligence* 298 (2021), 103437.

[21] M. Dastani and E. Lorini. 2012. A logic of emotions: from appraisal to coping. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*. IFAAMAS, 1133–1140.

[22] J. L. Fernandez Davila, D. Longin, E. Lorini, and F. Maris. 2021. A Simple Framework for Cognitive Planning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*. AAAI Press, 6331–6339.

[23] D. C. Dennett. 1987. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts.

[24] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2021), 755–781.

[25] F. Dorsch. 2013. Experience and Introspection. In *Hallucination: Philosophy and Psychology*, F. Macpherson and D. Platchias (Eds.). MIT Press, 175–220.

[26] O. Evans, A. Stuhlmüller, and N. D. Goodman. 2016. Learning the Preferences of Ignorant, Inconsistent Agents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAA'16)*. AAAI Press, 323–329.

[27] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. 1995. *Reasoning about Knowledge*. MIT Press, Cambridge.

[28] J. N. Foerster, H. F. Song, E. Hughes, N. Burch, I. Dunning, S. Whiteson, M. M. Botvinick, and M. Bowling. 2019. Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, (ICML 2019) (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 1942–1951.

[29] M. Frijda. 1986. *The Emotions*. Cambridge University Press.

[30] D. Galles and J. Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Foundation of Science* 3, 1 (1998), 151–182.

[31] A. I. Goldman. 1970. *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs NJ.

[32] A. I. Goldman. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

[33] U. Grandi, E. Lorini, and T. Parker. 2023. Moral Planning Agents with LTL Values. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)*. ijcai.org, 418–426.

[34] U. Grandi, E. Lorini, T. Parker, and R. Alami. 2022. Logic-Based Ethical Planning. In *Proceedings of Advances in Artificial Intelligence - XXIst International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022) (LNCS, Vol. 13796)*. Springer, 198–211.

[35] H. P. Grice. 1989. *Studies in the way of words (3rd edition)*. Harvard University Press.

[36] J. Y. Halpern. 2000. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research* 12 (2000), 317–337.

[37] J. Y. Halpern. 2016. *Actual causality*. MIT Press.

[38] J. Y. Halpern and Y. Moses. 1992. A Guide to Completeness and Complexity for Modal Logics of Knowledge and Belief. *Artificial Intelligence* 54, 2 (1992), 319–379.

[39] J. Y. Halpern and J. Pearl. 2005. Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for Philosophy of Science* 56, 4 (2005), 843–887.

[40] T. E. Higgins. 1998. Beyond pleasure and pain. *American Psychologist* 52, 12 (1998), 1280–1300.

[41] D. J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65–81.

[42] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence* 101, 1-2 (1998), 99–134.

[43] F. Kominis and H. Geffner. 2017. Multiagent Online Planning with Nested Beliefs and Dialogue. In *Proceedings of the Twenty-Seventh International Conference on Automated Planning and Scheduling (ICAPS 2017)*. AAAI Press, 186–194.

[44] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, and S. Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. https://arxiv.org/abs/2202.01875

[45] R. S. Lazarus. 1991. *Emotion and adaptation*. Oxford University Press, New York.

[46] E. Lorini. 2020. Rethinking epistemic logic with belief bases. *Artificial Intelligence* 282 (2020).

[47] E. Lorini. 2021. A Logic of Evaluation. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. ACM, 827–835.

[48] E. Lorini. 2021. A Qualitative Theory of Cognitive Attitudes and their Change. *Theory and Practice of Logic Programming* 21, 4 (2021), 428–458.

[49] E. Lorini. 2023. A Rule-Based Modal View of Causal Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)*. ijcai.org, 3286–3295.

[50] E. Lorini, N. Sabouret, B. Ravenet, J. Fernandez Davila, and C. Clavel. 2022. Cognitive Planning in Motivational Interviewing. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)*. SCITEPRESS, 508–517.

[51] B. F. Malle. 2004. *How the mind explains behavior: folk explanations, meaning, and social interaction*. MIT Press.

[52] J. McCarthy. 1995. Making Robots Conscious of Their Mental States. In *Machine Intelligence 15, Intelligent Agents*. Oxford University Press, 3–17.

[53] F. Meneguzzi and R. Fraga Pereira. 2021. A Survey on Goal Recognition as Planning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*. ijcai.org, 4524–4532.

[54] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. 1999. A Logical Approach to the Dynamics of Commitments. *Artificial Intelligence* 113(1-2) (1999), 1–40.

[55] T. Miller. 2019. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence* 267, 1 (2019), 1–38.

[56] S. Muggleton and L. de Raedt. 1994. Inductive logic programming: theory and methods. *Journal of Logic Programming* 19-20 (1994), 629–679.

[57] C. Muise, V. Belle, P. Felli, S. A. McIlraith, T. Miller, A. R. Pearce, , and L. Sonenberg. 2021. Efficient Multi-agent Epistemic Planning: Teaching Planners About Nested Belief. *Artificial Intelligence* 302 (2021).

[58] C. Muise, V. Belle, P. Felli, S. A. McIlraith, T. Miller, A. R. Pearce, and L. Sonenberg. 2015. Planning over multi-agent epistemic states: A classical planning approach. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*. AAAI Press, 3327–3334.

[59] A. Y. Ng and S. Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. Morgan Kaufmann, 663–670.

[60] A. Ortony, G. L. Clore, and A. Collins. 1988. *The cognitive structure of emotions*. Cambridge University Press.
[61] C. Pelling. 2011. Characterizing hallucination epistemically. *Synthese* 178 (2011), 437–459.
[62] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018) (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 4215–4224.
[63] A. Rago, H. Li, and F. Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023)*. 582–592.
[64] M. Ramírez and H. Geffner. 2009. Plan Recognition as Planning. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*. 1778–1783.
[65] I. J. Roseman and A. A. Antoniou. 1996. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory. *Cognition and Emotion* 10 (1996), 241–277.
[66] M. D. Sadek, P. Bretier, and F. Panaget. 1997. ARTIMIS: Natural Dialogue Meets Rational Agency. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97)*. Morgan Kaufmann, 1030–1035.
[67] C. Sakama. 2021. Deception in Epistemic Causal Logic. In *Proceedings of the Second International Workshop on Deceptive AI (DeceptAI 2021)*. Communications in Computer and Information Science (CCIS), Vol. 1296. Springer, 105–123.
[68] J. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge.
[69] K. Sokol and P. A. Flach. 2020. One explanation does not fit all. *Künstliche Intelligenz* 34, 2 (2020), 235–250.
[70] R. Stalnaker. 2002. Common ground. *Linguistics and Philosophy* 25(5-6) (2002), 701–721.
[71] B. R. Steunebrink, M. Dastani, and J.-J. Ch. Meyer. 2012. A formal model of emotion triggers: an approach for BDI agents. *Synthese* 185, Supplement-1 (2012), 83–129.
[72] D. R. Traum. 1999. *Speech Acts for Dialogue Agents*. Springer Netherlands, Dordrecht, 169–201.
[73] D. R. Traum and S. Larsson. 2003. The Information State Approach to Dialogue Management. In *Current and New Directions in Discourse and Dialogue*, J. van Kuppevelt and R. W. Smith (Eds.). Springer, Dordrecht, 325–353.
[74] J. van Benthem and F. Liu. 2007. Dynamic logic of preference upgrade. *Journal of Applied Non Classical Logics* 17, 2 (2007), 157–182.
[75] G. H. Von Wright. 1971. *Explanation and understanding*. Routledge and Kegan Paul.
[76] D. Walton. 2004. A new dialectical theory of explanation. *Philosophical Explorations* 7, 1 (2004), 71–89.
[77] F. R. Ward, F. Toni, and F. Belardinelli. 2023. Defining Deception in Structural Causal Games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. ACM, 2902–2904.
[78] M. Winikoff, G. Sidorenko, V. Dignum, and F. Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artificial Intelligence* 300 (2021).