Responsible Agent Deliberation

DOUGLAS WALTON
Centre for Research in Argumentation, Reasoning and Rhetoric,
University of Windsor
dwalton@uwindsor.ca

ALICE TONIOLO
School of Computer Science, University of St Andrews
a.toniolo@st-andrews.ac.uk

Our question is whether by adhering to the designed rules of today's deliberation dialogue protocols, intelligent rational agents can act fairly, transparently and responsibly. In this paper, we propose some reflections and guidelines on how deliberation dialogue should be held to these principles using norms to define protocols.

KEYWORDS: Deliberation dialogue, multiagent systems, responsibility

1. INTRODUCTION

In recent years, advances in the development of Artificial Intelligence systems have called for a reflection on principles that such systems should adhere to. Fairness, responsibility and transparency in decision making among others are essential in today's design of agent systems. Agents engaging in deliberation to make a decision or to solve a problem act on behalf of their users and the development of such systems should be held to these principles. In this paper, we reflect on the effect that current design choices in deliberation dialogue frameworks have towards fairness, responsibility and transparency of protocols for such a dialogue.

Formal dialogue protocols define how to move forward in a dialogue prescribing how an agent might respond to a particular statement, and when they are allowed to speak. This approach is commonly used in current deliberation systems, such as the McBurney, Hitchcock, and Parsons (MHP) model (2007). Commenting on these conversational policies among agents, Maudet et al (2002) suggest that they may or may not represent deliberations of the kind found in natural conversation. In a more general context, Shi et al (2010) show that even in a flexible protocol, undefined sequences or unexpected additional sub-sequences of speech acts occur in natural dialogue. Later protocols have suggested that additional elements are required to represent more natural deliberation (Walton et al, 2016). However, by observing instances of human dialogue, we note that the design of dialogue protocols rules on how the agents are required to act might have consequences on the information shared, on the decisions taken, and beyond, to affect the resulting actions.

Hence our question is whether by adhering to the designed rules of today's deliberation dialogue protocols, agents will behave fairly, transparently and responsibly. In this paper, we propose some reflections and guidelines on how deliberation dialogue should be held to these principles using norms to define protocols. Fairness requires that protocols are designed to not discriminate against agents. Our initial observations show that the turn-taking function may cause unfair behavior, and that this function is not typical of deliberation, even though it is typical of persuasion dialogue. In deliberation dialogue, dialogues need to be more transparent so all can see the reasons given supporting or attacking the various alternatives (Yu et al, 2018).

Responsibility of agents has been studied in many contexts, predominantly in social and ethical behavior. Here we reflect on the problem of omission of information and attribution of responsibility due to protocol prescriptions. Castelfranchi (2000) holds that agents will inevitably deceive each other, and one way is by making an agent ignore something crucial for them. We show that the dialogue protocol rules may cause agents to be unable to state crucial information about an action. An agent may then be held responsible later if that action causes serious negative consequences. Responsibility should be considered as one of the principles for protocol design. We conclude our paper with some desirable properties that deliberation dialogue protocols should adhere to, to achieve better fairness, responsibility and transparency in decision making.

For the purposes of this paper, we define an *intelligent autonomous agent* (IAA) as an entity minimally having the following five capabilities, following in broad outline the approach of (Wooldridge, 2009). First, an IAA has the capability for perception and for collecting information. Second, an IAA can foresee some (but not all) of the future consequences of its actions and can change its planned actions accordingly. Third, an IAA can communicate with other agents so they can act together. Fourth, it can be inferred from the speech acts of an IAA that it is committed to a proposition, an action or a goal. Fifth, an IAA has the capability to add or retract commitments from its commitment store.

2. AGENT DELIBERATION

Deliberation can mean a wide variety of things in natural language (pretty much any activity involving some kind of thought can be called deliberation), but in recent computer science it has been given a more precise meaning. McBurney et al. (2007) cite three characteristics that have been widely adopted. First, deliberation is concerned with actions rather than propositions (and so is different from inquiry). Second, there are no initial commitments on either side (and so it is different from persuasion). Third, deliberation is cooperative rather than adversarial. The object is to achieve consensus, rather than conversion (persuasion) or compromise (negotiation).

Below is a simple example of a deliberation dialogue adapted from Kok et al (2011).

> Ann: Where should we go for dinner?
> Bob: We should go to the Italian restaurant.
> Ann: Why?
> Bob: It serves very tasty pizza.
> Ann: But it is too expensive. We should go to the Japanese restaurant.
> Bob: Why?
> Ann: It's close to my place.
> Bob: But I have to go home early and the Japanese restaurant is too far.

So far the deliberation dialogue has reached an impasse. But suppose the dialogue continues when Ann offers some new information which gives rise to a new option.

> Ann: I noticed this new Greek restaurant on my way to work today, it is close to your place, and much cheaper.
> Bob: OK.

Ann has offered two arguments supporting this new option. Bob indicates that he is OK with the proposal she has made, and so the dialogue has reached a successful resolution. They can go ahead with this proposal.

What made the dialogue terminate with a good outcome based on the arguments and proposals put forward by both sides? The introduction of the new information that Bob intended to go home early guided the subsequent identification of a new option revising the initial issue, helping the parties to find a suitable agreement.

In order for agents to engage in this dialogue, an agent model requires a representation of plans, actions, commitments and goals. A model of arguments is then required for agents to construct instantiated arguments about plans and actions to put forward in the dialogue. Finally, a dialogue protocol must be defined to identify when one agent is allowed to speak and what arguments can be stated. At each turn, the agent will identify from the protocol the possible speech acts that can be used to respond to a previous speech act. These will include arguments that could be exchanged, identified according to plans, actions, and goals. A selection of the next move is then to be made among the set of potential answers available. Figure 1 shows the layered representation of the agent knowledge as adapted from Prakken and Sartor (2002).

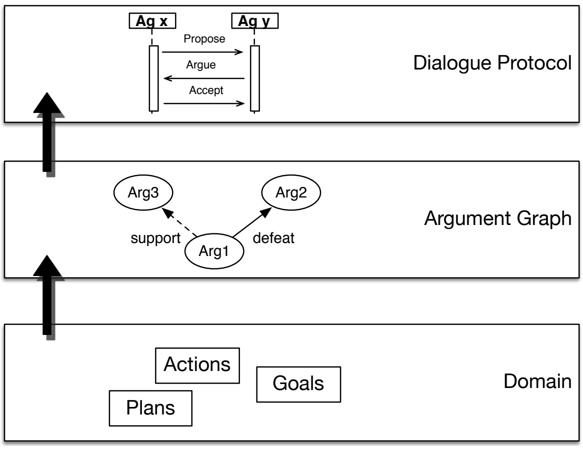

Figure 1. An agent layered dialogue model.

Arguments exchanged during the dialogue can be woven into an argumentation structure represented by an argument map (argument diagram). So we can use the familiar argumentation tools to evaluate the whole sequence of connected argumentation to get a big map showing all the supporting and attacking arguments for each of the proposals.

3. FORMAL DIALOGUE MODELS OF DELIBERATION

The seminal MHP model (McBurney, Hitchcock and Parsons, 2007) has three stages: an opening stage, an argumentation stage and a closing stage.[1] During the opening stage the issue is settled concerning the choice to be made. During the argumentation stage, there are four kinds of distinct intervals. During the first interval the agents seek for information concerning the circumstances of the case where the decision is to be made. During the second interval, the agents put forward proposals offering potential solutions to the problem that is to be solved in the deliberation dialogue. During the third interval, the agents consider and revise the proposal that has been put forward. During the fourth interval the agents recommend a particular proposal as the one best suited to solve the problem or to make the best decision based on the information that has been collected and assessed. The third stage of the dialogue is the closing stage where the agents reach agreement on what action to take, based on the evidence collected and the recommendations made during the argumentation stage.

An interesting problem with applying the MHP model to realistic cases of deliberation is that the knowledge bases that the agents have tend to be incomplete, and may need to be updated once new information comes in. For this reason Walton, Toniolo, and Norman (2016) proposed a model in which an open knowledge base enables information about changed circumstances to come in. During the argumentation stage there is a cyclical flow of argumentation as new knowledge comes that requires re-evaluation of proposals.

According to this revised model of deliberation dialogue, an additional feature is a knowledge base that is continually collecting new information about the circumstances as the agents are deliberating. In the Walton, et al. model, this information is used to continually update the knowledge base as new circumstances are retrieved. Naturally, as new knowledge comes in, this will affect the framing of the choice to be made, which may have to be updated as some options turn out to be unrealistic while others are supported by the new evidence. In the example dialogue in Section 2, the Ann's solution to the problem was based on new information that came in.

The argumentation stage of the revised model is comparable to that of the MHP model. In the first interval, where the agents find the circumstances of the decision to be made, new information continually streaming in from the updated knowledge base affects the other three intervals during the argumentation stage where proposals are put forward, revised and evaluated.

Based on this reconstruction of the argumentation stage, the revised model moves to a closing stage in which the best proposal is

[1] Subsequent models of deliberation dialogue include (Kok, Meyer, Prakken, & Vreeswijk, 2011), (Medellin-Gasque, Atkinson, McBurney, & Bench-Capon 2011) and (Walton, Toniolo, & Norman, 2016)

accepted as the course of action best suited to the findings carried out in the argumentation stage.

4. CONTROL OF AN INTELLIGENT AGENT

To move toward providing a framework defining moral responsibility in section 5, we introduce the technical term 'control' to stand for the capacity of an agent to act, as represented by the set of capabilities of an IAA defined in Section 1.

Control, in this sense of the term is *"the capacity to intervene in the course of events so as to be able both to make something happen and to preclude it from happening, this result being produced in a way that can be characterized as in some sense intended or planned or foreseen"* (Rescher, 1969, 329). On this view, a rational agent has control over its actions (or refraining from actions) of a sort that can change its circumstances. It can set goals for itself, direct its actions based on these goals, and can retract or modify its goals, for example if it sees that its goals conflict.

To extend the notion of control beyond the account of the capacities of an agent in Section 1, we can add seven further capabilities relating to goals and actions by expressing them in the language of control.

(1) An agent has control over carrying out actions (or refraining from actions) of a kind that can change its circumstances.
(2) An agent has goals, can set goals for itself, and direct its actions based on these goals.
(3) An agent can retract or modify its goals, as it might do if it sees that its goals conflict.
(4) An agent can grasp how actions to achieve a goal fall into an ordered sequence where some actions are required to carry out others.
(5) An agent can organize goals and actions into a hierarchy of levels of abstraction.
(6) An agent will generally keep trying to achieve a goal even if it has previously failed (plasticity), unless it has reasons to stop trying.
(7) An agent will not continue trying to carry out an action that it knows is impossible.

These capabilities can be formulated even more precisely by classifying the different types of control that can be distinguished. A classification system has been drawn up (Walton, 1974, 163), distinguishing six types of control: (1) complete positive control, (2) complete negative control, (3) positive partial control, (4) partial negative control, (5) full partial control, and (6) complete full control. The only kind of control not defined yet is that of full partial control. Full partial control can be explained by looking at Figure 2.

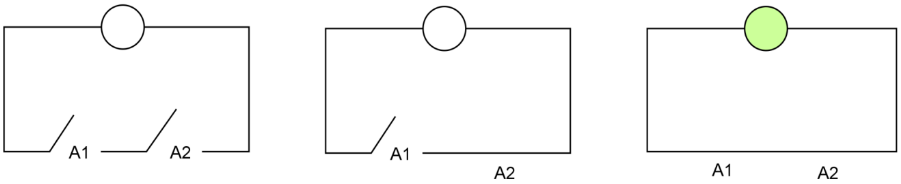

Figure 2: Series Circuit

Figure 2 represents the kind of case where agents A1 and A2 jointly, but not individually, have both positive and negative control over the outcome. An example of full partial control would be one where A1 and A2 are separately at the mercy of A3 with respect to the outcome, but where A1 and A2 can team up and get control of the outcome from A3. Individually, A1 and A2 are powerless to produce or prevent the outcome which is fully controlled by A3, yet jointly they can either produce or prevent the outcome. The existence of this type of control suggests the usefulness of modeling control as a teleological notion that needs to be defined within a framework of multiagent deliberation.

Next let us look at Figure 3. As shown by the two right circuits, A1 can keep the light on whether A2 turns her switch off or on. The only way for the light to be off is if both agents keep their switch in the open position, as shown in the left circuit.

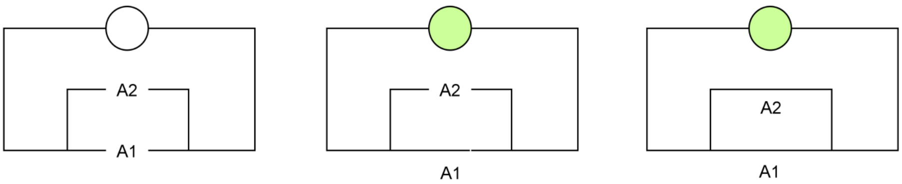

Figure 3: Parallel Circuit

Each agent, for example A1, can illuminate the light by closing his/her switch, thus exercising positive control over the outcome that the light is on, but A1 cannot make it so that the light is off unless A2 also decides to keep her switch open. Thus although A1 has positive control, he lacks full control because he lacks negative control. Neither agent individually has negative control. But each has positive control. Each has partial control. If they act together, they can exert positive control on whether the light is on or off (Walton, 1974, 164).

However, the requisite notion of control has not yet been defined completely enough to handle all problem cases. To conclude this section we set out two problems for defining the notion of control in this technical sense more fully.

The first example is posed by asking the question: does the moon control the tides (Rescher, 1969, 332)? Rescher answers that, in the proper meaning of the term 'control', it does not. Although he concedes that the movements of the moon determine the ebb and flow of the tides, he asserts that it is not proper to say that in the sense of the word 'control' he has in mind, the moon controls the tides. We agree with Rescher that in order to retain the intuitive idea of control, there has to be some aspect of deliberative agency or goal-directed action on the part of a controller present in the background.

The second example is the case of the berserk traffic light. Rescher argues this is not really an instance of control because *"it is not possible to retain the intuitive idea of control without retaining some aspect of deliberative agency or purpose contrivance on the part of the controller"* (Rescher, 1969, 332). He argues that the traffic light may still determine the flow of traffic, but the flow is no longer a controlled one, so in his (and our) sense of the word, the terminology of control has become inappropriate.

5. MORAL RESPONSIBILITY OF INTELLIGENT AGENTS

An IAA is only morally responsible for actions that actions it controls (carries out voluntarily, could have done otherwise). Such an agent must be autonomous (have self-control). Traditionally in philosophy this factor is called "free will" (a contested term. One way to reframe this notion so it can be made more precise for application to multiagent systems is to say that such an agent can control its actions). It is generally assumed in moral and legal philosophy that moral responsibility is "*the status of morally deserving praise, blame, reward, or punishment for an act or omission, in accordance with one's moral obligations*" (Eshleman, 2016, 1).

An intelligent rational ethical agent is an IAA that is committed to social (ethical) norms specifying that certain actions, or kinds of actions, are obligatory, permitted or forbidden in a group it is part of, in addition to the defining features of an IAA given above. This means that to have a formal deliberation system in which to frame ethical judgments about responsibility, deontic logic has to be brought in.

An open question, however, remains on how a deliberation model for an ethical IAA should be designed to yield a fair, responsible and transparent deliberation, particularly if that agent's role is to deliberate on behalf of a user in a team of agents or a mixed team of agents and users.

Desirable properties in our dialogue model include the ability to explain why a decision was taken, walking back through the dialogue exchange. Key information needs to be exchanged to identify a new option, as well as the selection of a new option. Reasons why a particular dialogue step was taken need to be allowed, according to the dialogue protocol formalization. We note that with the use of argumentation-based deliberation dialogue, dialogues are more transparent through reasons supporting various alternatives (Yu et al, 2018). Argumentation-based explainable AI (e.g. Fan and Toni, 2015) can be used to compute a set of arguments that form relevant explanations to the acceptability of an argument. Tintarev and Kutlak (2014) propose a system of dialogue to better understand the steps of a plan for example "Why does the system NOT say that I should do Y?" The user can ask why an option is rejected.

Consider a follow up to our example in Section 2. Ann and Bob agree that they will go to the Greek restaurant, but next they have to decide how to get there. Ann suggests that the fastest way to get there is to take the tube to the place, but in the end they decide to walk to the main square and then take the bus from there because the tube is too busy. Assume that Charlie joins the discussion later. If so, Charlie should be provided with an explanation on why they are not taking the fastest route.

The second desirable property is that of fairness, which requires that protocols are designed not to discriminate against agents. From one side, agents should be allowed to exchange actions and plans that better represent their interests and that of the group. Our focus however is to understand whether this is always possible given a specific protocol. By using a group turn-taking algorithm, agents can eventually voice all the proposals that they have available, provided that they can continue to discuss previously moved proposals, skip a turn, or advance new proposals.

However, we noticed that this function together with other constraints might prevent agents to exchange proposals or information leading to an unfair situation. For example, in Toniolo, Norman, and Sycara (2012), adopting components from Kok et al's (2011) dialogue framework, an agent can only make a relevant move in a dialogue. A relevant move is one that changes the acceptability status of a proposal, but this may prevent an agent from stating other proposals or further information. The dialogue protocol rules may then prevent agents from being able to state crucial information about an action because the statement no longer contributes to changing the acceptability of the proposal. However, this information may be essential to identify a different proposal. An agent may then be held responsible later if the action or plan chosen causes negative consequences. In our example, assume that Bob has also a different reason for not wanting to go to the Japanese restaurant: not only is it far, but it can only be reached on foot, and his knee is painful. Assume that Bob shares this second reason, instead of stating that he wants to go home early. Note that in this example, only one of these reasons would be considered relevant. Ann would not receive the critical information that the place needs to be close by. Hence the dialogue could take longer to explore the space of proposals with closer locations, and might end with a less favorable option or end with no acceptable option.

6. ELEMENTS OF MORAL RESPONSIBILITY

Aristotle (1969) suggested that knowledge is an element of moral responsibility (*Nichomachean Ethics*, 1110 b1 17) when he wrote that everything that is done by reason of ignorance is nonvoluntary. Aristotle argued that for a man to have acted voluntarily in the ethical sense of the term, he must know what he was doing when he acted (1110 b1 18). He also holds that an action can be a candidate for praise or blame only if it was voluntarily undertaken by its agent.

This approach suggests a way of modeling the inferential structure of the sequence of evidence-based argumentation used in legal and ethical cases to arrive at a conclusion about how to assign praise or blame. We argue that an agent acts voluntarily only if the action he carried out was under his control at the time.

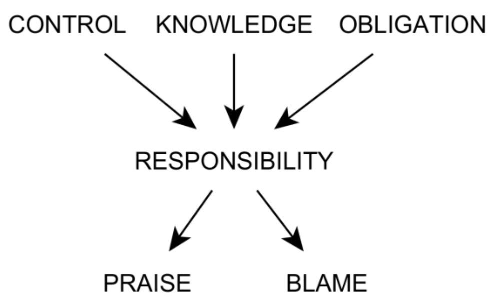

Figure 4: Factors for Arguing from Responsibility to Praise and Blame

On this approach, the three elements required to draw a conclusion about the responsibility of an IAA are control, knowledge and obligation, as shown in Figure 4. However, after some discussion at the

ECA conference in Groningen, we became convinced that in many cases, the notion of causation has to be factored in as well. Our initial reaction was that the concept of causation is too complex and not always needed for judging ethical responsibility. However, we accepted that a partial definition of causation could be used based on the INUS conditions.[2] On this approach, one event or action *A causes* another event or action *B* if and only if *A* is a member of a set of necessary conditions that, taken together, provide a sufficient condition for the occurrence of *B*. However, following the theory of causation in law of Hart and Honoré, such a selected event is generally a voluntary (human) action or an event or action that is "abnormal".

The next question is how to define the concept of a voluntary action. It is a contested concept and there is much literature on it in law, philosophy, and other fields. But H. L. A. Hart had a way around this. He saw remarkably (in 1949) that voluntariness is best defined in law as a defeasible concept (Hart, 1949, 180). That is, instead of seeing voluntariness as some elusive internal event or state in the human mind, he wrote that it should be defined by excluding a number of other concepts. This means that it serves to *exclude* a heterogeneous range of cases such as physical compulsion, coercion by threats, accidents mistakes, etc. In today's terms, he saw it as a defeasible concept. This insight anticipates the later AI view that case-based reasoning of the kind used in ethical and legal reasoning is inherently defeasible.

---

[2] INUS conditions are **i**nsufficient but non-redundant parts of a condition which is itself unnecessary but **s**ufficient for the occurrence of the effect according to the account of (Mackie, 1974).

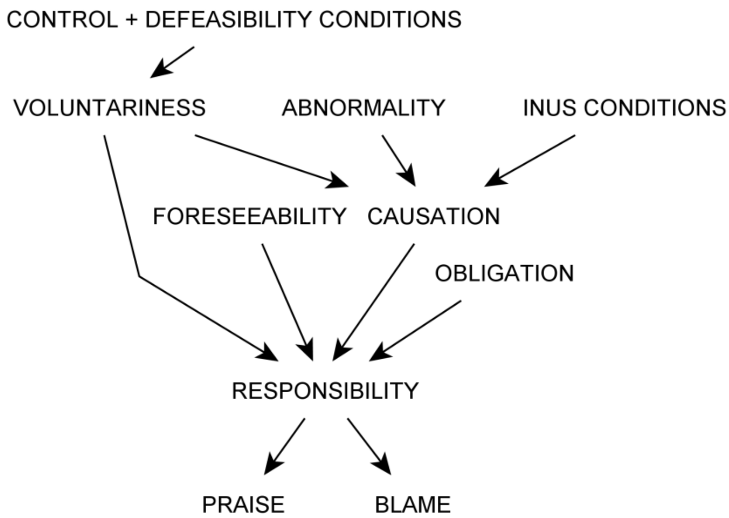

Figure 5: Factors for Arguing from Responsibility to Praise and Blame

What advantage can we derive from the insights of these early British analytical philosophers? To accommodate them we could modify the attempted defining conditions on how responsibility should be arrived at shown in Figure 5 as follows.

As indicated in Figure 5, we distinguish four basic components of responsibility. These are voluntariness, foreseeability, causation and obligation. We define causation in an admittedly simplistic way by using the INUS conditions, leaving the concept open to further refinement. Although causation is not always required to be considered, it is important in some cases, and therefore useful to include.

Foreseeability fits in well with the account of capabilities of an IAA listed in Section 1. In such systems, a rational agent has only a bounded rationality: it is aware of some (but not all) of the consequences of the actions it carries out or is contemplating carrying out. How foreseeable such a rational agent is expected to be in legal and ethical settings is variable. It is circumscribed by the granularity of the common sense reasoning that needs to be applied to the given situation in which the agent is situated. Our proposal then is that when the notion of ethical responsibility is redefined in this way, it would integrate formal models of deliberation with requirements for defining the notion of responsibility in a manner suitable for use in artificial intelligence.

Responsibility for omissions should be considered as one of the principles for protocol design based on the scheme for blame for omissions shown below.

Argumentation Scheme for Blame for Omissions

Major Premise: Agent A1 failed to carry out action S1.
Minor Premise 1: A had control over carrying out action S1.
Minor Premise 2: A1's failure to carry out S1 had negative consequences (NC).
Conclusion: A is to blame for NC.

Critical Questions

CQ1: What kind of control was involved, such as full control or partial control?
CQ2: Did A1 have knowledge about the reasonable likelihood of NC?
CQ3: Did A1 have an obligation to carry out S1 or otherwise to prevent NC?
CQ4: Did A1's carrying out S1 run into conflict with some other obligation of A1?
CQ5: Could NC have been prevented by other agents who were involved?

If we consider this scheme in the context of an agent deciding which argument or proposal to move forward in the deliberation, an agent may fail to inform another agent about an action T1 that could have been carried out at a particular time. If we assume that this lack of information leads to carry out T2 instead, which is revealed to yield negative consequences, the scheme above can be used to reason about whether A1 is to be blamed for these consequences.

7. CONCLUSIONS

We have presented a typology of deliberation dialogue that can be used to model ethical and legal responsibility in agent deliberation. Fairness, responsibility and transparency in AI decision making, among other properties, are essential in today's design of agent systems. We have shown that argument-based models of dialogue are useful for achieving more transparent decisions and that responsibility has many components that should be considered, including control and obligations. Protocols (and related constraints) should be further studied to consider how to block unfair moves in deliberation dialogues.

REFERENCES

Aristotle (1984). Nicomachean Ethics (trans. W. D. Ross, revised by J. O. Urmson). In Jonathan Barnes (ed.), *The Complete Works of Aristotle.* Princeton: Princeton University Press.

Castelfranchi, C. (2000). Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology, 2*(2), 113-119.

Eshleman, A. (2016). Moral Responsibility, *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2016/entries/moral-responsibility.

Fan, X., and Toni, F. (2015). On computing explanations in argumentation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Hart, H. L. A. (1949). The Ascription of Responsibility and Rights, *Proceedings of the Aristotelian Society, New Series, 49*(1948-1949), 171-194. Published by: Wiley on behalf of The Aristotelian Society. http://www.jstor.org/stable/4544455.

Hart, H. L. A. and Honoré, A. M. (1959). *Causation in the Law*. Oxford: Clarendon Press.

Kok, E. M., Meyer, J-J., Prakken, H. and Vreeswijk, G. (2011). A Formal Argumentation Framework for Deliberation Dialogues. In *Argumentation in Multi-Agent Systems*, ed. McBurney, P., Rahwan, I. and Parsons, 31–48 . Springer Berlin Heidelberg.

Maudet, M. and Chaib-draa, B. (2002). Commitment-based and Dialogue-based Protocols: New Trends in Agent Communication Languages: http://www2.ift.ulaval.ca/~chaib/publications/maudet-chaib-ker-revised2.pdf

Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation.* Oxford: Clarendon Press.

McBurney, P., Hitchcock, D., and Parsons, S. (2007). The Eightfold Way of Deliberation Dialogue. *International Journal of Intelligent Systems, 22*(1), 95–132.

Medellin-Gasque, R., Atkinson, K., McBurney, P., and Bench-Capon, T. J. M. (2011). Arguments over co-operative plans. In *Theory and Applications of Formal Argumentation, Lecture Notes in Computer Science, 7132,* 50–66. Springer Berlin Heidelberg.

Prakken, H. and Sartor, G. (2002). The role of logic in computational models of legal argument: a critical survey. In *Computational Logic: Logic Programming and Beyond, Lecture Notes in Computer Science, 2408*, 342–381. Springer Berlin Heidelberg.

Rescher, N. (1969). *The Concept of Control. Essays in Philosophical Analysis.* Pittsburgh. Pittsburgh University Press.

Shi, H., Ross, R., Tenbrink, T. and Bateman, j. (2010). Modelling illocutionary structure: Combining empirical studies with formal model analysis. In *Proceeding of the 11th International Conference in Computational Linguistics and Intelligent Text Processing*, 340-353.

Tintarev, N., and Kutlak, R. (2014). SAsSy: Making Decisions Transparent with Argumentation and Natural Language Generation. In *Proceedings of the International Conference on Intelligent User Interfaces*, 29-32.

Toniolo, A., Norman, T. J., and Sycara, K. (2012). An empirical study of argumentation schemes for deliberative dialogue. In *Proceedings of the Twentieth European Conference on Artificial Intelligence*, 242, 756–761.

Walton, D. (1974). *Control, Behaviorism, 2*(2), 162-171.

Walton, D., Toniolo, A. and Norman, T. J. (2016). Towards a Richer Model of Deliberation Dialogue: Closure Problem and Change of Circumstances, *Argument and Computation, 7*(2-3), 155-173.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems, 2nd ed.*, John Wiley & Sons.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., and Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 5527-5533.