Journal of Artificial Intelligence Research 44 (2012) 223-273 Submitted 11/11; published 05/12

Modeling Social Causality and Responsibility Judgment in Multi-Agent Interactions

Wenji Mao WENJI.MAO@IA.AC.CN
State Key laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
No.95 Zhongguancun East Road, Beijing 100190, China

Jonathan Gratch GRATCH@ICT.USC.EDU
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094, U.S.A.

Abstract

Social causality is the inference an entity makes about the social behavior of other entities and self. Besides physical cause and effect, social causality involves reasoning about epistemic states of agents and coercive circumstances. Based on such inference, responsibility judgment is the process whereby one singles out individuals to assign responsibility, credit or blame for multi-agent activities. Social causality and responsibility judgment are a key aspect of social intelligence, and a model for them facilitates the design and development of a variety of multi-agent interactive systems. Based on psychological attribution theory, this paper presents a domain-independent computational model to automate social inference and judgment process according to an agent's causal knowledge and observations of interaction. We conduct experimental studies to empirically validate the computational model. The experimental results show that our model predicts human judgments of social attributions and makes inferences consistent with what most people do in their judgments. Therefore, the proposed model can be generically incorporated into an intelligent system to augment its social and cognitive functionality.

1. Introduction

Recent years have seen an explosion of research at the intersection of computing and human social behavior. Topics such as human-centered (Jaimes, Sebe, & Gatica-Perez, 2006), social (Wang, Zeng, Carley, & Mao, 2007) and affective computing (Picard, 1997, 2010) emphasize the role of computers as partners or facilitators of human social activity, and highlight the challenge of computationally understanding and participating in human social interactions. Traditional artificial intelligence, with its emphasis on individual problem solving and reasoning of rational behavior, is not obviously suitable for the social, emotional, and human-like characteristics of social interaction. In this paper, we demonstrate how AI reasoning methods can be applied to understanding, modeling and predicting human social judgments, with applications in human-centric social interaction.

The specific challenge we focus on in this paper is reasoning about social causality. *Social causality* refers to the inference an entity makes about the social behavior of other entities and self. Such inference differs dramatically from how traditional artificial intelligence methods (e.g., planning) reason about physical reality. Besides physical cause and effect, social causality includes reasoning about mental states (e.g., did the actor intend to cause the outcome? could she foresee the outcome?) and social power (e.g., did the actor have the freedom to act or was she coerced by circumstances or other individuals?). *Responsibility judgment* is the process whereby one forms judgment results about responsibility, credit or blame based on the inference of social causality. Social causality and responsibility judgment underlie how we act on and make sense of the social world around us: they lead to emotional expressions of praise or rage; they justify public applause or prison terms. In short, they lie at the heart of social intelligence.

With the advance of multi-agent interactive systems, adaptive user interfaces and applications that socially interact with people, it is increasingly important to model and reason about this human-centric form of social intelligence. Social causal reasoning facilitates multi-agent planning by augmenting classical planners with the ability to reason about which entities have the power to effect changes. It facilitates adaptive learning by appraising praiseworthy or blameworthy behavior, and reinforcing the praiseworthy. In modeling the communicative and social behavior of human-like agents, responsibility judgment helps inform models of social emotions by characterizing which situations evoke anger, guilt or praise (Gratch, Mao, & Marsella, 2006). As people are usually adept at taking credit and deflecting blame in social dialogue (e.g., negotiation), the information helps guide natural language conversation strategies (Martinovski, Mao, Gratch, & Marsella, 2005).

Social causal inference helps reason about the social and cognitive states of an entity, and responsibility judgment helps form the assessment of the observed social behavior of an entity (either a human user, a computer program or an agent). They thus can facilitate various forms of interactions including human-computer, human-agent and agent-agent interactions. They can also facilitate human-human interaction by identifying the underlying cognitive process and principles of human judgments. In a multi-agent environment, social causality and responsibility judgment help share responsibility in multi-agent organization (Jennings, 1992), evaluate social power and dependence (Castelfranchi, 1990; Sichman, Conte, Demazeau, & Castelfranchi, 1994), automate after-action review for group training (Gratch & Mao, 2003; Johnson & Gonzalez, 2008), and support social simulation of agent society.

Our primary goal is to develop a faithful computational framework for human-like intelligent agents so as to drive realistic behavior modeling and generation (Swartout et al., 2006). Psychological and philosophical studies agree on the broad features people use in their everyday behavioral judgment. Our work is particularly influenced by *attribution theory*, a body of research in social psychology exploring folk explanation of behavior. Based on psychological attribution theory, we have developed a general computational framework for inferring social causality and forming responsibility judgment according to an agent's causal knowledge and observations of communication and task execution, and empirically validated our approach using human data.

The rest of this paper is organized as follows. In *Section 2*, we review previous computational work on social causality, responsibility and blame/credit. In *Section 3*, we introduce two influential attributional models of behavioral judgment, Weiner's (1995) model for responsibility judgment and Shaver's (1985) model for blame attribution. Based on these attributional models, *Section 4* presents our computational framework for social causality and responsibility judgment.

We provide the computational representation, inferences and algorithm in our proposed model, and illustrate our approach using an example from our system development. Then in *Section 5*, we report our empirical studies on model validation. *Section 6* further discusses some research issues. The paper concludes in *Section 7*.

2. Related Work

Since the rise of cognitive science (Newell & Simon, 1972), computational methods and metaphors have been applied to modeling and understanding human behavior. Several lines of research have addressed aspects of social cognition, including natural language dialogue (Cassell, Sullivan, Prevost, & Churchill, 2000; Ferguson & Allen, 2007), collaborative problem solving (Rich, Sidner, & Lesh, 2001; Schurr, Marecki, Tambe, & Scerri, 2005), modeling emotions (Marinier & Laird, 2004; Gratch, Marsella, & Petta, 2009), simulating human negotiation processes (Kraus, Hoz-Weiss, Wilkenfeld, 2008; Martinovski & Mao, 2009), and understanding human social networks (Golbeck & Hendler, 2006; Wang et al., 2010). When modeling human social behavior, it is useful to distinguish between *normative*, *descriptive* and *legal* perspectives. Normative models attempt to prescribe how people *should* assign responsibility and blame/credit. Descriptive models characterize what people do *in practice*, which may differ considerably from normative prescriptions. Legal models refer to the formalized processes society uses for responsibility assignment, which can be seen as the amalgam of normative and practical considerations. Before presenting our descriptive model of social causality and responsibility judgment, we motivate this work by examining each of these perspectives.

2.1 Normative Models

Normative (or prescriptive) models typically put forward a set of rational principles that should universally guide decision-making. For example, Bayesian decision theory is proposed as the optimal method for deciding between alternative courses of actions. Game theory is proposed as the ideal method for arriving at certain social decisions, such as whether or not to cooperate with another, possibly deceptive, party. While game theoretic approaches model group decision making itself in a rational way, social causality and responsibility judgment model the reasoning and assessment of social causes and consequences resulting from such decision making. For the judgment of causality, responsibility and blame/credit, research on normative models largely resides on moral philosophy where the aim is to identify rational principles to govern the assignment of social credit and blame. For example, Kant (1998) argued that, unlike what is often observed in practice, it would be rational to assign the same standards of responsibility regardless of the valence (i.e., praiseworthy or blameworthy) or severity of a social act. Within computer science and artificial intelligence, we are unaware of any other complete models based on the normative principles, with the exception of the computational model proposed by Chockler and Halpern (2004).

2.2 Legal Models

Legal models attempt to formalize responsibility judgment and inferences realized within judicial systems, typically with the aim of automating or verifying human legal judgments. This is a fertile research field at the intersection of artificial intelligence and law. The field has continuously been progressing since the development of early legal systems such as TAXMAN & TAXMAN-II (McCarty & Sridharan, 1981; McCarty, 1995), HYPO (Rissland & Ashley, 1987), CABARET (Rissland & Skalak, 1991) and CATO (Aleven & Ashley, 1995). There are similarities in the judgments of normative and legal responsibility, and some researchers have suggested using legal model as a direct analogue for normative model of responsibility judgment (e.g., Fincham & Jaspars, 1980). However, there are fundamental differences between these two kinds of responsibility judgment. Legal judgment largely depends on *specific* circumstances. That is why most legal reasoning systems are *case-based*, whereas evaluating moral responsibility identifies *general* theories that fall within the broad studies of cognitive functionalism[1] (e.g., clarifying the roles of cause, belief and intention in explaining behavior).

In addition to case-based legal reasoning systems, researchers have proposed *logic-based* approaches that focus on general reasoning mechanism, typically defeasible inference using non-monotonic reasoning and defeasible argumentation (e.g., Hage, 1997; Prakken, 1997). The main efforts in logic-based legal systems are on the representation of complex legal rules (e.g., contradictory, nonmonotonic and priority rules), inference with rules and exceptions, and handling conflict rules (Prakken & Sartor, 2002). McCarty (1997) argued whether in real cases, a judge would apply formal theory to evaluate complex rules, and thereby arrive at correct results. He called for a more intuitive version of legal rules, which would be "simple and clear". Furthermore, we argue that a layman's judgment of behavior in everyday situations is not quite the same as that made in the court. Not only does it occur in richer forms of social interaction, but it follows different set of rules.

2.3 Descriptive Models

Descriptive models attempt to characterize how people form social judgments *in practice*, which can differ from both the presumed normative principles and legal judgments. For example, in contrast to Kant's prescription to adopt uniform principles, people use different criteria when assigning blame versus credit and often form different judgments depending on the severity of an outcome. Descriptive models also differ in their criteria for validation. Whereas normative models are judged by their consistency with universal principles such as fairness and legal models are judged by their consistency with past legal decisions, descriptive models are assessed by their agreement with the judgments people form in their day-to-day lives. In this sense, descriptive models are most relevant to the field of human-centered or social computing, where the goal is to adapt computation to human norms of practice, rather than forcing humans to adapt to prescriptive norms of behavior. Research on descriptive models largely resides on social psychology (Heider, 1958; Shaver, 1985; Weiner 1995, 2001, 2006) and there is little work within artificial intelligence on attributing responsibility and blame/credit in a human-like fashion.

2.4 Computational Approaches

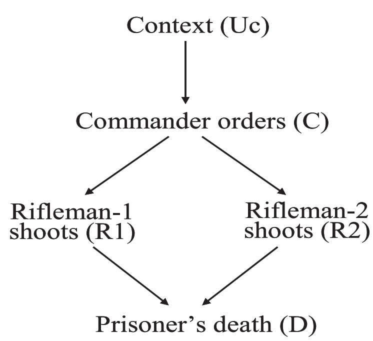
In AI and causality research, computational approaches were developed to address the problem by extending causal models (Halpern & Pearl, 2001; Chockler & Halpern, 2004). Halpern and Pearl (2001) presented a definition of *actual cause* within the framework of structural causal

[1] The doctrine that views theories of behavior as complex mental states, introduced and individualized by the functions or the roles they play in producing the behavior to be explained.

models. As their approach can extract more complex causal relationships from simple ones, their model is capable of inferring indirect causal factors including social cause. A *causal model* (or a structural model) is a system of equations over a set of random variables. There are two finite sets of variables: exogenous (U) and endogenous (V). The values of exogenous variables are determined by factors outside the model, thus they have no corresponding equations. Each endogenous variable has exactly one *causal equation* (or structural equation) that determines their value. A causal model can be expressed as a causal diagram, with nodes corresponding to the variables, and edges from the parents of each endogenous variable (indicated by the causal equations) to the endogenous variable. Take the *two-man firing squad* example (Pearl, 1999):

> There is a two-man firing squad; on their captain's order, both riflemen shoot simultaneously and accurately, and the prisoner dies.

Figure 1 illustrates the causal model for the firing squad example, where $U=\{Uc\}$ and $V=\{C, R1, R2, D\}$. A vector of values for the exogenous variables in U (called a context) in the causal model represents a specific situation (i.e., a causal world). For instance, if we assume $Uc=1$ (i.e., the captain's order is true) in the causal model below, then the resulting causal world describes the *two-man firing squad* story above. Causal inference is based on *counterfactual dependence* under some contingency. Roughly speaking, B is counterfactually dependent on A if, had A not happened then B would not have happened. For example, in the above firing squad scenario, given the context that the captain orders, under the contingency that *rifleman-2* did not shoot, the prisoner's death is counterfactually dependent on *rifleman-1*'s shooting. So *rifleman-1*'s shooting ($R1=1$) is an actual cause of the death. Similarly, *rifleman-2*'s shooting ($R2=1$) is an actual cause of the death. Besides the two riflemen who physically cause the death, Halpern & Pearl's model can find the captain's order ($C=1$) as an actual cause for the death as well.



Causal equations:

$Uc = C$

$C = R1$

$C = R2$

$R1 \vee R2 = D$

Figure 1: Causal Model for the Firing-Squad Example

Chockler and Halpern (2004) further extended this notion of causality, to account for *degree of responsibility*. They provide a definition of degree of responsibility based on the consideration of contingencies. Given a causal model M, a variable $X \in V$ and a context μ, the degree of responsibility of a formula $X=x$ for an outcome φ is measured by the minimal number of changes k that have to be made in μ in order to make φ counterfactually depend on $X=x$. If $X=x$ is not an actual cause of φ, then the degree of responsibility of $X=x$ for φ is 0; Otherwise the degree of responsibility of $X=x$ for φ is $1/(k+1)$. If φ counterfactually depends on $X=x$, then the degree of responsibility of $X=x$ for φ is 1. For example, if a person wins an election 11-0, then each voter who votes for her is a cause for the victory, and the degree of responsibility of each voter for the victory is *1/6*. However, in a *6-5* victory, the degree of responsibility of each voter is *1*.

Based on this notion of responsibility, Chockler and Halpern (2004) then defined the *degree of blame*, using the expected degree of responsibility weighed by the epistemic state of an agent. An agent's epistemic state is represented as a pair (K, Pr), where K is a situation with the form (M, μ) and Pr is a probability distribution over K. The degree of blame of $X=x$ for φ relative to an agent's epistemic state (K, Pr) is computed as the sum of multiplying the expected degree of responsibility of $X=x$ for φ in each possible situation in ($M_{X\leftarrow x}$, μ) and the agent's epistemic state of the probability of the situation. To illustrate this, they provide the *ten-man firing squad* example:

> There is a firing squad consisting of ten excellent marksmen. Only one of them has live bullets in his rifle; the rest have blanks. The marksmen do not know which of them has the live bullets. The marksmen shoot at the prisoner and he dies.

Suppose that an agent knows that exactly one marksman has live bullets in his rifle, and that all the marksmen will shoot. Then the agent considers *10* possible situations, depending on who has the bullets. Let $\{p_1, \ldots, p_{10}\}$ be the probability distribution over these situations, where p_i is the agent's prior probability that marksman-*i* has live bullets. Thus, according to the agent's epistemic state, the expected degree of responsibility of marksman-*1*'s shot for the death is *1* under the situation when he has the bullets (and *0* under other situations), and the degree of blame of marksman-*1*'s shot for the death is p_1.

Grounded on the *philosophical* principle (i.e., counterfactual reasoning), Chockler & Halpern's extended definition of responsibility accounts better for multiple causes and the extent to which each cause contributes to the occurrence of a specific outcome. Another advantage of their model is that their definition of degree of blame takes an agent's epistemic state into consideration. However, they only consider one epistemic variable, that is, an agent's knowledge prior to action performance. Important concepts in moral responsibility, such as intention and freedom of choice are excluded from their definition. As a result, their model uses one epistemic state as the only determinant for blame assignment, which is inconsistent with psychological theories.

As Chockler & Halpern's (2004) model is the extension of counterfactual reasoning within the structural-model framework, and structural-model approach represents all the events as random variables and causal information as equations over the random variables, there are several other limitations in their model. For instance, causal equations do not have direct correspondence in computational systems, so it is hard to obtain them for practical applications. As communicative events are also represented as random variables in their model (which are propositional), it is difficult to construct equations for communicative acts and infer intermediate beliefs (e.g., beliefs about desires, intentions, etc) that are important for social causal reasoning.

3. Attribution Theory for Behavioral Judgment

Most contemporary psychological studies of social causality and responsibility judgment draw on attribution theory (Heider, 1958). In over 50 years of research, attribution theory has progressed significantly and became a core area of social psychology (Malle, 2001; Weiner, 2006). Attribution research views that social perceivers make sense of the world by attributing behavior and events to their underlying causes. Attribution therefore refers to the process of ascribing a cause to an event or explaining the event, as well as the inferences or judgments made. Two influential attributional models for social causality, responsibility and blame (or credit) are those proposed by Shaver (1985) and Weiner (1995), which identify the underlying key factors (i.e., *attribution variables*) people use in behavioral judgment. Below we summarize their theories (we adopt the terminology of Shaver's model in this paper).

The assessments of *physical causality* and *coercion* identify the responsible party. Physical causality refers to the connection between events and the outcomes they produce, which includes personal causality (i.e., human agency) and impersonal causality (i.e., environmental factors). Only when human agency is involved, does an event become relevant to the investigation of responsibility and blame/credit. In the absence of coercion, the actor whose action directly produces the outcome is regarded as responsible. However, in the presence of coercion (as when some external force, such as a more powerful individual or a socially sanctioned authority, limits an agent's freedom of choice), some or all of the responsibility may be deflected to the coercive force. For example, in the two-man firing squad example, if the captain's order does limit the riflemen's freedom to avoid the prisoner's death, the captain should take some or all of the responsibility, depending on the degree of coercion.

Intention and *foreseeability* determine the degree of responsibility. Intention is generally conceived as the commitment to work towards a certain act or outcome. Most theories view intention as the major determinant of the degree of responsibility. Foreseeability refers to an agent's foreknowledge about actions and their effects. For example, although the riflemen foresaw that shooting a gun leads to the prisoner's death, they may not intend shooting and killing the prisoner. However, if an agent intends an action to achieve a certain outcome, then the agent must have the foreknowledge that the action brings about the outcome. The higher the degree of intention, the greater the responsibility assigned. If the riflemen have no intention of killing the prisoner, for instance, they should be assigned much less responsibility than in the case when they really intend so.

Weiner (2001) distinguished between act intentionality and outcome intent. An agent may intentionally perform an action, but may not intend all the action effects. For example, the riflemen may intentionally shoot the enemy, but may not intend the side effect of exposing themselves to the enemy force. It is *outcome intention* (i.e., intended action effect), rather than *act intention* (i.e., intended action) that are key in responsibility and behavioral judgment. Similar difference exists in *outcome coercion* (i.e., coerced action effect) and *act coercion* (i.e., coerced action). Furthermore, an agent's intentional action and action effect may fail. However, as long as it manifests intentions, a *failed attempt* can be blamed or credited almost the same as a successful one (Zimmerman, 1988).

The result of the judgment process is the assignment of certain blame or credit to the responsible party. Shaver's model of blame assignment follows a strict sequential process. In his model, first one assesses physical causality. If human agency is involved, the judgment process proceeds by assessing other key variables. Finally, the perceiver takes possible mitigating factors (i.e., justifications or excuses) into consideration and assigns proper blame to the responsible agent (mitigating factors are not modeled yet in our work). Weiner's model is similar, but it is more relaxed in that the sequential processing in Shaver's model is not presumed (we follow the implications of Weiner's model and relax the strict sequential feature in Shaver's model). The *intensity* of blame or credit is determined by the severity or positivity of the outcome as well as the degree of responsibility. The latter is based on the assessed values of attribution variables.

4. Proposed Computational Model

Attribution theory identifies the general process and key variables people use in judging social behavior. However, this process and the variables are not directly applicable to computational systems, as they are described at an abstract conceptual level that is insufficiently precise from a computational perspective. On the other hand, current intelligent systems are increasingly sophisticated, usually involving natural language communication, multi-agent interactions, goal-directed reasoning to generate and execute plans, and methods to explicitly model beliefs, desires and intentions of agents (Pollack, 1990; Grosz & Kraus, 1996; Gratch et al., 2006; Ferguson & Allen, 2007; Swartout et al., 2010).

To bridge the gap between conceptual descriptions of the theory and actual components in current intelligent systems, we need to develop the computational mechanisms that automatically convert the implications of the conceptual descriptions into a functionally workable model in use for intelligent systems. The computational model functions as the inferential mechanism to derive the conceptual variables in the theory from information and context available in practical systems. Ideally, the computational model should be based on the data structures and representations that are typically used in practical systems, and rely as little as possible on additional structural or representational features.

In constructing our computational model, we follow the basic dimensions in Shaver's model but relax its strict sequential feature. We follow the implications of Weiner's model, considering both the actions of agents and the outcomes they produce. We adopt plan representation used by most intelligent systems, especially in agent-based systems. This representation provides a concise description of the causal relationship between events and states. It also provides a clear structure for exploring alternative courses of actions, recognizing intentions, and assessing coercive situations and plan interventions.

We take advantage of artificial intelligence modeling and reasoning techniques, in particular, the *Belief-Desire-Intention model* (Bratman, 1987; Georgeff & Lansky, 1987) and *commonsense reasoning* (Gordon & Hobbs, 2004; Mueller, 2006). The BDI concepts help us map sometimes vague psychological terms into widely accepted concepts in AI and agent research, and research in commonsense reasoning informs the design of the inferential mechanism that generally operates on these conceptual representations. We use logic as a formal representation tool, focusing on the design of a small number of inference rules to capture the intuitions in people's judgments of social behavior[2].

[2] Note that our focus here is not the definition of a logical language, but rather, we aim at identifying the commonsense intuitions in people's behavioral judgment so as to come up with the computational modeling of social causality and responsibility attribution.

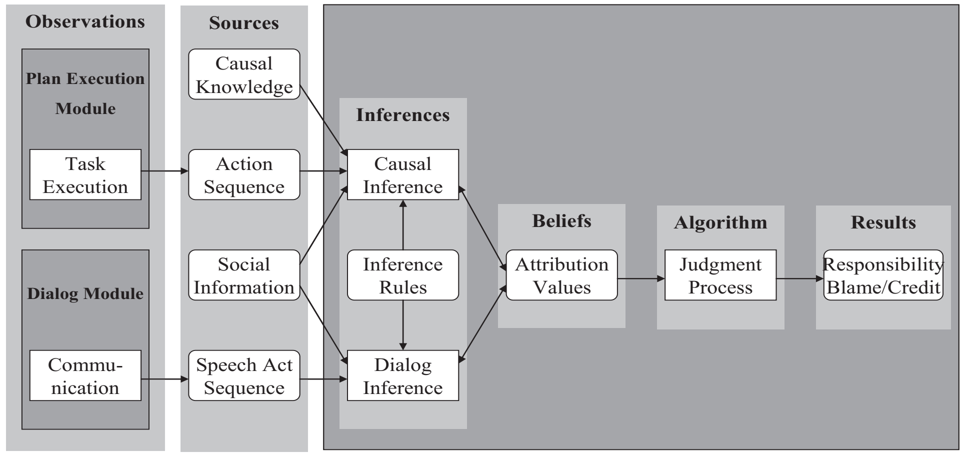


Figure 2: Overview of the Computational Model

We have developed a computational model that can automatically derive judgments underlying responsibility and blame attribution from knowledge and observations about social acts. *Figure 2* illustrates an overview of the computational model. Two sources of information contribute to the inference process. One source is the actions performed by the agents involved in the social situation (including physical acts and communicative acts). The other is the general causal knowledge about actions and states of the world (i.e., *causal knowledge*), social roles and power relationship of agents (i.e., *social information*). *Causal inference* derives beliefs from causal evidence. *Dialog inference* derives beliefs from communicative evidence. Both inferences make use of commonsense rules and generate beliefs of attribution variables. These beliefs serve as inputs for the judgment process, which is described as an *algorithm*. Finally, the algorithm forms an overall judgment and assigns proper credit or blame to the responsible agents.

4.1 Representations

Our computational representation is based on the plan descriptions that are widely applied to the applications and architecture design of intelligent systems (e.g., Georgeff & Lansky, 1987; Veloso et al., 1995; Fischer, Mueller, & Pischel, 1996; Rao, 1996; d'Inverno, Kinny, Luck, & Wooldridge, 1997; Huber, 1999; Gil, Deelman, Blythe, Kesselman, & Tangmurarunkit, 2004; Marsella & Gratch, 2009). More specifically, we adopt the classical STRIPS operators (Fikes & Nilsson, 1971) with the hierarchical plan representation (Erol, Hendler, & Nau, 1994; Nau, Cao, Lotem, & Muoz-Avila, 1999).

4.1.1 CAUSAL KNOWLEDGE

In our approach, causal knowledge is encoded via a hierarchical plan representation. An *action* has a set of propositional *preconditions* and *effects* (including *conditional effects*). Actions can be either *primitive* (i.e., directly executable by agents) or *abstract*. An abstract action may be decomposed in multiple ways and each decomposition is one *choice* of executing the action. Different choices of action execution are *alternatives* each other. If an abstract action can be decomposed in multiple ways, it is a *decision node* (i.e., *or node*) and an agent must decide amongst the alternatives. Otherwise, if an abstract action can only be decomposed in one way, it is a *non-decision node* (i.e., *and node*) and execution of the action is realized via executing all its *subactions*.

A *plan* is a set of actions to achieve certain intended *goal(s)*. As a plan may contain abstract actions (i.e., an abstract plan), decomposing the abstract actions into primitive ones in an abstract plan results in a set of primitive plans (i.e., plans composed of only primitive actions), which are directly executable by agents. Consequences or *outcomes* (we use them as exchangeable) are those desirable or undesirable action effects (i.e., effects having positive or negative significance to an agent). The desirability of action effects is represented by utility values (Blythe, 1999). To represent the hierarchical organizational structure of social agents, each action in a plan is associated with a *performer* (i.e., the agent capable of performing the action) and an agent who has *authority* over its execution. This is used to model the power relationships of agents.

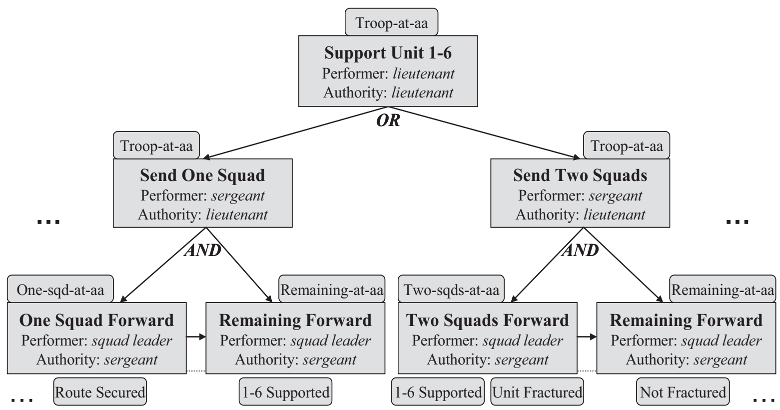


Figure 3: Partial Plan Representation for an Agent Team

Figure 3 illustrates an example of plan representation from a team training system we developed (we shall discuss more on this example in *Section 4.4*). In the example, a *lieutenant*, a *sergeant* and *squad leaders* work as a team in fulfilling the task of supporting a sister unit (i.e., unit *1-6*). The lieutenant is the leader of the troop. Two alternative ways are available to *support unit 1-6*, either *sending one squad* or *sending two squads*. Each alternative can be performed by the sergeant if authorized. The alternatives can be further decomposed into subsequent primitive actions that are directly executable by the squad leaders. Action execution brings about certain effects, for example, *two squads forward* (meaning that two of the four squads in the troop leave the scene) fractures the unit (meaning that the troop forces are split and weakened), which is undesirable to the troop. (Unit) *1-6 supported* (meaning that the sister unit is reinforced by the departing squads) is a desirable team goal.

4.1.2 Communicative Events

Communication between agents is a rich source of information for inferring social causality. We represent communicative events as a sequence of *speech acts* (Austin, 1962; Searle, 1969). For our purpose, we consider the speech acts commonly used in agent communication, and especially those that help infer dialogue agents' desires, intentions, foreknowledge and choices in acting. We thus focus on the acts *inform*, *request*, *order*, *accept*, *reject* and *counter-propose*.

4.1.3 ATTRIBUTION VARIABLES

Attributional models employ a set of key variables to determine social cause and responsibility. *Causality* refers to the relationship between cause and effect. For the investigation of responsibility attribution, the involvement of human agency is required (Weiner, 1995; Shaver, 1985). In our approach, we encode causal knowledge about actions (i.e., human agency) and the effects they produce via plan representation.

We consider both act intentionality and outcome intent of agents. Act intention is represented using *intend* and *do*, outcome intention using *intend* and *achieve*, and the connection between act and outcome intentions using *intend* and *by*. We use *know* and *bring about* to represent foreseeability. Two concepts are important in modeling *coercion*[3]. One concept is social *obligation*. The other is (*un*)*willingness*. For example, if some authorizing agent commands another agent to perform a certain action, then the latter agent has the obligation to do so. But if the latter agent is actually willing to, this is a voluntary act rather than a coercive one. We use *coerce* and *do* to represent act coercion and *coerce* and *achieve* for outcome coercion.

4.1.4 NOTATIONS

Now we provide the symbolic expressions of the notations used in our model[4].

Predicates

Let x and y be different agents, A and B be actions, e be an action effect, p and q be propositions, E be an effect set and t be a time. We adopt the following predicates in the model:

P1. *primitive*(A): A is a primitive action.
P2. *and-node*(A): action A is a non-decision node in plan structure.
P3. *or-node*(A): action A is a decision node in plan structure.
P4. *alternative*(A, B): actions A and B are alternatives of performing a higher-level action.
P5. *do*(x, A): agent x performs an action A.
P6. *achieve*(x, e): agent x achieves an effect e.
P7. *bring-about*(A, e): action A brings about an effect e.
P8. *by*(A, e): by acting A to achieve an effect e.
P9. *execute*(x, A, t): agent x executes an action A at time t.
P10. *occur*(e, t): effect e occurs at time t.
P11. *inform*(x, y, p, t): agent x informs agent y that p at time t.
P12. *request*(x, y, p, t): agent x requests agent y that p at time t.
P13. *order*(x, y, p, t): agent x orders agent y that p at time t.
P14. *accept*(x, p, t): agent x accepts that p at time t.
P15. *reject*(x, p, t): agent x rejects that p at time t.
P16. *counter-propose*(x, p, q, y, t): agent x counters that p and proposes that q to agent y at time t.
P17. *cause*(x, e, t): agent x causes an effect e at time t.

[3] Coercion sometimes means physical coercion, such as pushing someone's hand to pull the trigger of a gun. Here we mean psychological coercion, which emphasizes its impact on the psychological states of agents.
[4] Although we represent these notations in first-order predicate calculus, we treat them as semi-formal notations in our model and do not conduct theorem-proving type of inference with them in strict logical sense.

P18. *assist-cause*(x, y, e, t): agent x assists agent y in achieving an effect e at time t.
P19. *know*(x, p, t): agent x knows that p at time t.
P20. *want*(x, p, t): agent x wants that p at time t.
P21. *obligation*(x, p, y, t): agent x has the obligation that p created by agent y at time t.
P22. *intend*(x, p, t): agent x intends that p at time t.
P23. *coerce*(x, y, p, t): agent x coerces agent y that p at time t.
P24. *superior*(x, y): agent x is a superior of agent y.
P25. *enable*(x, E, t): agent x makes an effect set E true at time t (*enable*(x, $\neg E$, t) means that agent x disables effect set E by making at least one effect in E false at time t).
P26. *can-enable*(x, E, t): agent x is capable of making an effect set E true at time t (*can-enable*(x, $\neg E$, t) means that agent x can disable effect set E by making at least one effect in E false at time t).
P27. *true*(E, t): effect set E is true at time t (this means that every effect in E is true at time t, and $\neg$*true*(E, t) means at least one effect in E is false at time t).

Predicates *P1–P10* denote the features related to plan structure and action execution. Predicates *P11–P16* represent communicative acts. These predicates are used to express task knowledge and observations of action execution and agent communication. Predicates *P17–P23* describe the epistemic variables (including attributions) used for inferring intermediate beliefs. Predicates *P24–P26* represent the power relationship and capabilities of agents.

Functions

Let A be an action, e be an action effect and DT be the domain theory[5]. We adopt the following functions in the model:

F1. *subaction*(A): subaction set of an abstract action A.
F2. *choice*(A): choice set for performing an abstract action A.
F3. *precondition*(A): precondition set of an action A.
F4. *effect*(A): (definite) effect set of an action A.
F5. *conditional-effect*(A): conditional effect set of an action A.
F6. *antecedent*(e): antecedent set of a conditional effect e.
F7. *consequent*(e): consequent of a conditional effect e.
F8. *indefinite-effect*(A): indefinite effect set of an action A.
F9. *relevant-action*(e, DT): relevant action set to achieve an effect e based on the domain theory DT.
F10. *relevant-effect*(e, DT): relevant effect set to achieve an effect e based on the domain theory DT.
F11. *side-effect*(e, DT): side effect set to achieve an effect e based on the domain theory DT.
F12. *performer*(A): performing agent(s) of an action A.
F13. *authority*(A): authorizing agent(s) of an action A.
F14. *primary-responsible*(e): primary responsible agent(s) for an effect e.
F15. *secondary-responsible*(e): secondary responsible agent(s) for an effect e.

[5] Domain theory is a general term used in planning and plan-based systems, specifying the actions performed in a domain and state affairs (typically described as preconditions and effects) that are causally linked to the actions. Domain theory is the general knowledge of the domain represented using a given plan representation.

Among these functions, *F1–F7* denote the generic features in (hierarchical) plan representation. Functions *F8–F11* describe indefinite effect set, relevant action/effect and side effect, and functions *F12–F15* represent the agents involved.

4.2 Reasoning about Social Causality

Social causality and responsibility judgment involve evaluating outcomes of events with personal significance to an agent. This evaluation is always from a perceiving agent's subjective perspective. The perceiver uses her knowledge about the observed agents and observation of behavior to infer beliefs of social attributions. We show how automatic methods of causal and dialogue reasoning can provide such a mechanism.

4.2.1 Dialogue Inference

Conversation between agents is a rich source of information for deriving attribution values. Early attribution theorists (Kidd & Amabile, 1981; Hilton, 1990) have pointed out the importance of language communication in attributing behavior. Within AI research community, there has been much related work on intentions in agent communication (Cohen & Levesque, 1990; Smith & Cohen, 1996), plan inference (Allen & Perrault, 1980; Litman & Allen, 1990), discourse structure (Grosz & Sidner, 1986; Lochbaum, Grosz, & Sidner, 2000) and speech act theory (Perrault, 1990). Although some previous research have partially addressed the issue of inferring intentions under different formalism, our focus here is on identifying the generic commonsense reasoning rules of attribution variables as well as their interrelations from social communication.

Natural language communication can be seen as a collaborative activity between conversational agents. Successful communication requires the participants to follow the basic conversation principles (Grice, 1975) and reach some degree of common ground (Clark & Schaefer, 1987). Thus we assume communication between agents is *grounded* (Traum, 1994), and conversation conforms to Grice's maxims of *Quality*[6] and *Relevance*[7]. In a conversational dialogue, the participating agents exchange information alternatively. A perceiving agent (who can be one of the participating agents or another agent) forms and updates beliefs according to the observed speech acts and previous beliefs.

We design commonsense rules that allow a perceiving agent to derive beliefs about the epistemic states of the observed agents. We also take social information (i.e., social roles and relationship) into consideration. For example, an order can be successfully issued only to subordinates, but a request can be made of any agent; and same request performed by agents with different social status may lead to different belief derivations.

Hobbs (1985) proposed a first-order logic notation, using *eventuality*[8] to reify events and conditions. To avoid expressing higher-order properties in first-order logic, our formalism has adopted this notation; but for simplification and ease of illustration, we still keep the higher-

[6] The quality maxim states that one ought to provide true information in conversation.
[7] The relevance maxim states that one's contribution to conversation ought to be pertinent in context.
[8] Eventuality is an extra argument used in each predication referring to the condition that exists when that predication is true. For every predicate $P(x)$, P is true of x if and only if there is an eventuality or possible situation e' of P being true of x (called P') and e' really exists, i.e. $(\forall x)P(x)\Leftrightarrow(\exists e')P'(e',x)\wedge Exist(e')$. The work of Hobbs (1985) provided further explanation on the ontological assumptions of the notation.

order expressions in this paper (note that they are actually handled using Hobb's notation in our approach). Also, to simplify logical forms, universal quantifiers are omitted in the rules, and we substitute *A* and *e* for *do*(x, A) and *achieve*(x, e) respectively, when causing no confusion.

If at time *t1*, a speaker (s) *informs* (or *tells*) a hearer (h) the content p, then after *t1*, it can be inferred that the speaker knows that proposition p as long as there is no intervening contradictory belief (*Rule D1*). As conversations between agents are grounded, it can be inferred that the hearer also knows that p (*Rule D2*). To further simplify the expressions of rules, we introduce a predicate *etc*[9] which stands for the absence of contradictory situations.

Rule *D1* [*inform*]:
$\text{inform}(s, h, p, t1) \wedge t1<t2 \wedge \text{etc}_1 \Rightarrow \text{know}(s, p, t2)$
Rule *D2* [*inform-grounded*]:
$\text{inform}(s, h, p, t1) \wedge t1<t2 \wedge \text{etc}_2 \Rightarrow \text{know}(h, p, t2)$

A *request* shows what the speaker wants (*Rule D3*). An *order* (or *command*) shows what the speaker intends (*Rule D5*). An order can only be successfully issued by someone higher in social status. If requested or ordered by a superior, it creates a social obligation for the hearer to perform the content of the act (*Rules D4 & D6*).

Rule *D3* [*request*]:
$\text{request}(s, h, p, t1) \wedge t1<t2 \wedge \text{etc}_3 \Rightarrow \text{want}(s, p, t2)$
Rule *D4* [*superior-request*]:
$\text{request}(s, h, p, t1) \wedge \text{superior}(s, h) \wedge t1<t2 \wedge \text{etc}_4 \Rightarrow \text{obligation}(h, p, s, t2)$
Rule *D5* [*order*]:
$\text{order}(s, h, p, t1) \wedge t1<t2 \wedge \text{etc}_5 \Rightarrow \text{intend}(s, p, t2)$
Rule *D6* [*order*]:
$\text{order}(s, h, p, t1) \wedge t1<t2 \wedge \text{etc}_6 \Rightarrow \text{obligation}(h, p, s, t2)$

The hearer may *accept, reject* or *counter-propose* an order (or request). Various inferences can be made depending on the response of the hearer and the social relationship between the speaker and the hearer. For instance, if the hearer accepts, and there is no obligation beforehand or the hearer is willing to (i.e., wants), it can be inferred that the hearer intends (*Rules D7 & D8*).

Rule *D7* [*accept*]:
$\neg\text{obligation}(h, p, s, t1) \wedge \text{accept}(h, p, t2) \wedge t1<t2<t3 \wedge \text{etc}_7 \Rightarrow \text{intend}(h, p, t3)$
Rule *D8* [*willing-accept*]:
$\text{want}(h, p, t1) \wedge \text{accept}(h, p, t2) \wedge t1<t2<t3 \wedge \text{etc}_8 \Rightarrow \text{intend}(h, p, t3)$

If there is no clear evidence of an agent's willingness, yet the agent accepts the obligation, there is evidence of coercion (*Rule D9*). In another case, if an agent is obviously unwilling to (i.e., unintended) but accepts the obligation, there is *clear* evidence of coercion (*Rule D10*).

Rule *D9* [*accept-obligation*]:

[9] This is similar to the notation used in the work of Hobbs, Stickel, Appelt, and Martin (1993). It essentially means that there is no contradictory belief in between.

$\neg(\exists t1)(t1<t3 \wedge \text{intend}(h, p, t1)) \wedge \text{obligation}(h, p, s, t2) \wedge \text{accept}(h, p, t3) \wedge t2<t3<t4 \wedge \text{etc}_9$
$\Rightarrow \text{coerce}(s, h, p, t4)$

Rule *D10* [*unwilling-accept-obligation*]:
$\neg\text{intend}(h, p, t1) \wedge \text{obligation}(h, p, s, t2) \wedge \text{accept}(h, p, t3) \wedge t1<t3 \wedge t2<t3<t4 \wedge \text{etc}_{10} \Rightarrow \text{coerce}(s, h, p, t4)$

If the hearer *rejects*, infer that the hearer does not intend (*Rule D11*). If the hearer *counters A* and *proposes B* instead, both the speaker and the hearer are believed to know that *A* and *B* are alternatives (*Rules D12 & D13*). It also implies what the hearer wants and does not intend (*Rules D14 & D15*).

Rule *D11* [*reject*]:
$\text{reject}(h, p, t1) \wedge t1<t2 \wedge \text{etc}_{11} \Rightarrow \neg\text{intend}(h, p, t2)$

Rule *D12* [*counter-propose*]:
$\text{counter-propose}(h, A, B, s, t1) \wedge t1<t2 \wedge \text{etc}_{12} \Rightarrow \text{know}(h, \text{alternative}(A, B), t2)$

Rule *D13* [*counter-propose-grounded*]:
$\text{counter-propose}(h, A, B, s, t1) \wedge t1<t2 \wedge \text{etc}_{13} \Rightarrow \text{know}(s, \text{alternative}(A, B), t2)$

Rule *D14* [*counter-propose*]:
$\text{counter-propose}(h, p, q, s, t1) \wedge t1<t2 \wedge \text{etc}_{14} \Rightarrow \neg\text{intend}(h, p, t2)$

Rule *D15* [*counter-propose*]:
$\text{counter-propose}(h, p, q, s, t1) \wedge t1<t2 \wedge \text{etc}_{15} \Rightarrow \text{want}(h, q, t2)$

If the speaker has *known* the *alternatives* and still *requests* (or *orders*) one of them, infer that the speaker wants (or intends) the chosen action and does not intend the alternative (*Rules D16 & D17*). (Here *z* can be *s* or *h*.)

Rule *D16* [*know-alternative-request*]:
$\text{know}(s, \text{alternative}(A, B), t1) \wedge \text{request}(s, h, \text{do}(z, A), t2) \wedge t1<t2<t3 \wedge \text{etc}_{16} \Rightarrow \neg\text{intend}(s, \text{do}(z, B), t3)$

Rule *D17* [*know-alternative-order*]:
$\text{know}(s, \text{alternative}(A, B), t1) \wedge \text{order}(s, h, A, t2) \wedge t1<t2<t3 \wedge \text{etc}_{17} \Rightarrow \neg\text{intend}(s, \text{do}(h, B), t3)$

4.2.2 CAUSAL INFERENCE

Plan representation gives further information for inferring agency, intention and coercion, in both direct and indirect cases. Causal inference is a plan-based evaluation based on the causal information provided by plan representation.

Agency. In a plan execution environment where multiple agents inhabit, agents' plans can interact in various ways. The preconditions of an agent's action may be established by the activities of other agents, and thus these other agents indirectly help cause the outcome. Given the domain theory *DT*, observed executed actions and an outcome *e*, the performer of an action *A* that directly causes *e* is the *causal agent* (*Rule C1*). Other performers of relevant actions to achieve *e* have *indirect agency* (*Rule C2*). In the absence of coercion, causal agent is deemed responsible for *e*, while other agents assist causing *e* should share responsibility with this causal agent. (The computation of relevant actions and effects to achieve *e* is given in *Appendix A*.)

Rule *C1* [*cause-action-effect*]:
$\text{execute}(x, A, t1) \wedge e\in\text{effect}(A) \wedge \text{occur}(e, t2) \wedge t1<t2<t3 \wedge \text{etc}_{18} \Rightarrow \text{cause}(x, e, t3)$

Rule *C2* [*cause-relevant-effect*]:

$\text{cause}(y, e', t1) \wedge e' \in \text{relevant-effect}(e, DT) \wedge \text{cause}(x, e, t2) \wedge t1<t2<t3 \wedge \text{etc}_{19} \Rightarrow \text{assist-cause}(y, x, e, t3)$

Intention. Attribution of intention is essential to people's explanations of behavior (Heider, 1958; Malle & Knobe, 1997). As we have discussed in *Section 4.2.1*, intentions can be inferred from evidence in natural language conversation. Causal inference helps infer outcome intention from evidence of act intention. For example, if an agent intends an action A voluntarily, the agent must intend at least one action effect of A (*Rule C3*).

Rule *C3* [*intend-action*]:

$\text{intend}(x, \text{do}(z, A), t1) \wedge \neg(\exists y)\text{coerce}(y, x, A, t1) \wedge t1<t2 \wedge \text{etc}_{20} \Rightarrow \exists e(e \in \text{effect}(A) \wedge \text{intend}(x, e, t2))$

In more general cases, when an action has multiple effects, in order to identify whether a specific outcome is intended or not, a perceiver may examine action *alternatives* the agent intends and does not intend, and compare the effects of intended and unintended alternatives. If an agent intends an action A voluntarily and does not intend its alternative B, we can infer that the agent either intends (at least) one action effect that only occurs in A or does not intend (at least) one effect that only occurs in B, or both. If the effect set of A is a subset of that of B, or if the effect set of B is a subset of that of A, they can be further simplified (*Rules C4 & C5*).

Rule *C4* [*intend-one-alternative*]:

$\text{intend}(x, \text{do}(z, A), t1) \wedge \neg\text{intend}(x, \text{do}(z, B), t1) \wedge \neg(\exists y)\text{coerce}(y, x, A, t1) \wedge \text{alternative}(A, B) \wedge \text{effect}(A) \subset \text{effect}(B) \wedge t1<t2 \wedge \text{etc}_{21} \Rightarrow \exists e(e \notin \text{effect}(A) \wedge e \in \text{effect}(B) \wedge \neg\text{intend}(x, e, t2))$

Rule *C5* [*intend-one-alternative*]:

$\text{intend}(x, \text{do}(z, A), t1) \wedge \neg\text{intend}(x, \text{do}(z, B), t1) \wedge \neg(\exists y)\text{coerce}(y, x, A, t1) \wedge \text{alternative}(A, B) \wedge \text{effect}(B) \subset \text{effect}(A) \wedge t1<t2 \wedge \text{etc}_{22} \Rightarrow \exists e(e \in \text{effect}(A) \wedge e \notin \text{effect}(B) \wedge \text{intend}(x, e, t2))$

If there is no clear belief of intention derived from causal and dialogue inferences, we can employ *intention recognition* as a general approach to detecting intentions. Given the observed executed actions of agent(s) and a plan library, if the observed action sequence matches the actions in a primitive plan, then we can certainly infer that the primitive plan is pursued by the agent(s). In most situations, however, the observed action sequence can only partially match a specific plan. To find a hypothesized plan that best explains the observed actions, most intention recognition algorithms use probabilistic models for the inference. We have developed a general intention recognition algorithm based on probabilistic plan inference (Mao, Gratch, & Li, in press). Our algorithm recursively uses causal information in the plan representation to compute the best candidate plan. Here we provide the criteria for determining intended actions and effects.

If an agent intends a certain *plan* to achieve the *goal* of the plan, then the agent should intend those actions and effects that are relevant to achieving the goal in the plan context (*Rules C6 & C7*). The goal itself should be intended by definition. Other side effects are not intended by the agent (*Rule C8*). (The computation of relevant actions and effects as well as side effects in the plan context is given in *Appendix A*.)

Rule *C6* [*intend-plan*]:

$\text{intend}(x, \text{by}(plan, goal), t1) \wedge A \in \text{relevant-action}(goal, plan) \wedge t1<t2 \wedge \text{etc}_{23} \Rightarrow \text{intend}(x, A, t2)$

Rule *C7* [*intend-plan*]:
intend(x, by($plan$, $goal$), $t1$) $\wedge$ $e\in$relevant-effect($goal$, $plan$) $\wedge$ $t1<t2$ $\wedge$ etc$_{24}$ $\Rightarrow$ intend(x, e, $t2$)
Rule *C8* [*intend-plan*]:
intend(x, by($plan$, $goal$), $t1$) $\wedge$ $e\in$side-effect($goal$, $plan$) $\wedge$ $t1<t2$ $\wedge$ etc$_{25}$ $\Rightarrow$ $\neg$intend(x, e, $t2$)

Foreknowledge. As foreknowledge belongs to an agent's epistemic state, it is mainly derived from dialogue inference. Speech act such as *inform* or *tell*, gives the evidence that the conversants know the content of the act. Intention recognition also helps infer an agent's foreknowledge, as intention entails foreknowledge: if an agent intends an action A to achieve an effect e of A, then the agent must *know* that A brings about e (*Rule C9*).

Rule *C9* [*intent-foreknowledge-relation*]:
intend(x, by(A, e), $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{26}$ $\Rightarrow$ know(x, bring-about(A, e), $t2$)

In addition, an agent *should know* what her action would bring about, if the action and its effects are general knowledge in the plan representation and the perceiver does not have contradictory belief of the specific knowledge the involved agents have (*Rules C10 & C11*).

Rule *C10* [*foreknowledge-performer*]:
$e\in$effect(A) $\wedge$ etc$_{27}$ $\Rightarrow$ know(performer(A), bring-about(A, e), $t1$)
Rule *C11* [*foreknowledge-authority*]:
$e\in$effect(A) $\wedge$ etc$_{28}$ $\Rightarrow$ know(authority(A), bring-about(A, e), $t1$)

Coercion. A causal agent could be absolved of responsibility if she was coerced to cause some outcome by other forces. But just applying coercive force does not mean outcome coercion actually occurs. What really matters is whether this force truly constrains the causal agent's freedom to avoid the outcome. Causal inference helps infer outcome coercion from evidence of act coercion.

If an agent is coerced to execute a primitive action, the agent is also coerced to achieve all the action effects (*Rule C12*). If being coerced to execute an abstract action and the action has only one decomposition (i.e., non-decision node), then the agent is also coerced to execute the subsequent actions and achieve all the subaction effects (*Rules C13 & C14*).

Rule *C12* [*coerce-primitive*]:
coerce(y, x, A, $t1$) $\wedge$ primitive(A) $\wedge$ $e\in$effect(A) $\wedge$ $t1<t2$ $\wedge$ etc$_{29}$ $\Rightarrow$ coerce(y, x, e, $t2$)
Rule *C13* [*coerce-non-decision-node*]:
coerce(y, x, A, $t1$) $\wedge$ and-node(A) $\wedge$ $B\in$subaction(A) $\wedge$ $t1<t2$ $\wedge$ etc$_{30}$ $\Rightarrow$ coerce(y, x, B, $t2$)
Rule *C14* [*coerce-non-decision-node*]:
coerce(y, x, A, $t1$) $\wedge$ and-node(A) $\wedge$ $e\in$effect(A) $\wedge$ $t1<t2$ $\wedge$ etc$_{31}$ $\Rightarrow$ coerce(y, x, e, $t2$)

If the coerced action has multiple decompositions (i.e., decision node), then the subsequent actions are not coerced (*Rule 15*). Since the agent has options, only the effects that appear in all alternatives are unavoidable (i.e., *definite*), and thus these effects are coerced (*Rule 16*); Other effects that only appear in some (but not all) alternatives are avoidable (i.e., *indefinite*), so they are not coerced (*Rule 17*). (The computation of definite and indefinite effects is given in *Appendix B*.)

Rule *C15* [*coerce-decision-node*]:
coerce(y, x, A, $t1$) $\wedge$ or-node(A) $\wedge$ $B\in$choice(A) $\wedge$ $t1<t2$ $\wedge$ etc$_{32}$ $\Rightarrow$ $\neg$coerce(y, x, B, $t2$)

Rule *C16* [*coerce-decision-node*]:
$\text{coerce}(y, x, A, t1) \wedge \text{or-node}(A) \wedge e \in \text{effect}(A) \wedge t1<t2 \wedge \text{etc}_{33} \Rightarrow \text{coerce}(y, x, e, t2)$

Rule *C17* [*coerce-decision-node*]:
$\text{coerce}(y, x, A, t1) \wedge \text{or-node}(A) \wedge e \in \text{indefinite-effect}(A) \wedge t1<t2 \wedge \text{etc}_{34} \Rightarrow \neg\text{coerce}(y, x, e, t2)$

Given a *conditional effect* is coerced, if its antecedents are initially true, its consequent is also coerced (*Rule C18*). Otherwise, if its antecedents are false initially, then the consequent is not coerced (*Rule C19*). If the antecedents are established by self (i.e., the performer), then the consequent is not coerced, as she could choose to do otherwise (*Rule C20*). If some other agent(s) establish the antecedents, then these other agents assist coercing the consequent (*Rule C21*).

An agent can be indirectly coerced (e.g., by enabling/disabling action preconditions, or blocking other action alternatives). If among the choices of the coerced action, there is only one executable alternative available or the coerced agent can enable only one alternative (i.e., by making action preconditions true), then the agent is coerced to execute the only alternative (*Rules C22 & C23*). If the only available alternative is *enabled* by some other agent(s), then these other agents assist coercing the only alternative (*Rule C24*). If some other agent(s) *block* other action alternatives (by disabling action preconditions), then the only alternative left is coerced and these blocking agents are also coercers (*Rule C25*).

Coercion entails intention. Handing over one's wallet under the threat of "your money or your life" may well be seen as intentional: one decides to do so, albeit unwillingly, with the goal of saving life.

Rule *C26* [*coerce-intend-relation*]:
$\text{coerce}(y, x, p, t1) \wedge t1<t2 \wedge \text{etc}_{43} \Rightarrow \text{intend}(x, p, t2)$

The complete inference rules are given in *Appendix C*.

4.3 Attribution Algorithm

The beliefs derived from dialogue and causal inferences are used in the attribution process to form an overall judgment. Different perceivers may have different observations, different knowledge and preferences, thus they may form different beliefs and judge the same situation differently. Despite individual differences, the posited attribution process is *general*, and applies uniformly to different perceivers. If an action performed by an agent brings about a positive or negative effect, and the agent is not coerced to achieve the action effect, then the *performer* of the action is the *primary responsible* agent. Other agents who indirectly assist the performer are the *secondary responsible* agents. In the presence of external coercion, the primary responsible agent is redirected to the *coercer* (Note that coercion may occur in more than one level of action hierarchy, and so the process may need to trace several levels up to find the ultimate source of responsibility). Other agents who indirectly assist the coercer are the secondary responsible agents. They should share responsibility with the primary responsible agent.

We have developed an algorithm to find the responsible agent(s) for a specific outcome (consequence e). First, based on the speech act (SA) sequence, the algorithm infers from dialogue evidence (*Step 1*). Then it applies causal inference rules (*Step 2*). For each executed action that potentially leads to the consequence, if the action does cause the outcome occurrence or the performer of the action intends to bring the outcome about (i.e. failed attempt) (*Step 3.1*), then assign the performer to the primary responsible agent. Other agents who assist the performer (by enabling action preconditions) are secondary responsible agents (*Step 3.2*). To trace the coercing agent(s), the evaluation process starts from the primitive action (*Step 3.3*), and works up the action hierarchy (*Step 3.4*). During each pass through the main loop, if there is evidence of outcome coercion (*Step 3.4.2*), the authority is deemed responsible (*Step 3.4.3*). If current action is not the root node in action hierarchy and outcome coercion is true, the algorithm assigns the parent node to current action (*Step 3.4.4*) and evaluates the next level up. If the outcome is intended by the responsible agent (*Step 3.5*), the degree of responsibility is high (*Step 3.6*). If the outcome is not intended (*Step 3.7*), then the degree assigned is low (*Step 3.8*). Otherwise, assign medium degree of responsibility (*Step 3.9*). At last, the algorithm returns the primary and secondary responsible agents as well as the degrees of responsibility (*Step 4*).

Attribution Algorithm (*SA sequence S, domain theory DT, consequence e, observations*):

```
1.     Based on the speech act sequence S, apply dialog inference rules
2.     Based on DT in the plan representation, apply causal inference rules
3.     FOR  each executed action A in observations
3.1          IF   cause(performer(A), e)  OR  intend(performer(A), by(A, e))   THEN
3.2               primary-responsible(e) = performer(A)
                  secondary-responsible(e) = performer(relevant-action(e, DT))
3.3               P = A
3.4               DO
3.4.1                  B = P
3.4.2                  IF   coerce(authority(B), performer(B), e)   THEN
3.4.3                         primary-responsible(e) = authority(B)
3.4.4                         P = parent of node B in DT
                       END-IF
                  WHILE   B ≠ root of action hierarchy  AND  coerce(authority(B), performer(B), e)
3.5               IF   intend(primary-responsible(e), e)   THEN
3.6                    Assign high degree of responsibility
3.7               ELSE IF   ¬intend(primary-responsible(e), e)   THEN
3.8                    Assign low degree of responsibility
3.9               ELSE   assign medium degree of responsibility
             END-IF
       END-FOR
4.     RETURN  primary-responsible(e) ∪ secondary-responsible(e);  Degrees of responsibility
```

We adopted the categorical model of responsibility assignment. If the outcome is intended by the responsible agent, the degree of responsibility is *high* (Recall that as long as it manifests intentions, a failed attempt can be blamed or credited almost the same as a successful one). If the outcome is not intended by the responsible agent, then the degree of responsibility is *low*. Otherwise, if there is no clear evidence of outcome intention, assign *medium* degree of responsibility. The *intensity* of credit or blame is computed by multiplying the degree of responsibility and the utility of the outcome. Events may lead to more than one desirable/undesirable outcomes. For evaluating multiple outcomes, we apply the algorithm the same way, focusing on one outcome each time during its execution. Finally, to form an overall judgment, the results are aggregated and grouped by the responsible agents.

4.4 Illustrative Example

We use an example from the Mission Rehearsal Exercise (*MRE*) leadership training system (Swartout et al., 2006) to illustrate how the model works. In the *MRE* system, a human trainee can practice decision making skills through interactions with virtual autonomous agents. To train students in high-stake social situations, these virtual agents not only have figures that resemble humans, they should also make sense of the perceived social events and exhibit human-like social reasoning ability. The training scenario opens with a *lieutenant* (played by the student), who lead a troop of soldiers to fulfill a peacekeeping mission. On his way to reinforce another unit, one of the troop's vehicles has seriously injured a civilian boy. The boy's mother and a medic are in the accident area, and a crowd is gathering around. The student is faced with the dilemma of whether to continue his mission or to render aid to the boy. Many decisions are possible, and each decision he makes will lead to different outcomes as the scenario unfolds. Here the important question for our work is that when some good or bad outcomes occur, how to ensure the agents make reasonable judgments and react like people in such social situations.

In one training exercise, for example, the student (i.e. lieutenant) decided to split his forces. He ordered his *sergeant* (acted by an autonomous agent) to send half of his squads to assist another unit. The sergeant informed of the bad consequence and tried to negotiate for a better alternative. However, the student persisted with his decision, and finally, the sergeant ordered the *squad leader* (Lopez) to perform the act. Three social actors are involved in this example. The lieutenant acts as an authority over the sergeant. The squad leader acts as a subordinate of the sergeant. The following dialogue is extracted from an actual run of the system. Below we illustrate how to attribute responsibility and blame based on the causal knowledge and observations of agents.

Student: Sergeant, send two squads forward. (Line *1*)
Sergeant: That is a bad idea, sir. We shouldn't split our forces. (Line *2*) Instead we should send one squad to recon forward. (Line *3*)
Student: Send two squads forward. (Line *4*)
Sergeant: Against my recommendation, sir. (Line *5*) Lopez! Send first and fourth squads to Eagle 1-6's location. (Line *6*)
Lopez: Yes, sir. Squads! Mount up! (Line *7*)

Within the *MRE* system, conversations between agents are represented as speech acts and a dialogue history is stored. Details on how this negotiation dialogue is automatically generated and how natural language is mapped into speech acts can be found in the work of Traum and his colleagues (2003, 2008). The dialogue above corresponds to the following speech acts, ordered by the time the speakers addressed them. (The symbols *lt*, *sgt* and *sld* stand for the lieutenant, the sergeant and the squad leader, respectively. $t1<t2<\ldots<t7$.)

Act 1: order(*lt*, *sgt*, do(*sgt*, *send-two-sqds*), *t1*) (Line *1*)
Act 2: inform(*sgt*, *lt*, bring-about(*send-two-sqds*, *unit-fractured*), *t2*) (Line *2*)
Act 3: counter-propose(*sgt*, do(*sgt*, *send-two-sqds*), do(*sgt*, *send-one-sqd*), *lt*, *t3*) (Line *3*)

Act 4: order(*lt*, *sgt*, do(*sgt*, *send-two-sqds*), *t4*) (Line *4*)
Act 5: accept(*sgt*, do(*sgt*, *send-two-sqds*), *t5*) (Line *5*)
Act 6: order(*sgt*, *sld*, do(*sld*, *two-sqds-fwd*), *t6*) (Line *6*)
Act 7: accept(*sld*, do(*sld*, *two-sqds-fwd*), *t7*) (Line *7*)

Figure 3 illustrates the causal knowledge of the troop underlying the example. Take the sergeant's perspective as an example. The sergeant has access to the partial plan knowledge of the troop, and perceives the conversation between the actors and task execution. He observed a physical action *two-squads-forward* executed by the squad leader and the occurrence of action effects. Two effects are salient to the sergeant, (*unit*) *1-6 supported* and *unit fractured*. Supporting unit 1-6 is a desirable team goal. Assume *unit fractured* is undesirable to the sergeant and so he assigns negative utility to it. This consequence serves as input of the algorithm.

Step 1. Based on sequence *1-7* in the dialogue history, the sergeant can derive a number of beliefs by inferring the observed speech acts (Here $t1<t1'<t2<t2'<\ldots<t7<t7'$):

Belief 1: intend(*lt*, do(*sgt*, *send-two-sqds*), *t1'*) (Act *1*, Rule *D5*)
Belief 2: obligation(*sgt*, do(*sgt*, *send-two-sqds*), *lt*, *t1'*) (Act *1*, Rule *D6*)
Belief 3: know(*sgt*, bring-about(*send-two-sqds*, *unit-fractured*), *t2'*) (Act *2*, Rule *D1*)
Belief 4: know(*lt*, bring-about(*send-two-sqds*, *unit-fractured*), *t2'*) (Act *2*, Rule *D2*)
Belief 5: know(*sgt*, alternative(*send-two-sqds*, *send-one-sqd*), *t3'*) (Act *3*, Rule *D12*)
Belief 6: know(*lt*, alternative(*send-two-sqds*, *send-one-sqd*), *t3'*) (Act *3*, Rule *D13*)
Belief 7: ¬intend(*sgt*, do(*sgt*, *send-two-sqds*), *t3'*) (Act *3*, Rule *D14*)
Belief 8: want(*sgt*, do(*sgt*, *send-one-sqd*), *t3'*) (Act *3*, Rule *D15*)
Belief 9: ¬intend(*lt*, do(*sgt*, *send-one-sqd*), *t4'*) (Act *4*, Belief *6*, Rule *D17*)
Belief 10: coerce(*lt*, *sgt*, do(*sgt*, *send-two-sqds*), *t5'*) (Act *5*, Beliefs *2&7*, Rule *D10*)
Belief 11: intend(*sgt*, do(*sld*, *two-sqds-fwd*), *t6'*) (Act *6*, Rule *D5*)
Belief 12: obligation(*sld*, do(*sld*, *two-sqds-fwd*), *sgt*, *t6'*) (Act *6*, Rule *D6*)
Belief 13: coerce(*sgt*, *sld*, do(*sld*, *two-sqds-fwd*), *t7'*) (Act *7,* Belief *12*, Rule *D9*)

Step 2. Based on the observations of task execution and the beliefs obtained in *Step 1*, causal inference further derives the following beliefs of the sergeant (Here *t0* is the initial time, $t0<t0'<t1$):

Belief 14: know(*sld*, bring-about(*two-sqds-fwd*, *unit-fractured*), *t0'*) (Rule *C10*)
Belief 15: know(*sgt*, bring-about(*two-sqds-fwd*, *unit-fractured*), *t0'*) (Rule *C11*)
Belief 16: intend(*lt*, *unit-fractured*, *t4'*) (Beliefs *1&9*, Rule *C5*)
Belief 17: coerce(*lt*, *sgt*, do(*sgt*, *two-sqds-fwd*), *t5'*) (Belief *10*, Rule *C13*)
Belief 18: coerce(*lt*, *sgt*, do(*sgt*, *remaining-fwd*), *t5'*) (Belief *10*, Rule *C13*)
Belief 19: coerce(*lt*, *sgt*, *1-6-supported*, *t5'*) (Belief *10*, Rule *C14*)
Belief 20: coerce(*lt*, *sgt*, *unit-fractured*, *t5'*) (Belief *10*, Rule *C14*)
Belief 21: coerce(*sgt*, *sld*, *unit-fractured*, *t7'*) (Belief *13*, Rule *C12*)

Step 3. Steps 3.1–3.2: As action *two-squads-forward* directly causes the evaluated outcome *unit-fractured*, and the action is performed by the squad leader, initially, assign the *squad leader* to the responsible agent.

Step 3.4: *Loop 1*: The algorithm starts from the primitive action *two-squads-forward*. The sergeant believes that he coerced the squad leader to fracture the unit (*Belief 21*). The sergeant also believes that both he and the squad leader should have foreseen the outcome *unit-fractured* (*Beliefs 14&15*). As outcome coercion is true, the *sergeant* is assigned to the responsible agent. Since outcome coercion is true and current node is not the root of the action hierarchy, the algorithm enters next loop.

Loop 2: The action is *send-two-squads*, performed by the sergeant. The sergeant believes that the lieutenant coerced him to fracture the unit (*Belief 20*). The sergeant also believes that the lieutenant intended *unit-fractured* (*Belief 16*). As outcome coercion is true, the *lieutenant* is assigned to the responsible agent. Since outcome coercion is true and current node is not the root of the action hierarchy, the algorithm enters next loop.

Loop 3: The action is *support-unit-1-6*, performed by the lieutenant. There is no relevant dialogue act in history, nor is there clear evidence of coercion. As current node is already the *root* of the action hierarchy, the algorithm exits the loop.

Steps 3.5–3.9: As the sergeant believes that the lieutenant intended *unit-fractured*, the *lieutenant* is assigned *high* degree of responsibility for the outcome.

5. Evaluation

To evaluate our computational framework, we need to assess the consistency between model predictions and human judgments of social cause, responsibility and blame/credit. In particular, we need to evaluate the consistency of the model's inferential mechanism underlying human attributions of responsibility and blame/credit – that is, whether our model uses the same sources of evidence and draws the same intermediate conclusions as people do. Thus, we design an experiment to test how our model performs in predicting the beliefs of intermediate variables (including attribution variables and other epistemic variables in the model) and evidence used for the inference process. We *claim* that our model predicts human judgments of social attributions and makes inferences consistent with what most people do in their judgments. As the alternative computational approaches are incapable of inferring the beliefs of intermediate variables, we directly compare the predictions of our model with human data.

5.1 Method

Participants and Procedure

The study consisted of *48* subjects that were either computer science graduate students or staff at the University of Southern California. Their ages range from *20* to *35*, and *30* of the subjects were male. Among them, *12* subjects each completed four scenarios of the survey. Other subjects each completed two scenarios. The survey was composed of four small scenarios where the order of the scenarios was randomized across subjects. Each scenario was followed by a questionnaire, asking questions about the assessments of internal variables including the characters' foreknowledge, desire, intentions, obligation and perceived coercions. In answering each question, the subjects were asked to mark the (multiple) lines in the scenario according to which they draw the answer. At the end of each questionnaire, there is a question asking the subjects to score how much blame the characters deserve in the scenario.

Materials

As a starting point, we adopt the "*company program*" scenario first used in (Knobe, 2003a). This scenario has received much attention in recent folk psychology and experimental philosophy research (Jones, 2009). We design three variants of the company program scenario and the questionnaires following each scenario. The original scenario (*Scenario 2*), its variants (*Scenarios 1*, *3* and *4*) and the complete questionnaires are given in *Appendix D*. For the convenience of assessing inference rules, descriptions of each scenario are organized into separate labeled lines of evidence (e.g., *E1*-*E6*).

Scenario 2:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.

E2 The vice president says, "The new program will help us increase profits,
E3 but according to our investigation report, it will also harm the environment."

E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"

E5 The vice president says, "Ok," and executes the new program.

E6 The environment is harmed by the new program.

Figure 4: Company Program Scenario 2

Experimental Design

As our model embodies the theoretical view that people will judge social cause and responsibility differently based on their perception of the key variables such as intention, foreknowledge and coercion, a good experimental design is to see how the model performs when evidence for such judgments is systematically varied. To this end, we take the description of a single social situation and systematically vary it, using the inference rules of our model as a guide. For example, if our model suggests that particular evidence supports the inference of coercion, then an obvious variation would be to add a line to the scenario encoding such evidence. By exploring the space of inference rules and generating the scenarios accordingly, we were able to incorporate information needed for different inference paths and to predict judgment results in a systematic way.

Based on the computational framework introduced in *Section 4*, the specific information utilized in the inference process includes those causal knowledge, goal identification, and observations of speech acts, physical actions and the occurrence of action effects. We encode the information into each line of the scenarios. The encoded information serves as the model's inputs and provides evidence for the specific inference. For example, in *Scenario 1*, the following information is encoded (*vp* and *chm* refer to the vice president and the chairman, respectively):

E1:	request(*vp*, *chm*, do(*vp*, *new-program*), *t1*)	(speech act)
E2:	inform(*vp*, *chm*, bring-about(*new-program*, *profit-increase*), *t2*)	(causal knowledge)
E3:	inform(*vp*, *chm*, ¬bring-about(*new-program*, *env-harm*), *t2*)	(causal knowledge)
E4:	accept(*chm*, do(*vp*, *new-program*), *t3*)	(speech act)
E5:	execute(*vp*, *new-program*, *t4*)	(action execution)
E6:	occur(*env-harm*, *t5*)	(outcome occurrence)

We design questions to test beliefs about different variables. Each question corresponds to the firing of an inference rule. We select to assess most groups of dialogue and causal inference rules (*D1-D17* and *C1-C17*). Some rules are tested in the virtual training system in *Section 4.4*. For dialogue inference, we design questions to test speech acts "*inform*", "*request*", "*order*", "*accept*", "*accept-obligation*" and "*counter-propose*". "*Know-alternative*" is tested in the virtual training scenario. For causal inference, we design questions to test "*intend-action*", "*intend-plan*", "*intent-foreknowledge-relation*", "*coerce-primitive*" and "*coerce-decision-node*". "*Intend-one-alternative*", "*foreknowledge*" and "*coerce-non-decision-node*" are tested in the virtual training scenario.

In *Scenario 1*, we manipulate evidence related to agents' foreknowledge of the outcome (i.e., no foreknowledge). We design questions to test the inference rules for foreseeability (*Question 4*, *Rule D1*), relation of intent and foreknowledge (*Question 5*, *Rule C9*), connection of act and outcome intentions (*Question 3*, *Rule C3*), etc. *Scenario 2* gives clear evidence of foreknowledge. The authority's goal is also stated. Correspondingly, questions are designed to test rules for intentional action/effect and side effect (*Questions 3-4*, *Rules C7*&*C8*), having foreknowledge (*Question 1*, *Rule D2*), and speech acts. In *Scenario 3,* we manipulate the degree of perceived coercion and unwillingness by introducing an alternative course of action that will not harm the environment and which the vice president prefers. Specifically, we add one line between *E3* and *E4* (and all the other lines remain the same as those in *Scenario 2*). Questions are designed to test the agent's willingness (*Question 2*, *Rules D14*&*D15*) and perceived coercion (*Questions 3-4*, *Rules D10*&*C12*). In *Scenario 4*, we manipulate the characters' freedom of choice. We introduce an alternative, but the preference of the vice president is based on a feature unrelated to the environment and the vice president is allowed to choose from the options. We design three questions to test other important rules for coercion (*Rules C15-C17*).

Model Predictions

For each question in the questionnaire, the model's prediction of belief and belief derivation are given in *Appendix E*.

5.2 Results

Here we provide the experimental results on assessing inferred beliefs and inference rules.

		Question 1		Question 2		Question 3		Question 4		Question 5		Question 6	
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Chair	VP
Scenario 1	Model	√		√		√			√		√		√
	People	**30**	0	**27**	3	**29**	1	2	**28**	0	**30**	3.00	**3.73**
Scenario 2	Model	√		√		√			√	√		√	
	People	**30**	0	**30**	0	**30**	0	10	**20**	**22**	8	**5.63**	3.77
Scenario 3	Model	√			√	√		√		N/A		√	
	People	**21**	9	2	**28**	**29**	1	**21**	9			**5.63**	3.23
Scenario 4	Model	√			√		√	N/A		N/A			√
	People	**21**	9	5	**25**	5	**25**					4.13	**5.20**

Table 1: Model Predictions and Subject Responses for Company Program Scenarios

5.2.1 ASSESSING INFERRED BELIEFS

Table 1 summarizes the experimental results. Results for questions *1* to *5* indicate the total number of subjects that gave a particular answer. For example, for *Scenario 1*, all thirty subjects reported that the vice president wanted to start the new program. Question *6* refers to the amounts of blame attributed to the chairman and the vice president on a scale of *1* (little) to *6* (lots), and the table lists the subjects' average reported values. The model's predictions are checked with '√' in the table. The data show that for most questions, people agree with each other quite well. But certain disagreements exist on some of the questions.

As our purpose is to assess the model's general agreement with people, we measure the agreement between the model and each subject using the *Kappa statistic*. The *Kappa* coefficient is the de facto standard to evaluate the agreement between raters, which factors out expected agreement due to chance (Carletta, 1996). The *K* coefficient is computed as:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ is the propositional agreement among raters. $P(E)$ is the expected agreement, that is, the probability that the raters agree by chance. Di Eugenio and Glass (2004) argued that the computation of *K* coefficient is sensitive to the skewed distribution of categories (i.e., prevalence). In our treatment, we account for prevalence and construct contingency tables for the calculation, and average the results of *Kappa* agreement of the model's predictions with each subject's answers. The average *Kappa* agreement between the model and subjects is ***0.732***. Based on the scales given by Rietveld and van Hout (1993), $0.6<K<0.8$ indicates substantial agreement. The empirical results show good consistency between the model's generation of intermediate beliefs and human data.

5.2.2 ASSESSING INFERENCE RULES

In our model, every belief is derived by a specific inference rule, so the answer to a question in the questionnaires corresponds to the firing of one rule (with the exception of three questions in the questionnaires designed to test two rules each). As the condition side of each rule is composed of a set of evidence, to assess the accuracies of the inference rules, we compare the conditions of each rule with the evidence people use in forming each answer. Accuracy of each rule is measured using standard *confusion matrix* (Kohavi & Provost, 1998). For every subject's evidence choice in each question, we build a confusion matrix to compute the number of true positive *TP* (i.e., evidence both the rule and the subject use), true negative *TN* (i.e., evidence both the rule and the subject ignore), false positive (i.e., evidence the rule incorrectly uses), and false negative (i.e., evidence the rule incorrectly ignores).

For each question Q_i, the correct selection of evidence by the corresponding rule with respect to subjects is measured by accuracy (*AC*), where N_s is the total number of subjects and N_e is the total number of evidence for Q_i.

$$AC(Q_i) = \frac{\sum_{j \in Subjects} AC(j, Q_i)}{N_s} = \frac{\sum_{j \in Subjects} (TP(j, Q_i) + TN(j, Q_i))}{N_s \times N_e}$$

Table 2 lists the accuracies of the tested rules. The average accuracy of these rules is ***0.85***. The empirical results show that the evidence the model uses for inference is consistent with human data. Thus the first experimental study generally supports our first claim of evaluation: our model predicts human judgments of social attributions and makes inferences consistent with what most people do in their judgments.

	Question	Inference Rule	Average Accuracy
Scenario 1	1	D3 [Request]	0.76
	2	D7 [Accept]	0.96
	3	C3 [Intend-Action]	0.85
	4	D1 [Inform]	0.94
	5	C9 [Intent-Foreknowledge-Relation]	0.91
Scenario 2	1	D2 [Inform-Grounded]	0.92
	2	D5 [Order]	0.96
	3	C7 [Intend-Plan]	0.86
	4	C8 [Intend-Plan]	0.70
	5	D6 & D9 [Order; Accept-Obligation]	0.84
Scenario 3	1	D13 [Counter-Propose-Grounded]	0.94
	2	D14 & D15 [Counter-Propose]	0.88
	3	D6 & D10 [Order; Unwilling-Accept-Obligation]	0.80
	4	C12 [Coerce-Primitive]	0.74
Scenario 4	1	C16 [Coerce-Decision-Node]	0.71
	2	C15 [Coerce-Decision-Node]	0.84
	3	C17 [Coerce-Decision-Node]	0.75

Table 2: Accuracies of Evidence Used by the Inference Rules

5.3 Discussion

Although the experimental results show fairly good consistency between our model's predictions and human data with respect to the inferred beliefs and inference rules, the results above also reveal several disagreements among the subjects and the accuracies of the evidence used by several inference rules are relatively lower. Now we briefly discuss the experimental findings from our first study.

In *Scenario 1*, the questionnaire specifically queries the perceived desire, foreknowledge and intentions of the characters. The accuracy of the rule tested in *Question 1* is lower than the others because, in addition to evidence *E1*, many people chose *E2* as well. Post-experiment interviews with the subjects uncovered that many subjects had assumed that making profits should be desirable to the vice president (because of his role), and therefore, she should want to start the new program to increase profits (which is supported by *E2*).

Scenarios 2 and 3 manipulate the degree of perceived coercion and willingness of the coerced agent. In *Question 4* of *Scenario 2*, one-third of the subjects think it the chairman's intention to harm the environment. Whether a side effect is intentional or not is controversial in philosophy, and other empirical studies show similar results as ours (Nadelhoffer, 2006). Also in *Question 5* of *Scenario 2*, some subjects think the vice president is not coerced to start the new program by the chairman, as the evidence is weaker than that in *Scenario 3*. Half of them referred to evidence *E5*, indicating that they expect the vice president to negotiate with the chairman rather than directly accept the order.

In the first question of *Scenario 3*, some subjects think the chairman does not know the alternative program, though the vice president clearly states this in the scenario. Most of these subjects (*80%*) referred to evidence *E5*, showing that they looked for grounding information. As our model infers grounded information from conversation, we should have considered this in the scenario design. In *Question 4* of *Scenario 3*, some subjects seemed reluctant to infer outcome coercion from evidence of act coercion. Nonetheless, they still assigned high degree of blame to the chairman.

In *Scenario 4*, the vice president has some freedom of choice. In *Question 1*, some subjects think that the vice president is not coerced to increase profits, for the same reason mentioned earlier. They think it the vice president's job to increase profits, so she must be willing to do so. The accuracies of the inference rules for *Question 1* and *Question 3* are relatively low. In our model, the evidence needed for the inference is *E3*, *E4* and *E5*. Many subjects ignore knowledge *E3* and this lowers the accuracies of the two rules (similar reason for the low accuracies of the rules used in *Question 4* of *Scenarios 2&3*).

Comparing the blame assignments in *Scenarios 2* and *3*, it shows that on the one hand, the higher the degree of coercion, the less blame is assigned to the actor − a result consistent with psychological findings. On the other hand, even when perceived coercion is not strong, people still assign high degree of blame to the coercer, as in *Scenario 2*. In *Scenario 4*, people assigned more blame to the vice president, as she could have done otherwise. This result is consistent with psychological findings (Shaver, 1985). However, people still assigned considerable blame to the chairman, though it was the vice president's choice to harm the environment.

5.4 Additional Experiment

In this section, we design an additional experiment to compare the overall judgment results by our model and the alternative models with human data. In *Section 2*, we have introduced Chockler and Halpern's (2004) model (abbreviated to *C&H model*) for responsibility and blame judgments. In addition to the *C&H* model, we also compare our model with two simple models. A *simple cause model* always assigns responsibility and blame to the actor whose action directly produces the outcome. This is the approach used by most current intelligent systems. Instead of picking up the actor, a slightly more sophisticated model captures the intuition that hierarchical structure is a universal characteristic of human society and organizations and social power always flows from the top in the organizational structure. A *simple authority model* can choose the highest authority as the responsible and blameworthy agent. Below we report our experiment with human data on the overall judgments and compare our model's predictions with the results by simple cause model, simple authority model and the *C&H* model.

5.4.1 METHOD

Participants and Procedure

Twenty-seven subjects participated in the experiment. They were either staffs or graduate students at the University of Southern California, with ages ranging from *20* to *45*, and *14* of the subjects were female. The subjects were presented with four similar scenarios. Each scenario was followed by a questionnaire, asking questions about the assessments of physical cause, responsibility, blame and perceived coercion of the characters. The order of the scenarios was randomly assigned.

Materials

We took as a starting point the "*firing squad*" scenario typically used in causality research. For the convenience of comparing with the related work, we used the original "*firing squad*" scenario in the work of Chockler & Halpern (2004) (*Scenario 1*), and designed its variants (*Scenarios 2*, *3* and *4*). Each scenario is followed by a questionnaire. The questions in the questionnaires are the same across scenarios. The original scenario, its variants and the wording of the questions are given in *Appendix F*.

Experimental Design

We designed the variants of *Scenario 1* to systematically vary the perception of the key variables such as intention and coercion. In each variant, we manipulate evidence of perceived coercion and intentions of agents. *Scenario 2* extends the example by including an authority - the commander, who orders the squad to shoot. *Scenario 3* further extends the example by presenting a negotiation dialogue between the commander and the marksmen. The marksmen first reject the commander's order. The commander insists and orders again. Finally the marksmen accept the order and shoot at the prisoner. In *Scenario 4*, the commander still orders, but each marksman has freedom to choose either using blanks or live bullets before shooting.

Model Predictions

Each alternative approach represents a typical way of handling social causality, responsibility and blame judgment. Below we give the predictions of our model (abbreviated to *M&G model*) and alternative models.

Simple cause model: The simple cause model uses physical causality as a substitute for social causality. So for each scenario, it predicts the marksman (or marksmen) with bullets as the responsible and blameworthy agent.

Simple authority model: The simple authority model judges social cause and responsibility from the top in power hierarchy, and regards the highest authority as being responsible. It assigns responsibility and blame to the commander in *Scenarios 2* to *4*.

C&H model: As each marksman is a real cause for the outcome, the *C&H* model predicts all marksmen share responsibility and blame in *Scenario 1*. For the similar reason, in *Scenarios 2 and 3*, the *C&H* model predicts both the commander and all marksmen are responsible and blameworthy. The model's prediction of *Scenario 4* depends on the context (We shall discuss more on this later).

M&G model: In *Scenario 1*, our model predicts the same result as that in the *C&H* model, but judges the commander as the sole responsible and blameworthy agent in *Scenarios 2 and 3*. In the last scenario, our model assigns responsibility and blame to the marksmen with bullets.

5.4.2 RESULTS

In answering the questions, the subjects choose the responsible and blameworthy agents from six categories. They are marksmen with bullets, all marksmen, commander, commander and marksmen with bullets, commander and all marksmen, and none of the above (see *Appendix E*). *Figure 5* shows the proportion of subjects that attribute blame and responsibility to different categories of agents, and the corresponding confidence intervals (α*=0.05*) (Rice, 1994). For example, in *scenario 1*, three subjects blame the marksman with live bullets in his rifle, *19* blame all the marksmen and the rest do not blame any of them. The analysis of the sample data and their confidence intervals show that a small percentage of the population will blame the marksman with live bullets, a significant majority will blame all the marksmen, and a small percentage won't blame any, with *0.95* confidence.

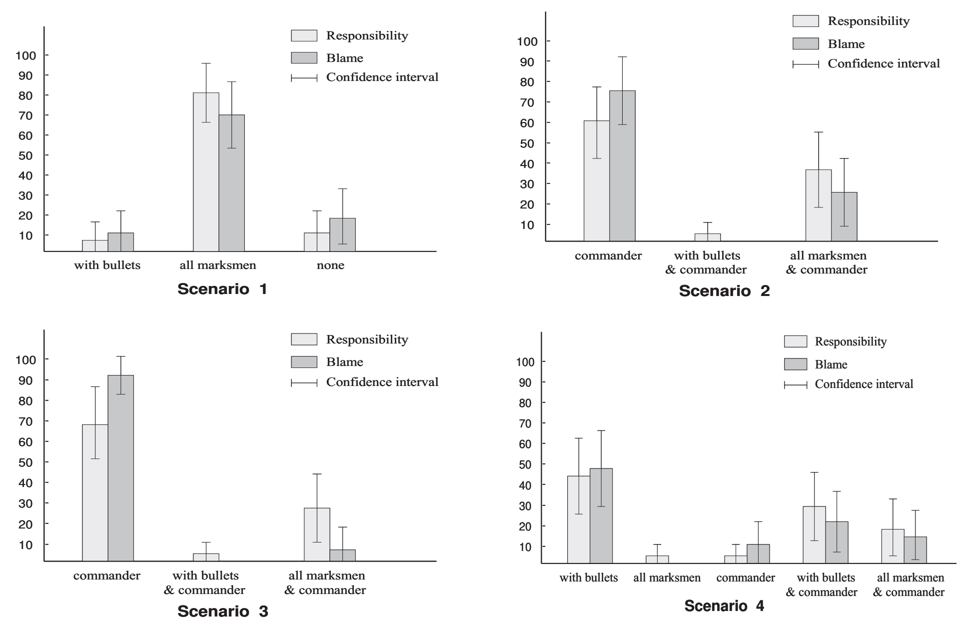


Figure 5: Proportion of Population Agreement on Responsibility/Blame in Scenarios

Blame	Simple Cause Model		Simple Authority Model		C&H Model		M&G Model		Human Majority Agreement
	Results	Match	Results	Match	Results	Match	Results	Match	
Scenario 1	with bullets	no	N/A	no	all marksmen	yes	all marksmen	yes	all marksmen
Scenario 2	with bullets	no	commander	yes	commander & all marksmen	no	commander	yes	commander
Scenario 3	with bullets	no	commander	yes	commander & all marksmen	no	commander	yes	commander
Scenario 4	with bullets	yes (partial)	commander	no	context dependent	—	with bullets	yes (partial)	with bullets/ with bullets & commander

Table 3: Comparison of Results by Different Models with Human Data

Table 3 summarizes the results of blame assignment generated by different models, and compares these results with the dominant proportion (i.e., majority) of human agreement. (In *Scenario 4*, however, the dominant proportion overlaps with another category; in this case, if a model's prediction falls into the majority category, we regard it as a partial match). The simple cause model partially matches the human agreement in *Scenario 4*, but is inconsistent with the data in *Scenarios 1* to *3*. The simple authority model matches the human data in *Scenarios 2 and 3*, but is inconsistent with the data in other scenarios. In general, simple models use invariant approaches to the judgment problem. Therefore, they are insensitive to the changing social situations specified in each scenario. The *C&H* model matches human judgments in *Scenario 1*. In the remaining scenarios, the results show that their blame model does not match human data very well. These empirical findings show that our model approximates human judgments of responsibility and blame/credit and performs better than other computational approaches.

5.4.3 Comparison and Discussion

We briefly discuss how our model appraises each scenario and compare our approach with the *C&H* model.

Scenario 1. Actions and plans are explicitly represented in our approach. In *Scenario 1*, each marksman performs a primitive action, *shooting*. The action has a conditional effect, with the antecedent *live bullets* and the consequent *death*. All marksmen's shooting actions constitute a team plan *squad firing*, with the definite (goal) outcome *death* (*Figure 6*). The shooting actions are observed executed, and the outcome *death* occurs. As all the observed primitive actions of the marksmen match the team plan, we can certainly infer that the plan is pursued by the squad[10] (i.e., certain case of intention recognition). The marksmen are believed to intend the actions in the plan and the plan outcome (i.e. *death*).

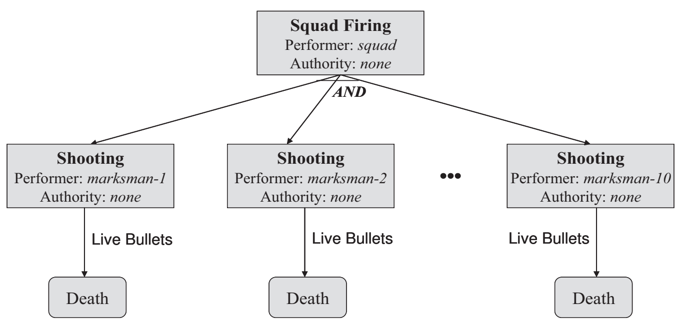


Figure 6: Team Plan for the Squad in Scenario 1

The marksman with the bullets is the sole *causal agent* for the death. This marksman intends the outcome, and thus deserves *high* degree of responsibility and blame. As other marksmen with blanks also intend the actions and the outcome, and shooting actions are observed executed but the antecedent of the conditional effect is false, their *failed attempt* can be detected. Therefore, other marksmen are also blameworthy for their attempt (recall that an unsuccessful attempt can be blamed or credited almost the same as a successful one, in *Section 3*).

The *C&H* model judges responsibility according to the actual cause of the event. As the marksman with the bullets is the only cause of the death, this marksman has degree of responsibility *1* for the death and others have degree of responsibility *0*. This result is

[10] Note that our intention recognition method is generally applied to a plan library and sequences of actions. This example is oversimplified.

inconsistent with human data. In determining blame, the *C&H* model draws the same conclusion as ours, but their approach is different. They consider each marksman's epistemic state before action performance (corresponding to foreknowledge). There are *10* situations possible, depending on who has the bullets. Each marksman is responsible for one situation (in which this marksman has bullets), with degree of responsibility *1*. Given that each situation is equally likely to happen (i.e., with possibility *1/10*), each marksman has degree of blame *1/10*.

As there is no notion of intention in their model, the *C&H* model uses foreknowledge as the only determinant for blame assignment. This is fine when there is no evidence of foreknowledge, as no foreknowledge entails no intention (*Rule C9*). When there is evidence of foreknowledge, however, the blame assigned is high, even if there is no intention manifested in the case. For example, in a context different from this example, if a marksman fires the gun by mistake, without any intention of causing or attempting the death, in the *C&H* model, this marksman will be blamed just the same as those who truly have such intention.

Scenarios 2&3. In our model, we take different forms of social interactions into account. The inference process reasons about beliefs from both causal and dialogue evidence. *Figure 7* illustrates the team plan of the squad in *Scenarios 2* and *3*, where a commander acts as an authority of the squad.

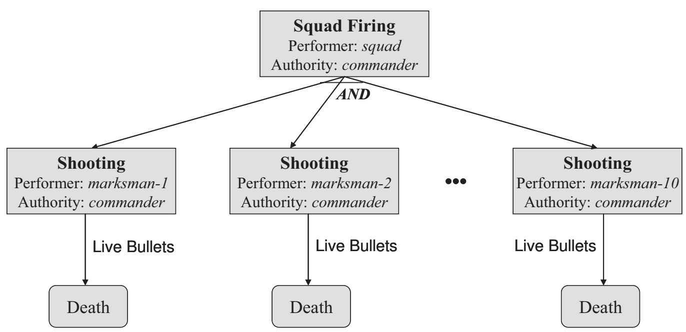


Figure 7: Team Plan for the Squad in Scenarios 2 and 3

The intermediate beliefs inferred from *Scenario 2* are given below. (The symbols *cmd*, *sqd* and *mkn* stand for the commander, the squad and the marksman with bullets, respectively. *t1<t1'<t2<t2'*.)

(1)	intend(*cmd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D5*)
(2)	obligation(*sqd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D6*)
(3)	intend(*cmd*, *death*, *t1'*)	(Belief *1*, Rule *C3*)
(4)	coerce(*cmd*, *sqd*, *squad-firing*, *t2'*)	(Act *accept* & Belief *2*, Rule *D9*)
(5)	coerce(*cmd*, *sqd*, *shooting*, *t2'*)	(Belief *4*, Rule *C13*)
(6)	coerce(*cmd*, *sqd*, *death*, *t2'*)	(Belief *4*, Rule *C14*)
(7)	coerce(*cmd*, *mkn*, *death*, *t2'*)	(Belief *5*, Rules *C14* & *C18*)

So in *Scenario 2*, the marksman causes the death due to coercion. The *commander* is responsible for the death. As the commander intends the outcome (*Belief 3*) and the severity of the outcome *death* is high, the commander is assigned *high* degree of responsibility and blamed with *high* intensity.

Scenario 3 includes a sequence of negotiation acts. The derived beliefs thus change to the following (*t4*<*t4'*):

(1)	intend(*cmd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D5*)
(2)	obligation(*sqd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D6*)
(3)	intend(*cmd*, *death*, *t1'*)	(Belief *1*, Rule *C3*)
(4)	¬intend(*sqd*, do(*sqd*, *squad-firing*), *t2'*)	(Act *reject*, Rule *D11*)
(5)	coerce(*cmd*, *sqd*, *squad-firing*, *t4'*)	(Act *accept* & Beliefs *2&4*, Rule *D10*)
(6)	coerce(*cmd*, *sqd*, *shooting*, *t4'*)	(Belief *5*, Rule *C13*)
(7)	coerce(*cmd*, *sqd*, *death*, *t4'*)	(Belief *5*, Rule *C14*)
(8)	coerce(*cmd*, *mkn*, *death*, *t4'*)	(Belief *6*, Rules *C14* & *C18*)

Clearly the marksmen do not intend firing (*Belief 4*). *Scenario 3* shows evidence of *strong* coercion. This is also reflected in the data. A greater proportion of subjects regard the commander as responsible and blameworthy in *Scenario 3* than in *Scenario 2*.

Assume *marksman-1* is the one with the live bullets. Using the *C&H* approach, the outcome is counterfactually dependent on *marksman-1*'s shooting, so *marksman-1*'s shooting is an actual cause of the death. Similarly, the commander's order is also an actual cause of the death. Based on the responsibility definition in the *C&H* model, both the commander and *marksman-1* are responsible for the death, and each has degree of responsibility *1*[11]. In assigning blame, there are ten situations altogether, and in each situation, the commander has expected responsibility *1*, so the commander is to blame with degree *1*. The marksmen each have degree of blame *1/10*. Thus the *C&H* model appraises that the commander and all marksmen are blameworthy for the outcome.

The *C&H* model represents all the relevant events in the scenarios as random variables. Thus, if we want to model the communicative acts in *Scenarios 2* and *3* using their approach, each act must be represented as a separate variable in their model (or a number of speech acts can be clumped together and represented as one variable). As conversational dialogue involves flexible contents and orders of the acts, it is difficult to come up with structural equations and represent the relationships between the variables. If we ignore some of the communicative acts in between, intermediate beliefs conveyed by them will be lost.

Scenario 4. Unlike the previous scenarios, in *Scenario 4*, the bullets are not initially set before the scenario starts. The marksmen can choose to use either bullets or blanks before shooting. Firing is still the *joint action* of the squad, but there is no team plan or common goal for the squad. As the commander orders the joint action, shooting actions and conditional effects are coerced. However, as the antecedents are enabled by a self agent (i.e., the marksmen with bullets), the consequent *death* is not coerced. The inferred beliefs are as follows.

(1)	intend(*cmd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D5*)
(2)	obligation(*sqd*, do(*sqd*, *squad-firing*), *t1'*)	(Act *order*, Rule *D6*)
(3)	coerce(*cmd*, *sqd*, *squad-firing*, *t3'*)	(Act *accept* & Belief *2*, Rule *D9*)
(4)	coerce(*cmd*, *sqd*, *shooting*, *t3'*)	(Belief *3*, Rule *C13*)

[11] Halpern and Pearl (2005) provide a refined definition of causality, where only the contingencies with allowable settings are considered. Under this refined definition, the commander is the only responsible agent for the death. But the results of blame assignment remain the same in each scenario.

(5) ¬coerce(*cmd*, *mkn*, *death*, *t3'*) (Belief *4*, Rules *C14* & *C20*)

In this case, the commander is not responsible for the outcome, but rather, the marksmen who choose to use bullets and cause the death are responsible and blameworthy. *Figure 5* shows that in *Scenario 4*, people's judgments somehow diffuse. There is an overlap between blaming the marksmen with bullets and blaming both the commander and the marksmen with bullets. Nonetheless, the category our model predicts is clearly better than the other three.

The *C&H* model requires all the structural equations to be deterministic. In essence, their model could not handle alternative courses of action, which inherently have nondeterministic properties. One remedy for this is to push the nondeterminism into the setting of the context (see *Section 2* for the explanation of context). For example, in *Scenario 4*, they could build a causal model to let the context determine whether the bullets are live or blank for each marksman, and then have a probability distribution over contexts. After that, they can compute the probability of an actual cause. However, since these contexts are treated as background variables whose values are assigned by the modeler, their approach could not construct the internal reasoning process to automate the inference for alternative courses of actions.

6. General Discussion

Based on the well-founded psychological attribution theory, we have built a general computational model for social causality and responsibility judgment. Our model takes different forms of social interaction into account and considers both the actions of agents and the outcomes they produce. We make use of commonsense reasoning to infer beliefs from dialogue communication and task execution. Our model is based on the general representation commonly used in intelligent systems. Causal inference is a plan-based evaluation over this representation. Both the inferences of social attributions and the overall judgments by our model have shown strong empirical support with respect to human data and in comparison with the alternative approaches.

Although the examples in this paper have focused on negative consequences and blame judgment, our model is capable of both credit and blame judgments. Currently we use a uniform model for these two types of judgments. However, several researchers made a distinction between them. D'Arcy (1963) pointed out that the criteria for judging benefit (i.e., credit assignment) are stricter than those for judging harm (i.e., blame assignment). The empirical findings in the work of Knobe (2003b) also show *credit and blame asymmetry* in people's judgments of behavior. These findings suggest us to consider using an asymmetry model for credit and blame assignments in our future extension.

Subjects tended to assign *shared blame* to the individuals involved. In the firing squad scenario *1*, for example, a portion of the subjects mentioned that they think the marksmen actually make group decisions together, and so they should be collectively responsible for the outcome. Sometimes this is true even when the individual is not causally connected to the creditworthy or blameworthy event (e.g., the chairman is blamed in the company program scenario *1*). Some researchers' work is relevant to this. Norman and Reed (2010) provided a logic formalism to account for *delegation* and responsibility. Our model's representational and inferential mechanism has the potential to incorporate these extensions.

Although attribution theory emphasizes subjective interpretation of events, it is a general theory of layman's judgment of behavior. We start from the general principles identified by attribution theory. However, it is also well known that responsibility judgment is influenced by the perceiver's emotional states, interpersonal goals such as impression management (Mele, 2001), and dispositional differences such as personality. People are notoriously biased when describing their involvement in creditworthy or blameworthy events (Bradley, 1978). These *biases* reveal subjective needs and motivational influence of the perceiver on responsibility judgment. Related work carried out in our lab has explored the influence of individual difference in the explanation of social events by modeling different explanatory styles according to agents' personalities (Oh, Gratch, & Woo, 2007).

In this paper, we have focused on the computational modeling of social causality and responsibility judgment in the context of multi-agent interactions. We produce the *first* general computational framework for social causality and responsibility judgment based on psychological attribution theory. One major contribution of our work is the identification of commonsense knowledge about the derivation of attributions from inter-agent communication and task execution. Another contribution of our work is the empirical validation of the model using human data. By producing the model, we also propose the computational account of coercion and design the algorithm to describe the attribution process and responsibility judgment. Because of the interdisciplinary nature of this work, it also takes a first step toward cognitive modeling of human social intelligence and helps advance our understanding of the process and principles of human social inference.

For practical applications of this work, we have taken a semi-formal approach and implemented our model mainly as a production system. Previously, there have been several versions of implementations and improvements regarding this work. The model was first implemented within Soar architecture in the context of virtual training environment described earlier. As in the virtual training system, the model was closely coupled with other system components using the blackboard representation, and belief update was handled using Soar's *JTMS* mechanism. We then moved to general-purpose programming language and implemented the inference engine in Java. The inference engine includes three parts: dialogue reasoner, intention recognizer and causal reasoner. We implemented dialogue inference rules and most of the causal inference rules in the model (*Rules C22-C25* were not implemented). Intention recognizer was implemented separately. Our experimental studies were based on the Java inference engine.

Other implementation and improvement efforts include the extension of the basic model in interactive environment by exploring different explanatory styles (Oh et al., 2007) and the improvement of the basic model by adding a model of negligence (Melissen, 2008). Tomai (2009) took the same attribution variables as ours and extended the basic model using qualitative process theory. His work translates attribution theory's implications of blame assignment into six views which impose ordinal constraints on blame assignment.

7. Conclusion

The social nature of computing is pervasive in every aspect of software research and development. With the advance of computer and communication technologies, social computing and intelligent system design will move toward emphasizing social intelligence (Wang et al., 2007). In this paper, we model a key aspect of social intelligence, by formalizing the underlying social reasoning process in people's behavioral judgment. We show how AI knowledge representation and reasoning methods can be utilized to automate social inference and judgment process. We also conduct human experiments to empirically validate our proposed model. The experimental results show that our model's predictions of the beliefs about intermediate variables, inferential mechanism and judgment results are consistent with people's responses. Therefore, our proposed model can be generally applied to the modeling of human-like social inference and behavioral judgment for intelligent entities.

Acknowledgments

We thank Jerry Hobbs, Paul Rosenbloom, Andrew Gordon, David Traum, Stephen Read, Joseph Halpern, Bernard Weiner and Joshua Knobe for the valuable discussions. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The work was supported in part by NNSFC grants #61175040, #71025001, #60921061, #70890084 and #91024030.

Appendix A. Computing Relevant Actions and Effects

Given the domain theory *DT*, an executed action set and a specific outcome *e*, the *relevant actions* to achieve *e* contain the following actions:

- The action *A* that causes *e* is relevant.
- The actions that enable a precondition of a relevant action to achieve *e* are relevant.
- If *e* is enabled by the consequent of a conditional effect of *A*, the actions that establish an antecedent of the conditional effect are relevant.
- If a precondition of a relevant action is enabled by the consequent of a conditional effect, the actions that establish an antecedent of the conditional effect are also relevant.

The preconditions of these relevant actions comprise the *relevant effects* to achieve *e*. Except for *e*, other effects of relevant actions are *side effects*.

If domain theory *DT* is confined to those actions, preconditions and effects in a specific *plan* (i.e., within the plan context), relevant actions and effects to achieve the *goal* of the plan can be derived based on the same computation as given above.

Appendix B. Computing Definite and Indefinite Effects

Let *A* be an action. If *A* is an abstract action and has only one decomposition, let a_i be a subaction of *A*. If *A* is an abstract action and has multiple decompositions, let a_i be a choice of *A*. The *definite effect set* of *A* is denoted as *effect*(*A*), and the *indefinite effect set* of *A* is denoted as *indefinite-effect*(*A*).

The definite effect set *effect*(A) is composed of those action effects, which occur in each way of decomposing A into primitive actions. It is computed recursively as follows:

- If A is a primitive action, *effect*(A) consists of all its action effects.
- If A is an abstract action and has only one decomposition, $effect(A) = \bigcup_{a_i \in subaction(A)} effect(a_i)$
- If A is an abstract action and has multiple decompositions, $effect(A) = \bigcap_{a_i \in choice(A)} effect(a_i)$

The indefinite effect set *indefinite-effect(A)* is composed of those action effects that only occur in some (but not all) ways of decomposing A into primitive actions. It is computed recursively as follows:

- If A is a primitive action, *indefinite-effect*(A) = ∅.
- If A is an abstract action and has only one decomposition,

$$indefinite-effect(A) = \bigcup_{a_i \in subaction(A)} indefinite-effect(a_i)$$

- If A is an abstract action and has multiple decompositions,

$$indefinite-effect(A) = \bigcup_{a_i \in choice(A)} (effect(a_i) \cup indefinite-effect(a_i)) - \bigcap_{a_i \in choice(A)} effect(a_i)$$

Appendix C. Inference Rules

For simplification, all universal quantifies are omitted. Variables x, y and z are different agents. Let s and h be a speaker and a hearer, p and q be propositions, and t, $t1$, …, $t4$ be time stamps. Let A, B and C be actions. Variable e is a state, denoting an action precondition, an effect, an antecedent or a consequent of a conditional effect. All the rules are from a perceiving agent's perspective.

Dialogue Inference Rules

D1 [inform]:
inform(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_1 \Rightarrow$ know(s, p, $t2$)

D2 [inform-grounded]:
inform(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_2 \Rightarrow$ know(h, p, $t2$)

D3 [request]:
request(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_3 \Rightarrow$ want(s, p, $t2$)

D4 [superior-request]:
request(s, h, p, $t1$) $\wedge$ superior(s, h) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_4 \Rightarrow$ obligation(h, p, s, $t2$)

D5 [order]:
order(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_5 \Rightarrow$ intend(s, p, $t2$)

D6 [order]:
order(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_6 \Rightarrow$ obligation(h, p, s, $t2$)

D7 [accept]:

$\neg$obligation(h, p, s, $t1$) $\wedge$ accept(h, p, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_7$ $\Rightarrow$ intend(h, p, $t3$)

D8 [willing-accept]:
want(h, p, $t1$) $\wedge$ accept(h, p, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_8$ $\Rightarrow$ intend(h, p, $t3$)

D9 [accept-obligation]:
$\neg(\exists t1)(t1<t3$ $\wedge$ intend(h, p, $t1$)) $\wedge$ obligation(h, p, s, $t2$) $\wedge$ accept(h, p, $t3$) $\wedge$ $t2<t3<t4$ $\wedge$ etc$_9$ $\Rightarrow$ coerce(s, h, p, $t4$)

D10 [unwilling-accept-obligation]:
$\neg$intend(h, p, $t1$) $\wedge$ obligation(h, p, s, $t2$) $\wedge$ accept(h, p, $t3$) $\wedge$ $t1<t3$ $\wedge$ $t2<t3<t4$ $\wedge$ etc$_{10}$ $\Rightarrow$ coerce(s, h, p, $t4$)

D11 [reject]:
reject(h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{11}$ $\Rightarrow$ $\neg$intend(h, p, $t2$)

D12 [counter-propose]:
counter-propose(h, A, B, s, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{12}$ $\Rightarrow$ know(h, alternative(A, B), $t2$)

D13 [counter-propose-grounded]:
counter-propose(h, A, B, s, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{13}$ $\Rightarrow$ know(s, alternative(A, B), $t2$)

D14 [counter-propose]:
counter-propose(h, p, q, s, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{14}$ $\Rightarrow$ $\neg$intend(h, p, $t2$)

D15 [counter-propose]:
counter-propose(h, p, q, s, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{15}$ $\Rightarrow$ want(h, q, $t2$)

D16 [know-alternative-request]:
know(s, alternative(A, B), $t1$) $\wedge$ request(s, h, do(z, A), $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_{16}$ $\Rightarrow$ $\neg$intend(s, do(z, B), $t3$)

D17 [know-alternative-order]:
know(s, alternative(A, B), $t1$) $\wedge$ order(s, h, A, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_{17}$ $\Rightarrow$ $\neg$intend(s, do(h, B), $t3$)

Causal Inference Rules

C1 [cause-action-effect]:
execute(x, A, $t1$) $\wedge$ $e\in$effect(A) $\wedge$ occur(e, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_{18}$ $\Rightarrow$ cause(x, e, $t3$)

C2 [cause-relevant-effect]:
cause(y, e', $t1$) $\wedge$ $e'\in$relevant-effect(e, DT) $\wedge$ cause(x, e, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc$_{19}$ $\Rightarrow$ assist-cause(y, x, e, $t3$)

C3 [intend-action]:
intend(x, do(z, A), $t1$) $\wedge$ $\neg(\exists y)$coerce(y, x, A, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc$_{20}$ $\Rightarrow$ $\exists e(e\in$effect(A) $\wedge$ intend(x, e, $t2$))

C4 [intend-one-alternative]:
intend(x, do(z, A), $t1$) $\wedge$ $\neg$intend(x, do(z, B), $t1$) $\wedge$ $\neg(\exists y)$coerce(y, x, A, $t1$) $\wedge$ alternative(A, B) $\wedge$ effect(A)$\subset$effect(B) $\wedge$ $t1<t2$ $\wedge$ etc$_{21}$ $\Rightarrow$ $\exists e(e\notin$effect(A) $\wedge$ $e\in$effect(B) $\wedge$ $\neg$intend(x, e, $t2$))

C5 [intend-one-alternative]:

intend(x, do(z, A), $t1$) $\wedge$ $\neg$intend(x, do(z, B), $t1$) $\wedge$ $\neg(\exists y)$coerce(y, x, A, $t1$) $\wedge$ alternative(A, B) $\wedge$ effect(B)$\subset$effect(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{22} \Rightarrow \exists e$($e\in$effect($A$) $\wedge$ $e\notin$effect(B) $\wedge$ intend(x, e, $t2$))

C6 [intend-plan]:
intend(x, by($plan$, $goal$), $t1$) $\wedge$ $A\in$relevant-action($goal$, $plan$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{23} \Rightarrow$ intend(x, A, $t2$)

C7 [intend-plan]:
intend(x, by($plan$, $goal$), $t1$) $\wedge$ $e\in$relevant-effect($goal$, $plan$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{24} \Rightarrow$ intend(x, e, $t2$)

C8 [intend-plan]:
intend(x, by($plan$, $goal$), $t1$) $\wedge$ $e\in$side-effect($goal$, $plan$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{25} \Rightarrow \neg$intend($x$, e, $t2$)

C9 [intent-foreknowledge-relation]:
intend(x, by(A, e), $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{26} \Rightarrow$ know(x, bring-about(A, e), $t2$)

C10 [foreknowledge-performer]:
$e\in$effect(A) $\wedge$ $\text{etc}_{27} \Rightarrow$ know(performer(A), bring-about(A, e), t)

C11 [foreknowledge-authority]:
$e\in$effect(A) $\wedge$ $\text{etc}_{28} \Rightarrow$ know(authority(A), bring-about(A, e), t)

C12 [coerce-primitive]:
coerce(y, x, A, $t1$) $\wedge$ primitive(A) $\wedge$ $e\in$effect(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{29} \Rightarrow$ coerce(y, x, e, $t2$)

C13 [coerce-non-decision-node]:
coerce(y, x, A, $t1$) $\wedge$ and-node(A) $\wedge$ $B\in$subaction(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{30} \Rightarrow$ coerce(y, x, B, $t2$)

C14 [coerce-non-decision-node]:
coerce(y, x, A, $t1$) $\wedge$ and-node(A) $\wedge$ $e\in$effect(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{31} \Rightarrow$ coerce(y, x, e, $t2$)

C15 [coerce-decision-node]:
coerce(y, x, A, $t1$) $\wedge$ or-node(A) $\wedge$ $B\in$choice(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{32} \Rightarrow \neg$coerce($y$, x, B, $t2$)

C16 [coerce-decision-node]:
coerce(y, x, A, $t1$) $\wedge$ or-node(A) $\wedge$ $e\in$effect(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{33} \Rightarrow$ coerce(y, x, e, $t2$)

C17 [coerce-decision-node]:
coerce(y, x, A, $t1$) $\wedge$ or-node(A) $\wedge$ $e\in$indefinite-effect(A) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{34} \Rightarrow \neg$coerce($y$, x, e, $t2$)

C18 [coerce-conditional-effect-initial-antecedent-true]:
$e\in$conditional-effect(A) $\wedge$ true(antecedent(e), $t1$) $\wedge$ coerce(y, x, e, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{35} \Rightarrow$ coerce(y, x, consequent(e), $t3$)

C19 [coerce-conditional-effect-initial-antecedent-false]:
$e\in$conditional-effect(A) $\wedge$ $\neg$true(antecedent(e), $t1$) $\wedge$ coerce(y, x, e, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{36} \Rightarrow$ $\neg$coerce(y, x, consequent(e), $t3$)

C20 [coerce-conditional-effect-self-establish-antecedent]:
$e\in$conditional-effect(A) $\wedge$ coerce(y, x, e, $t1$) $\wedge$ enable(x, antecedent(e), $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{37} \Rightarrow$ $\neg$coerce(y, x, consequent(e), $t3$)

C21 [coerce-conditional-effect-other-establish-antecedent]:
$e\in$conditional-effect(A) $\wedge$ coerce(y, x, e, $t1$) $\wedge$ enable(z, antecedent(e), $t2$) $\wedge$ $\neg$can-enable(x, $\neg$antecedent(e), $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{38} \Rightarrow$ coerce($y\cup z$, x, consequent(e), $t3$)

C22 [coerce-decision-node-initial-one-alternative]:
$A \in$choice(C) $\wedge$ true(precondition(A), $t1$) $\wedge$ ($B \in$choice(C)$\wedge B \neq A \Rightarrow \neg$true(precondition($B$), $t1$) $\wedge\neg$can-enable(x, precondition(B), $t1$)) $\wedge$ coerce(y, x, C, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{39} \Rightarrow$ coerce(y, x, A, $t3$)

C23 [coerce-decision-node-self-enable-alternative]:
coerce(y, x, C, $t1$) $\wedge$ $A \in$choice(C) $\wedge$ enable(x, precondition(A), $t2$) $\wedge$ ($B \in$choice(C)$\wedge B \neq A \Rightarrow$ $\neg$true(precondition(B), $t2$)$\wedge\neg$can-enable(x, precondition(B), $t2$)) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{40} \Rightarrow$ coerce(y, x, A, $t3$)

C24 [coerce-decision-node-other-enable-alternative]:
coerce(y, x, C, $t1$) $\wedge$ $A \in$choice(C) $\wedge$ enable(z, precondition(A), $t2$) $\wedge$ ($B \in$choice(C)$\wedge B \neq A \Rightarrow$ $\neg$true(precondition(B), $t2$)$\wedge\neg$can-enable(x, precondition(B), $t2$)) $\wedge$ $t1<t2<t3$ $\wedge$ $\text{etc}_{41} \Rightarrow$ coerce($y \cup z$, x, A, $t3$)

C25 [coerce-decision-node-disable-other-alternative]:
coerce(y, x, C, $t1$) $\wedge$ $A \in$choice(C) $\wedge$ true(precondition(A), $t2$) $\wedge$ ($B \in$choice(C)$\wedge B \neq A \Rightarrow$ enable(z, $\neg$precondition(B), $t3$)$\wedge\neg$can-enable(x, precondition(B), $t3$)) $\wedge$ $t1<t3<t4$ $\wedge$ $t2<t4$ $\wedge$ $\text{etc}_{42} \Rightarrow$ coerce($y \cup z$, x, A, $t4$)

C26 [coerce-intend-relation]:
coerce(y, x, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ $\text{etc}_{43} \Rightarrow$ intend(x, p, $t2$)

Appendix D. Company Program Scenarios

Scenario 1:

E1 The vice president of Beta Corporation goes to the chairman of the board and requests, "Can we start a new program?"
E2 The vice president continues, "The new program will help us increase profits,
E3 and according to our investigation report, it has no harm to the environment."
E4 The chairman answers, "Very well."
E5 The vice president executes the new program.
E6 However, the environment is harmed by the new program.

<u>Questions</u>:

1. Does the vice president want to start the new program?

Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

2. Does the chairman intend to start the new program?

Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

3. Is it the chairman's intention to increase profits?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

4. Does the vice president know that the new program will harm the environment?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

5. Is it the vice president's intention to harm the environment by starting the new program?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

6. How much would you blame the individuals for harming the environment?
Blame the chairman: 1 2 3 4 5 6
Blame the vice president: 1 2 3 4 5 6
Little Lots

Scenario 2:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.
E2 The vice president says, "The new program will help us increase profits,
E3 but according to our investigation report, it will also harm the environment."
E4 The chairman answers, "I only want to make as much profit as I can. Start the new program!"
E5 The vice president says, "Ok," and executes the new program.
E6 The environment is harmed by the new program.

<u>Questions:</u>

1. Does the chairman know that the new program will harm the environment?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

2. Does the chairman intend to start the new program?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

3. Is it the chairman's intention to increase profits?

Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

4. Is it the chairman's intention to harm the environment?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

5. Is the vice president coerced to start the new program (i.e. by the obligation of obeying the chairman)?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

6. How much would you blame the individuals for harming the environment?
Blame the chairman: 1 2 3 4 5 6
Blame the vice president: 1 2 3 4 5 6
Little Lots

Scenario 3:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.
E2 The vice president says, "The new program will help us increase profits,
E3 but according to our investigation report, it will also harm the environment.
E4 Instead, we should run an alternative program, that will gain us fewer profits than this new program, but it has no harm to the environment."
E5 The chairman answers, "I only want to make as much profit as I can. Start the new program!"
E6 The vice president says, "Ok," and executes the new program.
E7 The environment is harmed by the new program.

<u>Questions:</u>

1. Does the chairman know the alternative of the new program?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

2. Which program is the vice president willing to start?
Your answer: New program Alternative program
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

3. Is the vice president coerced to start the new program?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

4. Is the vice president coerced to harm the environment?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6 E7

5. How much would you blame the individuals for harming the environment?
Blame the chairman: 1 2 3 4 5 6
Blame the vice president: 1 2 3 4 5 6
Little Lots

Scenario 4:

E1 The chairman of Beta Corporation is discussing a new program with the vice president of the corporation.
E2 The vice president says, "There are two ways to run this new program, a simple way and a complex way.
E3 Both will equally help us increase profits, but according to our investigation report, the simple way will also harm the environment."
E4 The chairman answers, "I only want to make as much profit as I can. Start the new program either way!"
E5 The vice president says, "Ok," and chooses the simple way to execute the new program.
E6 The environment is harmed.

Questions:

1. Is the vice president coerced by the chairman to increase profits?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

2. Is the vice president coerced by the chairman to choose the simple way?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

3. Is the vice president coerced by the chairman to harm the environment?
Your answer: Yes No
Your confidence: 1 2 3 4 5 6
Low High
Based on which information (circle all that apply)? E1 E2 E3 E4 E5 E6

4. How much would you blame the individuals for harming the environment?

Blame the chairman:	1	2	3	4	5	6
Blame the vice president:	1	2	3	4	5	6
	Little					Lots

Appendix E. Belief Derivation of Company Program Scenarios

The symbols *chm* and *vp* refer to the chairman and the vice president, respectively. Time stamps $t1<t1'<t2<t2'<\ldots<t4<t5$. The severity of the outcome *environmental harm* is set to *medium*.

Scenario 1

Information Encoding:

E1 request(*vp*, *chm*, do(*vp*, *new-program*), *t1*)
E2 inform(*vp*, *chm*, bring-about(*new-program*, *profit-increase*), *t2*)
E3 inform(*vp*, *chm*, ¬bring-about(*new-program*, *env-harm*), *t2*)
E4 accept(*chm*, do(*vp*, *new-program*), *t3*)
E5 execute(*vp*, *new-program*, *t4*)
E6 *env-harm*∈effect(*new-program*); occur(*env-harm*, *t5*)

Question 1 (**Rule D3** [request]):
request(*vp*, *chm*, do(*vp*, *new-program*), *t1*)
⇒ want(*vp*, do(*vp*, *new-program*), *t1'*)

Question 2 (**Rule D7** [accept]):
accept(*chm*, do(*vp*, *new-program*), *t3*)
⇒ intend(*chm*, do(*vp*, *new-program*), *t3'*)

Question 3 (**Rule C3** [intend-action]):
intend(*chm*, do(*vp*, *new-program*), *t3'*) ∧ ¬coerce(*vp*, *chm*, *new-program*, *t3'*)
⇒ *profit-increase*∈effect(*new-program*) ∧ intend(*chm*, *profit-increase*, *t3'*)

Question 4 (**Rule D1** [inform]):
inform(*vp*, *chm*, ¬bring-about(*new-program*, *env-harm*), *t2*)
⇒ know(*vp*, ¬bring-about(*new-program*, *env-harm*), *t2'*)
⇒ ¬know(*vp*, bring-about(*new-program*, *env-harm*), *t2'*)

Question 5 (**Rule C9** [intent-foreknowledge-relation]):
¬know(*vp*, bring-about(*new-program*, *env-harm*), *t2'*)
⇒ ¬intend(*vp*, by(*new-program*, *env-harm*), *t2'*)

Question 6 (**Attribution Algorithm**):
Primary-responsible agent: *vp*
Degree of responsibility/Intensity of blame: *low*

Scenario 2

Information Encoding:

E2 inform(*vp*, *chm*, bring-about(*new-program*, *profit-increase*), *t1*)

E3 inform(*vp*, *chm*, bring-about(*new-program*, *env-harm*), *t1*)
E4 goal(*chm*, *profit-increase*); order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
E5 accept(*vp*, do(*vp*, *new-program*), *t3*); execute(*vp*, *new-program*, *t3*)
E6 occur(*env-harm*, *t4*)

Question 1 (**Rule D2** [inform-grounded]):
inform(*vp*, *chm*, bring-about(*new-program*, *env-harm*), *t1*)
⇒ know(*chm*, bring-about(*new-program*, *env-harm*), *t1'*)

Question 2 (**Rule D5** [order]):
order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
⇒ intend(*chm*, do(*vp*, *new-program*), *t2'*)

Question 3 (**Rule C7** [intend-plan]):
intend(*chm*, by(*new-program*, *profit-increase*), *t2'*) ∧ *profit-increase*∈relevant-effect(*profit-increase*, *new-program*)
⇒ intend(*chm*, *profit-increase*, *t2'*)

Question 4 (**Rule C8** [intend-plan]):
intend(*chm*, by(*new-program*, *profit-increase*), *t2'*) ∧ *env-harm*∈side-effect(*profit-increase*, *new-program*)
⇒ ¬intend(*chm*, *env-harm*, *t2'*)

Question 5 (**Rules D6** [order] & **D9** [accept-obligation]):
order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
⇒ obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*)
obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*) ∧ accept(*vp*, do(*vp*, *new-program*), *t3*)
⇒ coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*)

Question 6 (**Attribution Algorithm**):
Primary-responsible agent: *chm*
Degree of responsibility/Intensity of blame: *low*

Scenario 3

Information Encoding:

E2 inform(*vp*, *chm*, bring-about(*new-program*, *profit-increase*), *t1*)
E3 inform(*vp*, *chm*, bring-about(*new-program*, *env-harm*), *t1*)
E4 counter-propose(*vp*, do(*vp*, *new-program*), do(*vp*, *alternative-program*), *chm*, *t1*)
E5 goal(*chm*, *profit-increase*); order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
E6 accept(*vp*, do(*vp*, *new-program*), *t3*); execute(*vp*, *new-program*, *t3*)
E7 occur(*env-harm*, *t4*)

Question 1 (**Rule D13** [counter-propose-grounded]):
counter-propose(*vp*, do(*vp*, *new-program*), do(*vp*, *alternative-program*), *chm*, *t1*)
⇒ know(*chm*, alternative(*new-program*, *alternative-program*), *t1'*)

Question 2 (**Rules D14** & **D15** [counter-propose]):
counter-propose(*vp*, do(*vp*, *new-program*), do(*vp*, *alternative-program*), *chm*, *t1*)
⇒ ¬intend(*vp*, do(*vp*, *new-program*), *t1'*)

counter-propose(*vp*, do(*vp*, *new-program*), do(*vp*, *alternative-program*), *chm*, *t1*)
⇒ want(*vp*, do(*vp*, *alternative-program*), *t1'*)

Question 3 (**Rules D6** [order] & **D10** [unwilling-accept-obligation]):
order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
⇒ obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*)
¬intend(*vp*, do(*vp*, *new-program*), *t1'*) ∧ obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*) ∧ accept(*vp*, do(*vp*, *new-program*), *t3*)
⇒ coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*)

Question 4 (**Rule C12** [coerce-primitive]):
coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*) ∧ primitive(*new-program*) ∧ *env-harm*∈effect(*new-program*)
⇒ coerce(*chm*, *vp*, *env-harm*, *t3'*)

Question 5 (**Attribution Algorithm**):
Primary-responsible agent: *chm*
Degree of responsibility/Intensity of blame: *low*

Scenario 4

Information Encoding:

E2 inform(*vp*, *chm*, or-node(*new-program*), *t1*)
inform(*vp*, *chm*, *simple-way*∈choice(*new-program*), *t1*)
inform(*vp*, *chm*, *complex-way*∈choice(*new-program*,), *t1*)
E3 inform(*vp*, *chm*, bring-about(*simple-way*, *profit-increase*), *t1*)
inform(*vp*, *chm*, bring-about(*complex-way*, *profit-increase*), *t1*)
inform(*vp*, *chm*, bring-about(*simple-way*, *env-harm*), *t1*)
E4 goal(*chm*, *profit-increase*); order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
E5 accept(*vp*, do(*vp*, *new-program*), *t3*); intend(*vp*, *simple-way*, *t3*); ¬intend(*vp*, *complex-way*, *t3*);
execute(*vp*, *simple-way*, *t4*)
E6 occur(*env-harm*, *t5*)

Question 1 (**Rule C16** [coerce-decision-node]):
order(*chm*, *vp*, do(*vp*, *new-program*), *t2*)
⇒ obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*)
obligation(*vp*, do(*vp*, *new-program*), *chm*, *t2'*) ∧ accept(*vp*, do(*vp*, *new-program*), *t3*)
⇒ coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*)
coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*) ∧ or-node(*new-program*) ∧ *profit-increase*∈effect(*new-program*)
⇒ *coerce*(*chm*, *vp*, *profit-increase*, *t3'*)

Question 2 (**Rule C15** [coerce-decision-node]):
coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*) ∧ or-node(*new-program*) ∧ *simple-way*∈choice(*new-program*)
⇒ ¬coerce(*chm*, *vp*, *simple-way*, *t3'*)

Question 3 (**Rule C17** [coerce-decision-node]):

coerce(*chm*, *vp*, do(*vp*, *new-program*), *t3'*) $\wedge$ or-node(*new-program*) $\wedge$ *env-harm*$\in$indefinite-effect(*new-program*)
$\Rightarrow$ $\neg$coerce(*chm*, *vp*, *env-harm*, *t3'*)

Question 4 (**Attribution Algorithm**):
Primary-responsible agent: *vp*
Degree of responsibility/Intensity of blame: *high*

Appendix F. Firing Squad Scenarios

Scenario 1

Suppose that there is a firing squad consisting of ten excellent marksmen. Only one of them has live bullets in his rifle; the rest have blanks. The marksmen do not know which of them has the live bullets. The marksmen shoot at the prisoner and he dies.

Scenario 2

Suppose that there is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide by their leader's commands. Only one of them has live bullets in his rifle; the rest have blanks. The commanding officer and the marksmen do not know which marksman has the live bullets. The commander orders the marksmen to shoot the prisoner. The marksmen shoot at the prisoner and he dies.

Scenario 3

Suppose that there is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide by their leader's commands. Only one of them has live bullets in his rifle; the rest have blanks. The commanding officer and the marksmen do not know which marksman has the live bullets. The commander orders the marksmen to shoot the prisoner. The marksmen refuse the order. The commander insists that the marksmen shoot the prisoner. The marksmen shoot at the prisoner and he dies.

Scenario 4

Suppose that there is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide by their leader's commands. The commanding officer orders the marksman to shoot the prisoner, and each marksman can choose to use either blanks or live bullets. The commander and the marksmen do not know whether other marksmen have live bullets. By tradition, if the prisoner lives (i.e., everyone chooses blanks), he is set free. The marksmen shoot at the prisoner and he dies.

<u>Questions</u> (in *Scenario 1*, *Questions 1-3* only contain selections *a* and *b*):

1. Who physically *caused* the death?
   a) the marksmen who had live bullets in their rifles
   b) all the marksmen in the firing squad
   c) the commanding officer
   d) a) and c)

e) everybody
f) none of the above

2. Who would you think is *responsible* for the death?
   a) the marksmen who had live bullets in their rifles
   b) all the marksmen in the firing squad
   c) the commanding officer
   d) a) and c)
   e) everybody
   f) none of the above

3. Who deserves *blame* for the death?
   a) the marksmen who had live bullets in their rifles
   b) all the marksmen in the firing squad
   c) the commanding officer
   d) a) and c)
   e) everybody
   f) none of the above

4. In making your judgment, do you feel the marksmen were *coerced*?
   a) there was strong coercion
   b) there was weak coercion
   c) there was no coercion

References

Aleven, V., & Ashley, K. D. (1995). Doing Things with Factors. *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*.

Allen, J. F., & Perrault, C. R. (1980). Analyzing Intention in Utterances. *Artificial Intelligence*, 15(3):143-178.

Austin, J. (1962). *How to Do Things with Words*. Harvard University Press.

Blythe, J. (1999). Decision-Theoretic Planning. *AI Magazine*, 20(2):37-54.

Bradley, G. W. (1978). Self-Serving Biases in the Attribution Process: A Reexamination of the Fact or Fiction Question. *Journal of Personality and Social Psychology*, 36(1):56-71.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Intelligence*, 22(2):249-254.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.) (2000). *Embodied Conversational Agents*. Cambridge University Press.

Castelfranchi, C. (1990). Social Power. *Proceedings of the First European Workshop on Modeling Autonomous Agents in a Multi-Agent World*.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22:93-115.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on Contributions to Conversation. *Language and Cognitive Processes*, 2:1-23,.

Cohen, P. R., & Levesque, H. J. (1990). Intention is Choice with Commitment. *Artificial Intelligence*, 42(2-3):213-261.

D'Arcy, E. (1963). *Human Acts: An Essay in Their Moral Evaluation*. Oxford: Clarendon.

Di Eugenio, B., & Glass, M. (2004). The Kappa Statistic: A second Look. *Computational Linguistics*, 30(1):95-101.

d'Inverno, M., Kinny, D., Luck, M., & Wooldridge, M. (1997). A Formal Specification of dMARS. In: M. P. Singh, A. Rao and M. J. Wooldridge (Eds.). *Intelligent Agents IV*, pp. 155-176. Springer-Verlag.

Erol, K., Hendler, J., & Nau, D. S. (1994). UMCP: A Sound and Complete Procedure for Hierarchical Task-Network Planning. *Proceedings of the Second International Conference on Artificial Intelligence Planning Systems*.

Ferguson, G., & Allen, J. (2007). Mixed-Initiative Dialogue Systems for Collaborative Problem-Solving. *AI Magazine*, 28(2):23-32.

Fikes, R.E., & Nilsson, N. J. (1971). STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence*, 2(3-4).

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of Responsibility: From Man the Scientist to Man as Lawyer. In: L. Berkowitz (Ed.). *Advances in Experimental Social Psychology (Vol. 13)*, pp. 81-138. Academic Press.

Fischer, K., Mueller, J. P., & Pischel, M. (1996). A Pragmatic BDI Architecture. In: M. Wooldridge, J. P. Mueller and M. Tambe (Eds.). *Intelligent Agents II*, pp. 203-218. Springer-Verlag.

Georgeff, M. P., & Lansky, A. L. (1987). Reactive Reasoning and Planning. *Proceedings of the Sixth National Conference on Artificial Intelligence*.

Gil, Y., Deelman, E., Blythe, J., Kesselman, C., & Tangmurarunkit, H. (2004). Artificial Intelligence and Grids: Workflow Planning and Beyond. *IEEE Intelligent Systems*, 19(1):26-33.

Golbeck, J., & Hendler, J. (2006). Inferring Binary Trust Relationships in Web-Based Social Networks, *ACM Transactions on Internet Technology,* 6(4):497-529.

Gordon, A., & Hobbs, J. R. (2004). Formalizations of Commonsense Psychology. *AI Magazine*, 25(4):49-62.

Gratch, J., & Mao, W. (2003). Automating After Action Review: Attributing Blame or Credit in Team Training. *Proceedings of the Twelfth Conference on Behavior Representation in Modeling and Simulation*.

Gratch, J., Mao, W., & Marsella, S. (2006). Modeling Social Emotions and Social Attributions. In: R. Sun (Ed.). *Cognition and Multi-Agent Interaction*, pp. 219-251. Cambridge University Press.

Gratch, J., Marsella, S., & Petta, P. (2009). Modeling the Antecedents and Consequences of Emotion. *Journal of Cognitive Systems Research*, 10(1):1-5.

Grice, H. P. (1975). Logic and Conversation. In: P. Cole and J. Morgan (Eds.). *Syntax and Semantics: Vol 3, Speech Acts*. Academic Press.

Grosz, B., & Kraus, S. (1996). Collaborative Plans for Complex Group Action. *Artificial Intelligence*, 86(2):269-357.

Grosz, B. J., & Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175-204.

Hage, J. C. (1997). *Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying logic*. Kluwer Academic Publishers.

Halpern, J. Y., & Pearl, J. (2001). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *Proceedings of the Seventeenth Conference in Uncertainty in Artificial Intelligence*.

Halpern, J. Y., & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for Philosophy of Science*, 56(4):843-887.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons Inc.

Hilton, D. J. (1990). Conversational Processes and Causal Explanation. *Psychological Bulletin*, 107:65-81.

Hobbs, J. R. (1985). Ontological Promiscuity. *Proceedings of the Twenty-Third Annual Meeting of the Association for Computational Linguistics*.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as Abduction. *Artificial Intelligence*, 63(1-2):69-142.

Huber, M. J. (1999). JAM: A BDI-Theoretic Mobile Agent Architecture. *Proceedings of the Third International Conference on Autonomous Agents*.

Jaimes, A., Sebe, N., & Gatica-Perez, D. (2006). Human-Centered Computing: A Multimedia Perspective. *Proceedings of the Fourteenth Annual ACM International Conference on Multimedia*.

Jennings, N. R. (1992). On Being Responsible. In: E. Werner and Y. Demazeau (Eds.). *Decentralized A.I.*, pp. 93-102. North Holland Publishers.

Johnson, C., & Gonzalez, A. J. (2008). Automated After Action Review: State-of-the-Art Review and Trends. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*. 5(2):108-121.

Jones, D. (2009). The Good, the Bad and the Intentional. *The Psychologist*, 22(8):666-669, August.

Kant, I. (1998). *Groundwork of the metaphysics of morals*. Cambridge University Press.

Kidd, R. F., & Amabile, T. M. (1981). Causal Explanations in Social Interaction: Some Dialogues on Dialogue. In: J. H. Harvey, W. J. Ickes and R. F. Kidd (Eds.). *New Directions in Attribution Research (Vol. 3)*, pp. 307-328. Lawrence Erlbaum Associates.

Knobe, J. (2003a). Intentional Action and Side-Effects in Ordinary Language. *Analysis*, 63:190-193.

Knobe, J. (2003b). Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology*, 16:309-324.

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning*, 30(2/3):271-274.

Kraus, S., Hoz-Weiss, P., & Wilkenfeld, J. (2008), Resolving Crises through Automated Bilateral Negotiations. *Artificial Intelligence*, 172(1).

Litman, D. J., & Allen, J. F. (1990). Discourse Processing and Commonsense Plans. In: P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in Communication*, pp.365-388. The MIT Press.

Lochbaum, K. E., Grosz, B. J., & Sidner, C. L. (2000). Discourse Structure and Intention Recognition. In: R. Dale, H. Moisl and H. Somers (Eds.), *Handbook of Natural Language Processing*, pp.123-146.

Malle, B. F. (2001). Attribution processes. In N. J. Smelser and P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences Vol. 14*, pp. 913-917. Elsevier.

Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, 33:101-121.

Mao, W., Gratch, J., & Li, X. (in press). Probabilistic Plan Inference for Group Behavior Prediction. *IEEE Intelligent Systems*.

Marinier, R. P., & Laird, J.E. (2004). Towards a Comprehensive Computational Model of Emotions and Feelings. *Proceedings of the Sixth International Conference on Cognitive Modeling*.

Marsella, S., & Gratch, J. (2009). EMA: A Process Model of Appraisal Dynamics. *Journal of Cognitive Systems Research*, 10(1): 70-90.

Martinovski, B., & Mao, W. (2009). Emotion as an Argumentation Engine: Modeling the Role of Emotion in Negotiation. *Group Decision and Negotiation*, 18(3):235-259.

Martinovski, B., Mao, W., Gratch, J., & Marsella, S. (2005). Mitigation Theory: An Integrated Approach. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*.

McCarty, L. T., & Sridharan, N. S. (1981). The Representation of an Evolving System of Legal Concepts: II. Prototypes and Deformations. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*.

McCarty, L. T. (1995). An Implementation of Eisner v. Macomber. *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*.

McCarty, L. T. (1997). Some Arguments about Legal Arguments. *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*.

Mele, A. R. (2001). *Self-Deception Unmasked*. Princeton University Press.

Melissen, A. (2008). *Exploring Neglected Avenues in the Modeling of Attribution Theory*. Master Thesis, Department of Human Media Interaction, University of Twente.

Mueller, E. (2006). *Commonsense Reasoning*. Morgan Kaufmann Publishers.

Nadelhoffer, T. (2006). On Trying to Save the Simple View. *Mind & Language*, 21(5):565-586, November.

Nau, D. S., Cao, Y., Lotem, A., & Muoz-Avila, H. (1999). SHOP: Simple Hierarchical Ordered Planner. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall.

Norman, T. J., & Reed, C. (2010). A Logic of Delegation and Responsibility. *Artificial Intelligence*, 174(1):51-71.

Oh, S., Gratch, J., & Woo, W. (2007). Explanatory Styles for Socially Interactive Agents. *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*.

Pearl, J. (1999). Reasoning with Cause and Effect. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

Perrault, C. R. (1990). An Application of Default Logic to Speech Act Theory. In: P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in Communication*, pp.161-186. The MIT Press.

Picard, R. W. (1997). *Affective Computing*. The MIT Press.

Picard, R. W. (2010). Affective Computing: From Laughter to IEEE. *IEEE Transactions on Affective Computing*, 1(1):11-17, January-June.

Pollack, M. E. (1990). Plans as Complex Mental Attitudes. In: P. R. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in Communication*, pp.77-103. The MIT Press.

Prakken, H. (1997). *Logic Tools for Modeling Legal Argument: A Study of Defeasible Argumentation in Law*. Kluwer Academic Publishers.

Prakken, H., & Sartor, G. (2002). The Role of Logic in Computational Models of Legal Argument. In: A.Kakas and F. Sadri (eds.). *Computational Logic: Logic Programming and Beyond, Essays in Honor of Robert A. Kowalski, Part II*, pp. 342-380. Springer-Verlag.

Rao, A. S. (1996). AgentSpeak(L): BDI Agents Speak out in a Logical Computable Language. In: W. Van de Velde and J. W. Perram (Eds.). *Agents Breaking Away: Proceedings of the Seventh European Workshop on Modeling Autonomous Agents in Multi-Agent World*, pp. 42-55. Springer-Verlag.

Rice J. A. (1994). *Mathematical Statistics and Data Analysis (Second Edition)*. Duxbury Press.

Rich, C., Sidner, C. L., & Lesh, N. (2001). COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction*. AI Magazine*, 22(4):15-26.

Rietveld, T., & van Hout. R. (1993). *Statistical Techniques for the Study of Language and Language Behavior*. Mouton de Gruyter.

Rissland, E. L., & Ashley, K. D. (1987). A Case-Based System for Trade Secrets Law. *Proceedings of the First International Conference on Artificial Intelligence and Law*.

Rissland, E. L., & Skalak, D. B. (1991). CABARET: Statutory Interpretation in a Hybrid Architecture. *International Journal of Man-Machine Studies*, 34:839-887.

Schurr, N., Marecki, J., Tambe, M., & Scerri, P. (2005). Towards Flexible Coordination of Human-Agent Teams. *Multiagent and Grid Systems*, 1(1):3-16.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Shaver, K. G. (1985). *The Attribution Theory of Blame: Causality, Responsibility and Blameworthiness*. Springer-Verlag.

Sichman, J. S., Conte, R., Demazeau, Y., & Castelfranchi, C. (1994). A Social Reasoning Mechanism Based on Dependence Networks. *Proceedings of the Eleventh European Conference on AI*.

Smith, I. A., & Cohen, P. R. (1996). Toward a Semantics for Agent Communications Language Based on Speech-Acts. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*.

Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J., & Traum, D. (2006). Toward Virtual Humans. *AI Magazine*, 27(2):96-108.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J., Gerten, J., Chu, S., & White, K. (2010). Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. *Proceedings of the Tenth International Conference on Intelligent Virtual Agents*.

Tomai, E. (2009). *A Pragmatic Approach to Computational Narrative Understanding*. Ph.D. Thesis, Electrical Engineering and Computer Science Department, Northwestern University.

Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Thesis, Computer Science Department, University of Rochester.

Traum, D., Gratch, J., Marsella, S., Lee, J., & Hartholt, A. (2008). Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents. *Proceedings of the Eighth International Conference on Intelligent Virtual Agents*.

Traum, D., Rickel, J., Gratch, J., & Marsella, S. (2003). Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training. *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*.

Veloso, M., Carbonell, J., Perez, A., Borrajo, D., Fink, E., & Blythe, J. (1995). Integrating Planning and Learning: the Prodigy Architecture. *Journal of Theoretical and Experimental Artificial Intelligence*, 7(1):81-120.

Wang, F., Zeng, D., Carley, K., & Mao, W. (2007). Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligent Systems*, 22(2):79-83.

Wang, F., Zeng, D., Hendler, J. A., Zhang Q., Feng, Z., Gao, Y., Wang, H., & Lai, G. (2010). A Study of the Human Flesh Search Engine: Crowd-Powered Expansion of Online Knowledge. *Computer*, 43(8):45-53.

Weiner, B. (1995). *The Judgment of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press.

Weiner, B. (2001). Responsibility for Social Transgressions: An Attributional Analysis. In: B. F. Malle, L. J. Moses and D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, pp. 331-344. The MIT Press.

Weiner, B. (2006). *Social Motivation, Justice and the Moral Emotions: An Attributional Approach*. Lawrence Erlbaum Associates.

Zimmerman, M. J. (1988). *An Essay on Moral Responsibility*. Rowman & Littlefield.