To appear in the International Conference on Interactive Virtual Agents, Kos Greece, 2005

Social Causality and Responsibility: Modeling and Evaluation

Wenji Mao Jonathan Gratch

Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292, U.S.A.
{AAA, BBB}@ict.usc.edu

Abstract. Causality is a central issue in many AI applications. Social causality, in contrast to physical causality, seeks to attribute cause and responsibility to social events, and accounts for how an intelligent entity makes sense of the social behavior of others. Modeling the underlying process and inferences of social causality can enrich the cognitive and social functionality of intelligent agents. In this paper, we present a general computational model of social causality and responsibility. Our model incorporates the basic features people use in their judgments, including physical causality, coercion, intention and foreknowledge. We propose commonsense reasoning of these features from plan knowledge and observation, and empirically evaluate and compare the model with several other models.

1 Introduction

Causality is a central issue in many AI applications. Illuminating the actual cause of an event is crucial for achieving fundamental understanding of the world and ultimately for acting on the world to achieve specific ends. Whereas computational approaches (e.g., [Pearl, 2000]) have successfully exploited theories of *physical causality* for the explanation of physical phenomena, these theories are simply inadequate for exploiting and explaining social phenomena. In contrast, *social causality*, both in theory and as practiced in everyday folk judgments and in the legal system, emphasizes multiple causal dimensions, incorporates epistemic variables, and distinguishes between cause, responsibility and blame.

Recent approaches to social causality have addressed some of these differences by extending causal models [Chockler & Halpern, 2004], although it is unclear whether a full accounting of social causality will (or even should) result from such extensions. In contrast, we start with social causality theory and consider how this could be formalized in a computational model. This allows intelligent entities to reason about aspects of social causality not addressed by these extended causal models and provides a complementary perspective to the enterprise of causal reasoning about social events.

Psychological and philosophical theories identify key variables that mediate determinations of social causality. In these theories, social causality involves not only physical causality, with an emphasis on human agency, but also people's freedom of choice (e.g., coercion [Shaver, 1985] and controllability [Weiner, 1995]), intentions and foreknowledge [Shaver, 1985; Zimmerman, 1988]. Using these variables, social causality makes several distinctions not present in the determinations of physical

cause. For example, an actor may physically cause an event, but be absolved of responsibility and blame. Or a person may be held responsible and blameworthy for what she did not physically cause.

Our goal is to model the underlying process and inferences of social causality to enrich the cognitive and social functionality of intelligent agents. Such a model can help an agent to explain the observed social behavior of others, which is crucial for successful interactions among social entities. It can enrich the design components of human-like agents, guide strategies of natural language conversation and model social emotions.

To achieve this end, we take individual agent's perspective and explore the commonsense interpretation of human social inference, based on the broad variables people use in determining social causality and responsibility. Psychological and philosophical theories largely agree on these basic variables though they differ in terminology. In this paper, we adopt the terminology of Shaver [1985]. In Shaver's model, the judgment process proceeds by assessing several key variables: who *caused* the event; Did the actor *foresee* the consequence; Did she *intend* to bring the consequence about; Did she have *choices* or act under *coercion* (e.g., by an authority)?

Though the theory identifies the conceptual variables for social causality and responsibility judgment, in modeling social behavior of intelligent agents, we cannot assume that an agent has privileged access to the mental states of other agents, but rather, an agent can only make inferences and judgment based on the evidence accessible in the computational system it situates. Current intelligent systems are increasingly sophisticated, usually involving natural language conversation, interactions of multiple agents and a planning module to plan for sequence of actions, with methods that explicitly model beliefs, desires and intentions of agents. All these should play a role in deriving the conceptual variables underlying the judgment of social causality.

In order to bridge the conceptual descriptions of the variables and the computational realization in application systems, we need to model the inferential mechanism that derives the variable values needed for the judgment from information and context available in practical systems. This paper presents a domain-independent computational model of social causality and responsibility by inferring the key variables from plan knowledge and communication. To assess the veracity of the approach in modeling human social inference, we conduct empirical study to evaluate and compare the model with several other models of responsibility and blame.

In the rest of the paper, we first introduce the judgment process and how the key variables are utilized in the process. We then present the computational model, including the representation, inferences and the implementation module. We finally evaluate the model using empirical data and compare our approach with the related work. As we take individual perspective in this work, we model the general judgment process based on the implications of social attribution theory, a body of research in social psychology exploring subjective explanation of behavior.

2 Judgment Process and Key Variables

We base our work on the most influential attributional models of Shaver [1985] and Weiner [1995] for social causality and responsibility. Their models suggest that physical causality and coercion determine *who* is responsible for some outcome under

evaluation, whereas mental factors, intention and foreseeability, determine the *degree* of responsibility and blame/credit.

Physical causality refers to the connection between actions and the effects they produce. In the absence of external coercion, the actor whose action directly produces the outcome is regarded as responsible. However, in social situations, an agent may cause an outcome because she could not have done otherwise. *Coercion* occurs when some external force, such as a more powerful individual or a socially sanctioned authority, limits an agent's freedom of choice. The presence of coercion can deflect some or all of the responsibility to the coercive force, depending on the perceived degree of coercion.

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. If an agent intends an action to achieve an outcome, then the agent must have the foreknowledge that the action brings about the outcome. *Foreseeability* refers to an agent's foreknowledge about actions and their consequences. The higher the degree of intention, the greater the responsibility assigned. The lower the degree of foreseeability, the less the responsibility assigned.

An agent may intentionally perform an action, but may not intend all the action effects. It is outcome intent (i.e., intentional action effect), rather than act intentionality (i.e., intentional action) that are key in responsibility judgment. Similar difference exists in act coercion (i.e., coerced action) and outcome coercion (i.e., coerced action effect). The result of the judgment process is the assignment of certain blame or credit to the responsible agents. The intensity of blame/credit is determined by the degree of responsibility as well as the severity/positivity of the outcome.

3 The Social Inference Model

The judgment of social causality and responsibility is a subjective process. It is from the perspective of a perceiving agent (i.e., the agent who makes the judgment), and based on the perceiver's interpretation of the significance of events. The perceiver uses own knowledge about the observed agents' behavior to infer certain beliefs (in terms of the key variables). The inferred variable values are then applied to the judgment process to form an overall result.

3.1 Modular Structure

Two important sources of evidence contribute to the inferences of key variables. One source is the causal evidence about the actions and effects of the observed agents. The other is the observations of the actions performed by the observed agents, including both physical and communicative acts (e.g., in a conversational dialogue). The inference process acquires beliefs from communicative events (i.e., dialogue inference) and from the causal information about the observed action execution (i.e., causal inference). To construct a computational model, we need to represent the information source, and identify the inferential mechanism over the representation. We also need an algorithm to describe the judgment process.

We have designed the modular structure for evaluating social causality and responsibility (i.e., a social inference module), and interface it with other system compo-

nents. *Figure 1* illustrates the structure of the module. It takes the observed communicative events and executed actions as inputs. Causal information and social information are also important inputs. Causal information includes an action theory and a plan library (discussed below). Social information specifies the power relationship of roles. The inference process first applies dialogue inference, and then causal inference. Both inferences make use of the commonsense heuristics, and derive beliefs about the variable values. The values are then served as inputs for the algorithm (refer to [Mao & Gratch, 2004a]), which determines responsibility, and assigns certain blame or credit to the responsible agents.

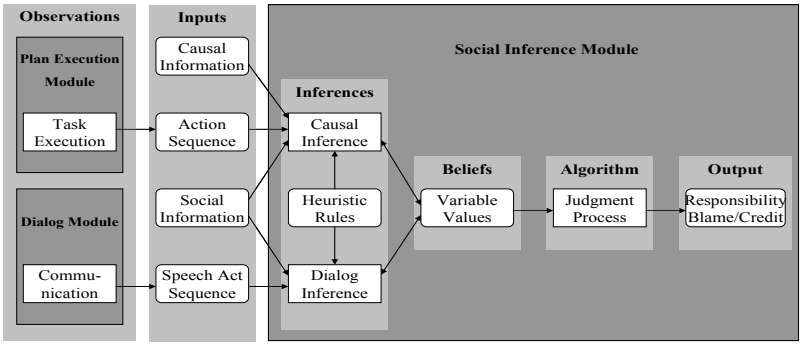

Figure 1 Structure of the Social Inference Module

3.2 Computational Representation

To represent an agent's causal information, we have adopted a hierarchical plan representation used in many intelligent agent systems. This representation provides a concise description of the physical causal relationship between events and world states. It also provides a clear structure for exploring alternative courses of actions and detecting plan interactions.

Actions and Plans

Physical causality is encoded via the hierarchical representation of actions and plans. *Actions* consist of a set of propositional preconditions and effects (including conditional effects). Each action step is either a *primitive* action (i.e., an action directly executable by some agent) or an *abstract* action. An abstract action may be decomposed hierarchically in multiple ways and each decomposition consists of a sequence of primitive and/or abstract sub-actions. Each decomposition is referred to as an action *alternative*. If an abstract action has only one decomposition, its effects are the union of those of its sub-actions; otherwise, its effects are those appear in all its sub-actions. The desirability of action effects is represented by utility values [Blythe, 1999].

A *plan* is a set of actions to achieve intended goal(s). As a plan may contain abstract actions (i.e., an abstract plan), each abstract plan indicates a *plan structure* of decomposition. Decomposing the abstract actions into primitive ones in an abstract plan results in a set of primitive plans (i.e., plans with only primitive actions), which is directly executable by agents. In addition, each action in the plan structure is asso-

ciated with the *performers* (i.e., agents capable of performing the action) and the *authority* (i.e., agent who authorizes the action execution). The performers cannot execute the action until authorization is given by the authority. This represents the hierarchical organizational structure of social agents.

Communicative Events

Communicative event is represented as speech act [Austin, 1962] sequence. For our purpose, we analyze the following speech acts typically found in negotiation dialogues (e.g., [Traum *et al*, 2003]) that help infer dialogue agents' intentions, foreknowledge and choices in acting (Variables x and y are different agents. Let p and q be propositions and t be the time stamp).

(1) *inform*(x, y, p, t): x informs y that p at t.
(2) *order*(x, y, p, t): x orders y that p at t.
(3) *request*(x, y, p, t): x requests y that p at t.
(4) *accept*(x, p, t): x accepts p at t.
(5) *reject*(x, p, t): x rejects p at t.
(6) *counter-propose*(x, p, q, t): x counters p and proposes q at t.

Action Execution

In task-oriented domain, an observed action is either intentionally performed by the actor or due to *negligence*. Besides, action execution can be *successful* or *fail*. Intentional action if successful, achieves the *intended effects* as well as some side effects. Otherwise, it ends up with *failed attempt* (Universal quantifiers are omitted below).

Negligence: $\neg\text{intend}(x, A, t1) \wedge \text{do}(x, A, t2) \wedge t1<t2$
Action success: $\text{do}(x, A, t1) \wedge (p \in \text{effect}(A) \rightarrow \text{occur}(p, t2)) \wedge t1<t2$
Action failure: $\text{do}(x, A, t1) \wedge \exists p(p \in \text{effect}(A) \wedge \neg\text{occur}(p, t2)) \wedge t1<t2$
Intended effect: $\text{intend}(x, \text{by}(A, p), t1) \wedge \text{do}(x, A, t2) \wedge \text{occur}(p, t3) \wedge t1<t2<t3$
Failed attempt: $\text{intend}(x, \text{by}(A, p), t1) \wedge \text{do}(x, A, t2) \wedge \neg\text{occur}(p, t3) \wedge t1<t2<t3$

3.3 Inferences

The inferences of physical causality, coercion, intentions and foreknowledge are informed by dialogue and causal evidence in social interactions. We introduce commonsense heuristics that allow an agent to make inferences based on this evidence.

Agency

A first step in attributing responsibility and blame is to identify which agents causally contribute to the occurrence of an outcome under evaluation. In multiagent plan execution environment, an actor often produces an outcome through the assistance of other agents. These other agents are viewed as indirect agency that helps producing the outcome. Given a specific outcome p and the observed action set S, the following actions in S are relevant to achieving p:

- The primitive action A that has p as its effect.
- The actions that establish a precondition of a relevant action to achieving p.
- If p or a precondition of a relevant action is enabled by the consequent of a conditional effect, the actions that establish the antecedent of the conditional effect are relevant.

In the absence of coercion, the actor is the *primary* responsible agent. Other performers of the relevant actions are *partially* responsible for *p*.

Intentions

Natural language conversation is a rich information source for inferring intentions. Assume conversations between agents are *grounded* [Traum, 1994] and they conform to Grice's maxims of *Quality* and *Relevance* (i.e., true and relevant information exchange in conversation).

An *order* or a *request* shows the speaker's *intent*. The two speech acts have different implications on the social status between the speaker and the hearer. If an order is successfully issued (i.e., there exists superior-subordinate relationship between the speaker and the hearer), it creates a social *obligation* for the subordinate to perform the content of the act. The hearer may *accept*, *reject* or *counter-propose*. Various inferences can be made depending on the response of the hearer and the power relationship between the speaker and the hearer. For example, if the hearer counters the order/request, and proposes another alternative, it can be inferred that both the speaker and the hearer *know* the *alternatives*. It is also believed that the hearer does *not intend* what is ordered/requested, but *want* the alternative. If the speaker has known the alternatives yet still requests (or orders) one, infer that the speaker *intends* the chosen action and *not* the alternative. The reader may refer to [Mao & Gratch, 2003] for the inference rules.

Outcome intent can also be partially inferred from evidence of act intention. If an agent intends an action voluntarily, the agent must intend at least one action effect. If there is only one action effect (significant to the agent), the agent must intend the only effect. As Plans provide context in evaluating intention, with association to the goals and reasons of an agent's behavior, in the absence of clear evidence from dialogue inference, we employ general plan-based algorithm to recognize intentions [Mao & Gratch, 2004b].

Foreknowledge

As foreknowledge refers to an agent's knowledge state, it is mainly derived from dialogue inference. For example, *inform* gives the evidence that the speaker knows the content of the act; if *grounded*, the hearer is also believed to know the content. Besides, intention recognition helps infer an agent's foreknowledge, as intentions entail foreknowledge (*Axiom 4* in [Mao & Gratch, 2004a]).

Coercion

Two concepts are important in understanding coercion. One is *social obligation*, created by utterance, role assigned, etc. However, there is difference between obligation and coercion. For example, if some authorizing agent commands another agent to perform a certain action, then the latter agent is obliged to perform the action. But if the latter agent is actually willing to perform the action, this is a voluntary act rather than a coercive one. Here, *willingness* is another important concept.

$\neg(\exists t1)(t1<t2 \wedge \text{want/intend}(x, p, t1)) \wedge \text{obligation}(x, p, y, t2) \wedge \text{accept}(x, p, t3) \wedge t1,t2<t3 \wedge t3<t4 \Rightarrow \text{coerce}(y, x, p, t4)$

If there is no clear evidence that an agent wants or intends, but the agent is obliged to do so, there is evidence of coercion. When there is clear evidence of the unwillingness, this is a strong case of coercion.

An actor could be absolved of responsibility if she was coerced by other forces, but just because an agent applies coercive force does not mean coercion actually occurs. What matters is whether this force truly constrained the actor's freedom to avoid the outcome. Causal inference helps evaluate outcome coercion from evidence of act coercion.

Given the action preconditions are initially true. If an agent is coerced to execute a primitive action, the agent is also coerced to achieve all the action effects. If being coerced to execute an abstract action and the action has only one decomposition, then the agent is also coerced to execute the sub-actions and achieve all the sub-action effects. If the coerced action has multiple decompositions, then the agent has choices: only the effects appear in all alternatives are unavoidable, and thus these effects are coerced; since other effects that only appear in some (but not all) alternatives are avoidable, they are not coerced. If some agents block other action alternatives (by disabling action preconditions), the only alternative left as well as its effects are coerced. These blocking agents are also viewed as coercers. If a conditional effect is coerced and its antecedent is initially true, its consequent is also coerced; otherwise, if the antecedent is initially false or disabled by some other agent, or the coerced agent is able to disable it, the consequent is not coerced.

4 Evaluation and Comparison

To test how our approach models human social inference process, we conducted a survey study to collect human data and analyze the majority agreement of the population from data. We then compare the results given by our model and those by other models with human majority to empirically validate our model.

4.1 Hypothesis

It is not uncommon to use a physical causality as a substitute for modeling social causality and responsibility, for example, previously we used one such model in our *MRE* system. A *simple causal model* always assigns responsibility and blame to the actor whose action directly produces the outcome. Instead of picking up the actor, a slightly more sophisticated model can choose the highest authority (if there is one) as the responsible and blameworthy agent. We can such model *simple authority model*.

Chockler and Halpern [2004] propose a structural-model approach to responsibility and blame (abbreviated to *C&H model* below). They give a definition of responsibility, which extends the definition of causality introduced by Halpern and Pearl [2001]. For example, if a person wins an election 11-0, then each voter who votes for her is a cause for the victory. But in the case of 11-0, each voter is less responsible for the victory than if she had won 6-5. Based on this notion of responsibility, they then defined the degree of blame, using the expected degree of responsibility weighed by the epistemic state of an agent.

Our *hypothesis* is that our computational model based on psychological attribution theory performs better than these existing models. By performing better, we mean that the judgment results given by our model fit the majority of people's answers better than those by other models.

4.2 Methods

Subjects were presented with a small survey consisting of 4 different scenarios (attached in the *Appendix*). Each scenario is followed by several questions, asking the subjects their evaluations of responsibility and blame. The subjects then choose answers from a list of categories. *Scenario 1* is the firing squad example used in the related work [Chockler & Halpern, 2004]: There is a ten-man firing squad. Only one marksman has live bullets in his rifle; the rest have blanks. The marksmen do not know who has the live bullets. They shoot together and the death occurs. *Scenario 2* extends the example to include an authority - the commander, who orders the squad to shoot. *Scenario 3* further extends the example. It involves a negotiation conversation between the commander and the marksmen. The marksmen first reject the commander's order. The commander orders again and finally the squad accepts the order and shoot. In the *Scenario 4*, the commander still orders. However, each marksman has freedom to choose either using blanks or live bullets before shooting.

4.3 Results

There are 27 subjects (most are university staffs including graduates) attending the survey, with ages ranging from 20 to 45 and evenly distributed genders. The sample result of each question in the survey is analyzed by proportion. Sample proportions are then used to estimate proportions for the whole population with confidence. This can be done by typical statistical approach (i.e., using sample as standard proportion, compute a confidence interval of proportion for the population. Refer to e.g. [Rice, 1994]). *Figure 2* illustrates the proportions of the population agreement on responsibility and blame in scenarios 1-4 based on the survey data (confidence level=0.95, α=0.05). For example, in *scenario 1*, 3 subjects blame the marksman with live bullets in his rifle, 19 blames all the marksmen and the rest do not blame marksmen. The analysis of sample data shows that less than 23 percent of the whole population blame the marksman with live bullets, 53 to 87 percent blame all the marksmen, and 4 to 33 percent do not blame any of them, with 0.95 confidence.

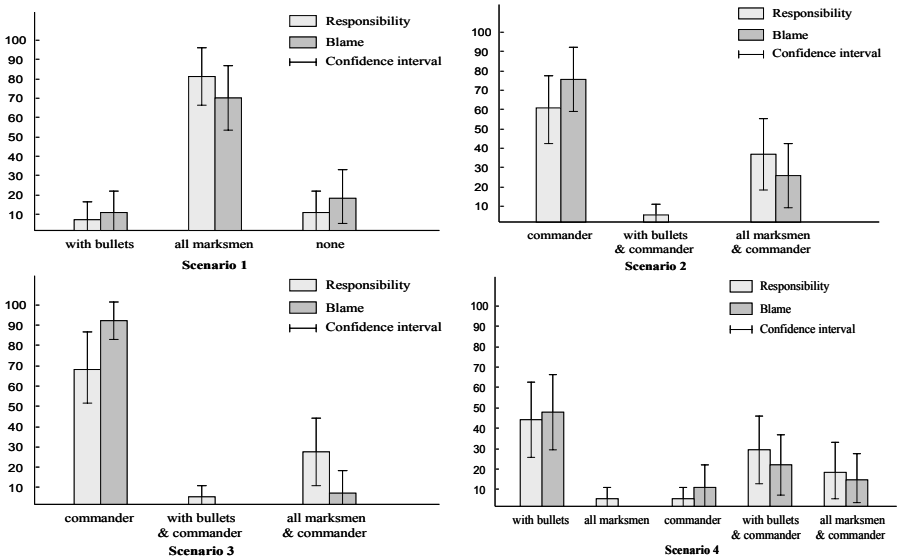

Figure 2 Proportions of the Population Agreement on Responsibility/Blame in Sample Scenarios

BLAME	Simple Cause Model		Simple Authority Model		C&H Model		Social Inference Model		Human Majority Agreement
	Results	Fit/Not	Results	Fit/Not	Results	Fit/Not	Results	Fit/Not	
S 1	with bullets	X	N/A	N/A	all marksmen	√	all marksmen	√	**all marksmen**
S 2	with bullets	X	commander	√	all marksmen& commander	X	commander	√	**commander**
S 3	with bullets	X	commander	√	all marksmen& commander	X	commander	√	**commander**
S 4	with bullets	√ (partial)	commander	X	N/A	N/A	with bullets	√ (partial)	**with bullets/ w. bullets & commander**

Table 1 Comparison of Results by Different Models with Human Data

Table 1 shows the results on blame generated by different models. All the results are compared with the dominant proportions (i.e. majority) of people's agreement (though in *Scenario 4*, there is an overlap between two categories. So we check our model as partial fit). Simple causal model always chooses physical causality. It only partially matches human agreement in *Scenario 4*, but is inconsistent with the data in *Scenarios 1-3*. Simple authority model always picks up the highest authority. It matches the human data in *Scenario 2 and 3*, but is inconsistent with the data in other scenarios. In general, simple models are insensible to the changing situation specified in each scenario and provide invariant types of answers.

C&H model does not perform very well either. It matches human agreement only in Scenario 1. In the rest scenarios, the results it returns are incompatible with the data. Like many other work in causality research, the underlying causal reasoning in C&H model is based on *philosophical* principles (i.e., counterfactual dependencies). Though their extended definition of responsibility accounts better for the extent to which a cause contributes to the occurrence of an outcome, the results show that their blame model does not fit human data well. The empirical findings generally support our hypothesis.

In the next section, we discuss how our model appraises each scenario and compare our approach with C&H model.

4.4 Comparison and Discussions

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 outcome *death*. The team plan is observed executed, and plan outcome occurs. Apply our intention recognition algo-

rithm[1] [Mao & Gratch, 2004b], the marksmen are believed to intend the actions and the only outcome.

The marksman with the bullets is the sole cause of 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, their failed attempt can be detected. As an unsuccessful attempt can be praised or blamed almost the same as a successful one [Zimmerman, 1988, pp. 92], we assign the same valence of utility to *attempting the death*. Therefore, the other marksmen are also blameworthy for their attempt.

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

As there is no notion of intention, C&H model uses foreknowledge as the only determinant for blame assignment. This is fine when there is no foreknowledge, as no foreknowledge entails no intention (as intentions entail foreknowledge). However, when there is foreknowledge, the blame assigned is high, even if there might be no intentions in the case. For example, if a marksman fires the gun by mistake, without any intention of shooting or attempting the death, in C&H model, still he will be blamed just the same as those who intend.

Scenarios 2 & 3

Our model takes different forms of social interactions into account. The inference process reasons about the beliefs of key variables from the perceived communicative and physical acts of agents and based on the plan representation of agents. *Figure 3* illustrates the team plan of the squad in the *Scenarios 2* and *3*, where a commander acts as an authority of the squad (*AND* denotes only one decomposition and *OR* denotes multiple decompositions).

The intermediate inference results for *Scenario 2* are given below (*cmd*, *sqd* and *mkn* stand for the commander, the squad and the marksman, respectively. Beliefs are ordered by the time).

(1)	intend(*cmd*, *do*(*sqd*, *firing*))	(Act *order*)
(2)	obligation(*sqd*, *firing*, *cmd*)	(Act *order*)
(3)	intend(*cmd*, *death*)	(Rule for *intention* & Result *1*)
(4)	coerce(*cmd*, *sqd*, *firing*)	(Act *accept* & Result *2*)
(5)	coerce(*cmd*, *mkn*, *shooting*)	(Rule for *coercion*)
(6)	coerce(*cmd*, *mkn*, *death*)	(Rule for *coercion*)

[1] Note that our intention recognition algorithm is generally applied to a plan library with multiple plans and sequences of actions, which is typical in intelligent agent applications. In this oversimplified example, intention recognition becomes trivial.

So in *Scenario 2*, the marksmen cause/attempt the death due to coercion. The commander is responsible for the death. As the commander intends the outcome, the commander is to blame with high degree.

Scenario 3 includes a sequence of negotiation acts. The above beliefs 4-6 thus change to the following.

(4) ¬intend(*sqd*, *firing*) (Act *reject* and Result *1*)
(5) coerce(*cmd*, *sqd*, *firing*) (Act *accept* and Results *2* & *4*)
(6) coerce(*cmd*, *mkn*, *shooting*) (Rule for *coercion*)
(7) coerce(*cmd*, *mkn*, *death*) (Rule for *coercion*)

Clearly the marksmen do not intend firing. *Scenario 3* shows strong coercion. This is also reflected in the data. More proportions of people regard the commander as responsible and blameworthy in *Scenario 3* than in *Scenario 2*.

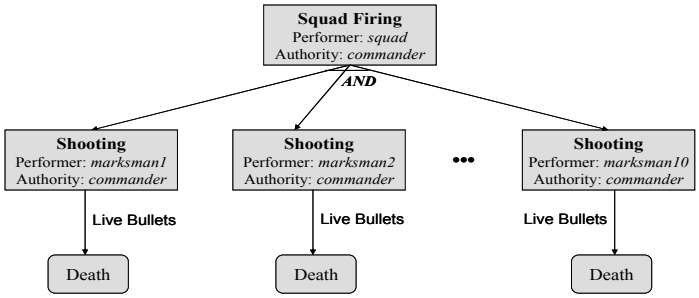

Figure 3 Team Plan of the Squad in Scenarios 2 and 3

C&H model represent all the relevant events in the scenarios as random variables. So if we want to model the communicative acts in *Scenarios 2* and *3*, each act would be a separate variable in their model. This is problematic when conversational dialogue is involved in a scenario. As the approach uses the structural equations representing the relationships between variables, and each equation in the model must be deterministic, it is difficult to come up with such equations for a dialogue sequence. For example, in *Scenario 3*, if we want to include communicative acts such as order, reject and accept in the reasoning process, we need to give deterministic relationship about them (e.g., if the commander orders, then the squad rejects). Such strict equations simply do not exist in a natural conversation. If we ignore some communicative acts in between, some important information conveyed by these acts will be lost.

Assume marksman 1 is the one with bullets. Using C&H approach, the outcome is counterfactually depends 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 C&H model, both the commander and marksman 1 are responsible for the death, each with degree of responsibility 1. This result is inconsistent with human data. Blame assignment is based on the epistemic states of the commander and the marksmen before action performance. There are ten situations altogether. In each situation, the commander has expected responsibility 1, so the commander is to blame with degree 1. The marksmen each has degree

of blame 1/10. So C&H model appraises that the commander and all marksmen are blameworthy for the outcome. As their model for responsibility and blame is the extension of counterfactual causal reasoning, which has been criticized as far too permissive [Hopkins & Pearl, 2003], the same problem is also reflected in their model of responsibility and blame.

Scenario 4

Different from 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, act coercion is true. However, based on the rules of inferring outcome coercion from act coercion, the marksmen are not coerced the outcome. So 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 2* 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 falls into is clearly better than the rest.

As C&H model requires all the structural equations are deterministic, essentially their model could not handle alternative courses of actions, which inherently have nondeterministic property. One way to compensate for this is to push the nondetrminism into the setting of the context. For example, in Scenario 4, they could have the model to let the context determine whether the bullets are live or blank for each marksman, and then have a probability over contexts. Then they can compute the probability of an actual cause. However, these contexts are only background variables. Their probabilities could not impact the reasoning process at all.

5 Summary

Causality is a central issue in many AI applications. This paper presents a domain-independent computational model of social causality and responsibility judgment. The approach bases on the broad features people use in behavior judgment, and models the commonsense interpretation of this social reasoning process. We present how the model derives key variable values for the judgment task, and conduct empirical study to evaluate and compare the model with several other models.

Appendix Firing Squad Scenarios

Scenario 1 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** There is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide 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 There is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide 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 reject the order. The commander insists that the marksmen shoot the prisoner. The marksmen shoot at the prisoner and he dies.

Scenario 4 There is a firing squad consisting of a commanding officer and ten excellent marksmen that generally abide 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. The group decides if the prisoner lives (i.e., everyone chooses blanks), he is set free. The marksmen shoot at the prisoner and he dies.

References

1. Austin, J. 1962. *How to Do Things with Words*. Harvard University Press, 1962.
2. Blythe, J. 1999. Decision-Theoretic Planning. *AI Magazine* 20(2):37-54.
3. Chockler, H. and Halpern, J. Y. 2004. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research* 22:93-115.
4. Halpern, J. Y. and Pearl, J. 2001. Causes and Explanations: A Structural-Model Approach – Part I: Causes. *Proceedings of the Seventeenth Conference in Uncertainty in Artificial Intelligence*.
5. Hopkins, M. and Pearl, J. 2003. Clarifying the Usage of Structural Models for Commonsense Causal Reasoning. *Proceedings of AAAI Spring Symposium on Logic Formulizations of Commonsense Reasoning*.
6. Mao, W. and Gratch, J. 2003. The Social Credit Assignment Problem (Extended Version). *ICT Technical Report ICT-TR-02-2003*.
7. Mao, W. and Gratch, J. 2004a. Social Judgment in Multiagent Interactions. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*.
8. Mao, W. and Gratch, J. 2004b. Utility-Based Approach to Intention Recognition. *AAMAS 2004 Workshop on Agent Tracking* (also available as *ICT Technical Report ICT-TR-01-2004*).
9. Pearl, J. 2000. *Causality: Models, Reasonig and Inference*. Cambridge University Press.
10. Rice, J. A. 1994. *Mathematical Statistics and Data Analysis (Second Edition)*. Duxbury Press.
11. Shaver, K. G. 1985. *The Attribution of Blame: Causality, Responsibility and Blameworthiness*. Springer-Verlag.
12. Traum, D., Rickel, J., Gratch, J. and 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*.
13. Weiner, B. 1995. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press.
14. Zimmerman, M. J. 1988. *An Essay on Moral Responsibility*. Rowman & Littlefield.