Appears in the 5th International Joint Conference on Autonomous Agents and Multiagent Systems, 2006

Evaluating a Computational Model of Social Causality and Responsibility

Wenji Mao
University of Southern California
Institute for Creative Technologies
13274 Fiji Way, Marina del Rey, CA 90292
mao@ict.usc.edu

Jonathan Gratch
University of Southern California
Institute for Creative Technologies
13274 Fiji Way, Marina del Rey, CA 90292
gratch@ict.usc.edu

ABSTRACT
Intelligent agents are typically situated in a social environment and must reason about social cause and effect. Such reasoning is qualitatively different from physical causal reasoning that underlies most intelligent systems. Modeling social causal reasoning can enrich the capabilities of multi-agent systems and intelligent user interfaces. In this paper, we empirically evaluate a computational model of social causality and responsibility against human social judgments. Results from our experimental studies show that in general, the model's predictions of internal variables and inference process are consistent with human responses, though they also suggest some possible refinement to the computational model.

Categories and Subject Descriptors
I.2 [**Computing Methodologies**]: Artificial Intelligence; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms
Algorithms, Experimentation

Keywords
Intelligent Agents, Cognitive Modeling, Causality, Commonsense Reasoning

1. INTRODUCTION
A growing number of applications seek to incorporate automatic reasoning techniques into intelligent agents. Many intelligent systems incorporate planning and reasoning techniques designed to reason about *physical* causality. Since intelligent agents are typically situated in a multiagent environment and multiagent interactions are inherently *social*, physical causes and effects are simply inadequate for explaining social phenomena. In contrast, social causality, both in theory and as practiced in everyday folk judgments, emphasizes multiple causal dimensions, involves epistemic variables, and distinguishes between physical cause, responsibility and blame.

AAMAS'06, May 8-12, 2006, Hakodate, Hokkaido, Japan.

The multiagent research community has considered many aspects of social reasoning (e.g., [1, 20]), but has largely ignored this crucial distinction between physical and social causal reasoning. A model of the process and inferences underlying social causality can enrich the cognitive and social functionality of intelligent agents. Such a model can help an agent to interpret the observed social behavior of others and impact the way an agent acts on the world, which is crucial for successful interactions among intelligent entities. With the advance of multi-agent systems and systems that socially interact with people, it is increasingly important to model this central form of human social inference. Social causal reasoning can inform the design of human-like agents, guide conversation strategies and help modeling and understanding social emotions [6].

We have developed a general computational model of social causality and responsibility [10, 11] that formalizes the factors people use in reasoning about social events. Psychological and philosophical theories identify intermediate constructs that determine social causality. In these theories, social causality involves not only physical causality, but also epistemic variables such as freedom of choice, intention and foreknowledge [19, 22, 24]. As a result, an actor may physically cause an event, but be absolved of responsibility and blame. Conversely, a person may be held responsible and blameworthy for what she did not physically cause. Our model serves as a bridge between such theoretical distinctions and their computational realization, by inferring these additional factors from a representation of the physical and social context.

To evaluate this computational model, we need to assess the model's consistency with human judgments of social cause and responsibility. This is challenging given that people's judgment results often vary, and sometimes, there might be even no consensus among people. But further, we are interested in the more challenging task of testing whether the inferential mechanism of the model is consistent with the intuition of human inference, that is, does our model infer judgment results *in the way* people actually do?

In the rest of the paper, we first review the psychological theory for social causality and responsibility, and the computational model we develop. We then discuss the details of the experimental evaluation of the model, including the methodology, results and some empirical findings.

2. ATTRIBUTION THEORY

Our work is based on the influential attribution theories of Shaver [19] and Weiner [22] of social causality and responsibility (we adopt the terminology of Shaver in this paper). Their theories argue that physical causality and coercion identify *who* is responsible for some outcome under judgment, whereas epistemic factors, intention and foreseeability, determine *how much* responsibility and blame/credit are assigned. Below we summarize their theories.

The assessments of physical causality and coercion identify the responsible party. *Physical causality* (including personal causality and environmental causality) refers to the connection between events and the outcomes they produce. 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.

Intention and forseeability 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. If an agent intends an action to achieve an 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. *Foreseeability* refers to an agent's foreknowledge about actions and their consequences. 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 intention* (i.e., intended action effect), rather than *act intention* (i.e., intended action) that are key in responsibility judgment [23]. Similar difference exists in *outcome coercion* (i.e., coerced action effect) and *act coercion* (i.e., coerced action). The result of the judgment process is the assignment of certain blame or credit to the responsible agent(s). 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.

3. THE COMPUTATIONAL MODEL (A REVIEW)

To model the process of social causality and responsibility attribution, we have constructed a computational model that can automatically derive the judgments that underlying attributions of responsibility and blame from observations and knowledge about social acts. Two important sources of information facilitate this inference process. One source is the actions performed by the observed agents (including physical acts and speech acts). The other is the causal knowledge about actions and their effects. To represent causal knowledge, we have adopted a hierarchical plan representation used by many intelligent 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, and plan interactions. The computational model is described in detail elsewhere [10, 11, 6]. Here we briefly review the inference techniques.

Figure 1 illustrates an overview of the computational model. The inference process infers beliefs from dialogue evidence and causal evidence. Both dialogue inference and causal inference make use of commonsense heuristics, and derive beliefs about the attribution values.

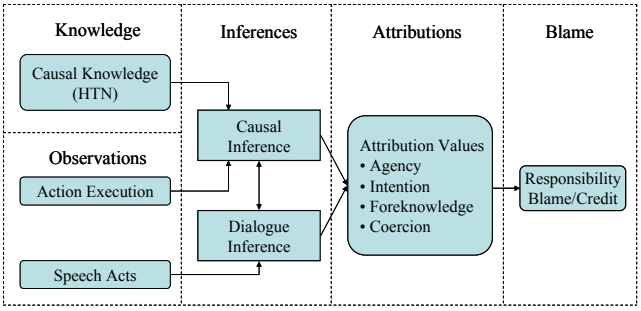

Figure 1. Computational model of responsibility attribution.

Dialogue inference reasons about beliefs of participating agents' desires, intentions, foreknowledge and choices from dialogue communication, represented as a sequence of speech acts. For example, *inform* gives evidence that both the speaker and the hearer know the content of the act, given the conversation between agents is *grounded* [21]. A *request* shows the speaker's desire (or want). An *order* shows the speaker's intent, and if it is successfully issued, it creates an obligation for the hearer to perform the content of the act. The hearer may *accept*, *reject* or *counter-propose* the request or order. Various inferences can be made depending on current beliefs and the response of the hearer [10].

Causal inference adopts a plan-based approach to reason about agency, coercion and intentions. To infer *agency*, the approach first identifies the performing agent whose action directly causes the outcome. Other agents who assist the performer by enabling action preconditions are viewed as indirect agency.

Determinations of *coercion* are informed by dialogue and causal inference. Two concepts are important in assessing coercion. One is social *obligation*, created by utterance, role assignment, etc. The other is *unwillingness*. For example, if an authorizing agent commands another agent to perform a certain action, then the latter agent has an obligation to do so. If there is no clear evidence that an agent intends beforehand, and the agent accepts the obligation, there is evidence of *weak* coercion. Note, intend(x, p, $t1$) denotes that agent x intends that proposition p at time $t1$, obligation(x, p, y, $t2$) represents that x has an obligation p by agent y at time $t2$, accept(x, p, $t3$) represents that x accepts that p at time $t3$, and coerce(y, x, p, $t4$) represents that y coerces x that p at time $t4$.

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

If there is clear evidence of the unwillingness (i.e., intend(x, p, $t1$) is false), there is evidence of *strong* coercion.

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

Causal inference helps infer outcome coercion from evidence of act coercion, by examining the plan structure and action alternatives. For example, assume an agent is coerced to perform an ab-

stract action. If there is only one way to realize (i.e., decompose) this act, the model concludes that the agent must perform all the subsequent actions within this realization and the occurrences of all the action effects are unaviodable. If, instead, the coerced action has multiple realizations (i.e., decompositions), only the effects shared across all of these alternatives are necessarily coerced. Non-shared effects only occur if the coerced agent chooses that particular realization of the coerced act. Outcomes can also be coerced indirectly, for example, some agents can block other alternatives by disabling action preconditions; or they can enable a conditional effect. Similarly, indirect coercion can be inferred from plan knowledge and action alternatives [13].

Intentions (i.e., act intention and outcome intention) can be inferred from dialogue evidence. For example, rejecting what the speaker wants or intends shows no intention. Besides, causal inference helps partially infer outcome intention from act intention. For example, 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), we can exactly infer which effect the agent intends. 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 a general plan-based algorithm to recognize intentions [12]. If a plan is intended by agents, then the actions and effects that are *relevant* to goal achievement (i.e., in the path from initial states to goal states of the plan) are intended. Other action effects are viewed as *side effects* in goal attainment and thus are not intended by the agents.

Foreknowledge refers to an agent's epistemic state before action execution. It is inferred from dialogue evidence (e.g., speech act inform, tell or assert). Intention recognition also helps infer an agent's foreknowledge, as intentions entail foreknowledge (*Axiom 4* in [11]).

The attribution values derived from causal inference and dialogue inference serve as inputs of the attribution process. The algorithm is detailed in [11]. As the last step, this algorithm determines responsibility and assigns proper blame or credit to the responsible agent(s) based on the implications of attribution theory.

4. RELATED WORK

Recent computational approaches have addressed social causality and responsibility judgment by extending causal models [7, 3]. Halpern and Pearl [7] proposed a definition of *actual cause* within the framework of structural causal models. As their approach can extract more complex causal relationships from simple ones, their model is capable of inferring indirect causal factors including social causes. Take the "firing squad" example [17]: there is a two-man firing squad; on their captain's order, both riflemen shoot simultaneously and accurately, and the prisoner dies. Besides the two riflemen who physically cause the death, Halpern & Pearl's model can find the captain as one actual cause for the death. Chockler and Halpern [3] extended this notion of causality, to account for degree of *responsibility*. They gave a definition of responsibility. For example, if a person wins an election 11-0, then each voter who votes for her is a cause for the victory, but each voter is less responsible for the victory than each of the voters in a 6-5 victory. Based on this definition of responsibility, they then defined the *degree of blame*, using the expected degree of responsibility weighed by the epistemic state of an agent.

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 (corresponding to foreseeability in Shaver's term). Important concepts in moral responsibility, such as intention and freedom of choice are excluded in their definition. As a result, their model uses foreknowledge as the only determinant for blame assignment, which is inconsistent with psychological theories. As their 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, this brings about 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 is 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.

In contrast, our approach is built on general plan representation commonly used in many intelligent systems. Causal inference is a plan-based evaluation over this representation. Our model takes different forms of interactions into account, and makes use of commonsense reasoning to infer beliefs from dialogue communication. In Mao & Gratch [13], we use four variants of the original firing squad scenario in the related work [3], to empirically compare our model with Chockler & Halpern's model and two other models (i.e., simple cause model and simple authority model). The results show that for responsibility and blame assignments, our model better approximates human judgments than these alternative models [13].

Social psychological studies show that people consider intentions, coercion and foreknowledge in their judgments. In our work, we have come up with computational account of all these variables. We would like to directly assess our model's ability in predicting the internal variables with respect to human results, and the veracity of the inference process that leads to the corresponding results.

5. EXPERIMENTAL EVALUATION

To evaluate the model, first we need to assess its consistency with human responses (i.e., given the same inputs, do people and the model produce the same outputs). Rather than simply viewing the model as a black box, however, we are also interested in assessing the consistency of the model's internal structure and processes underlying human attributions of responsibility and blame (i.e., do people use the same sources of evidence and generate the same intermediate conclusions). The results for the first task were already demonstrated in [13], though here we seek to extend these finding to additional scenarios. The second task is the focus of this paper.

5.1 Method

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 intentions and coercion. Thus, a good test is to see how the model performs when the evidence for

such judgments is systematically varied. In this study, we take a description of a single social situation and systematically generate several variants, using the inference rules of our model as a guide. For example, if our model suggests that a particular line of evidence is necessary to infer coercion, than an obvious variation would be to eliminate that line.

As a starting point, we adopt the well-known "company program scenario" (*Figure 3*: *Scenario 2*) in experimental philosophy research [5, 8]. In our study, descriptions of each scenario are organized into separate labeled statements of evidence (e.g., *E1*-*E6*). In *Scenario 1* (*Figure 2*), we manipulate evidence related to foreknowledge of the outcome:

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.

Questions:

1. Does the vice president want to start the new program?
   Your answer: Yes No
   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
   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
   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
   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
   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

Figure 2. Scenario 1 and the questionnaire

Each scenario is followed by a questionnaire. The questions in the questionnaires are designed to test the beliefs about different variables. *Figure 2* shows the wording of the questions after *Scenario 1*.

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 3. Scenario 2

In *Scenario 3,* we manipulate the degree of perceived coercion and willingness of the coerced agent by introducing an alternative course of action that will not harm the environment and which the vice president prefers. Specifically, we add one line to Scenario 2 between *E3* and *E4*:

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.

Figure 4. Scenario 3

In *Scenario 4*, we manipulate 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 the option:

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.

Figure 5. Scenario 4

Table 1. Model predictions and people responses for sample scenarios

		Question 1		Question 2		Question 3		Question 4		Question 5		Question 6	
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Chair	VP
S1	Model	√		√		√			√		√		√
	People	30	0	27	3	29	1	2	28	0	30	3.00	3.73
S2	Model	√		√		√			√	√		√	
	People	30	0	30	0	30	0	10	20	22	8	5.63	3.77
S3	Model	√			√	√		√		N/A		√	
	People	21	9	2	28	29	1	21	9			5.63	3.23
S4	Model	√			√		√	N/A		N/A			√
	People	21	9	5	25	5	25					4.13	5.20

5.2 Results

At the level of assessing internal variables, we have no computational alternative to compare with (as the related work only infers part of the variables). In our experiment, we directly compare the predictions of the model with human data.

5.2.1 Assessing Inferred Beliefs

Table 1 shows the number of responses to each question in the sample scenarios. The values for the last questions are the averages of people's answers (on a 6-point scale). The model's predictions are checked with '√'. The data show that for most questions, people agree with each other quite well. But certain disagreement exists on some of the questions.

Though people sometimes may disagree with each other on specific questions, our purpose is to assess the model's general agreement with people. We measure the agreement of the model and each subject using *Kappa statistic*. *Kappa* coefficient is the de facto standard to evaluate inter-rater agreement for classification tasks [1]. It corrects the raters' proportional agreement $P(A)$ due to chance agreement $P(E)$:

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

Di Eugenio & Glass [4] argue 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. We average the results of *Kappa* agreement of the model's predictions with each subject's answers. *Table 2* gives the results of *Kappa* agreement of the model's predictions with each subject's answers.

Table 2. The *Kappa* agreement of the model and subjects

Subjects	*P*(*A*)	*P*(*E*)	*K*
1	.824	.526	.628
2	.882	.543	.742
3	.706	.491	.422
4	.882	.509	.761
5	.941	.526	.876
6	.882	.543	.742
7	.941	.561	.866
8	.941	.561	.866
9	.882	.543	.742
10	.941	.526	.876
11	1	.543	1
12	.882	.543	.742
13	1	.543	1
14	.941	.526	.876
15	.765	.543	.485
16	.824	.491	.667
17	.824	.561	.598
18	.765	.543	.485
19	.882	.509	.761
20	.941	.526	.876
21	.824	.561	.598
22	.882	.543	.742
23	.765	.543	.485
24	.941	.526	.876
25	.941	.561	.866
26	.882	.509	.761
27	.765	.578	.443
28	.882	.543	.742
29	.824	.491	.667
30	.882	.509	.761
Average			.732

The average *Kappa* agreement between the model and subjects is *0.732*. According to [18], $0.61<K<0.80$ 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, each belief is derived by a specific inference rule, so each question in the questionnaire corresponds to the firing of one rule (with the exception of three questions in the questionnaire designed to test two rules each). Currently, we have 37 dialogue and causal inference rules in the model. This survey study covers 19 of them. To assess the inference rules, we compare the conditions of each rule with the evidence people use in forming each answer. For every subject and every question, we build a confusion matrix [9] to compute the numbers 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 prediction of the corresponding rule with respect to people's evidence choice 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.

$$AC(Q_i) = \frac{\sum_{j \in Subjects} AC(j, Q_i)}{Ns} = \frac{\sum_{j \in Subjects} (TP(j, Q_i) + TN(j, Q_i))}{Ns * Ne}$$

Table 3 gives the accuracy of each tested rules (*R**). Given that each question contains 6 or 7 lines of evidence and people choose multiple lines in most cases, the accuracy results are fairly good (well above those generated by a random model). The empirical results show that the sources of evidence the model uses for inference are consistent with human data.

Table 3. Accuracies of inference rules

	Question/Rule	Average Accuracy
Scenario 1	Q1/R7	0.76
	Q2/R13	0.96
	Q3/R2	0.85
	Q4/R5	0.94
	Q5/R1	0.91
Scenario 2	Q1/R6	0.92
	Q2/R10	0.96
	Q3/R24	0.86
	Q4/R25	0.70
	Q5/R11&R14	0.84
Scenario 3	Q1/R17	0.94
	Q2/R18&R19	0.88
	Q3/R11&R16	0.80
	Q4/R27	0.74
Scenario 4	Q1/R29	0.71
	Q2/R28	0.84
	Q3/R30	0.75

In the next section, we discuss how our model appraises each scenario and some experimental findings.

5.3 Discussions

5.3.1 Scenario 1

In *Scenario 1*, we manipulate the variable foreseeability to be false (from evidence *E3*). Evidence is encoded into the model as follows (*VP* and *CH* refer to the vice president and the chairman, respectively):

(E1) request(*VP*, *CH*, do(*VP*, new-program))
(E2) inform(*VP*, *CH*, profit-increase∈effect(new-program))
(E3) inform(*VP*, *CH*, env-harm∉effect(new-program))
(E4) accept(*CH*, do(*VP*, new-program))
(E5) do(*VP*, new-program)
(E6) occur(environmental-harm)

The evaluation shows that the model and people draw similar intermediate conclusions from this evidence, suggesting both that our representation of the evidence and the underlying inference rules are largely correct. For example, the questionnaire specifically queries the perceived desire, foreknowledge and intentions of the characters (see *Figure 2*). The belief that the vice president desires the new program can be inferred from speech act *request* in *E1*. The chairman's intention to start the new program can be inferred from speech act *accept* in *E4*. As starting *new program* has only one action effect (*E2*), we can infer outcome intention from act intention. The chairman must intend the only effect (i.e., *increase profits*). The vice president has no foreknowledge of the environmental harm can be inferred from the content of *inform* in *E3* (note that our approach assumes Grice's maxims of *Quality* and *Relevance*). According to a causal inference rule, no foreknowledge entails no intention.

As the goal of agent(s) was not explicitly expressed in this scenario, intention recognition method is not involved. However, from dialogue evidence and from causal connection of foreknowledge, act intention and outcome intention, the beliefs about intentions are properly inferred.

Subjects gave quite consistent answers to the questions in *Scenario 1*. Their answers to the last question show that blameworthiness is mitigated by no foreknowledge. This result is consistent with psychological findings. Though people assigned relatively more blame to the vice president, the data also suggest that the chairman should share blame with the vice president.

The accuracies of inference rules are also good in general. The accuracy of the rule tested in *Question 1* is lower than the others because besides 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, he should desire to start the new program to increase profits (which is supported by *E2*).

5.3.2 Scenarios 2&3

Scenarios 2 & 3 manipulate the degree of perceived coercion and willingness of the coerced agent. The agents have clear foreknowledge about the harm (*E3*). The chairman's goal of making *more profits* is also clearly informed (*E4* in *Scenario 2*, *E5* in *Scenario 3*).

The following beliefs can be inferred from *Scenario 2*:

(B1) know(*CH*, profit-increase∈effect(new-program))
(derived from *E2*: inform)
(B2) know(*CH*, env-harm∈effect(new-program))
(derived from *E3*: inform)
(B3) intend(*CH*, do(*VP*, new-program))
(derived from *E4*: order)
(B4) obligation(*VP*, do(*VP*, new-program), *CH*)
(derived from *E4*: order)
(B5) coerce(*CH*, *VP*, do(*VP*, new-program))
(derived from *B4* & *E5*: accept)
… …

As there is no evidence of unwillingness, perceived coercion (*B5*) is in a weak sense. The inferred beliefs *B2*, *B3*, *B4* and *B5* give predictions to *Questions 1, 2 & 5* in *Scenario 2*. As there is only one plan in this scenario and the chairman intends the action (i.e., starting *new program*) in the plan (*B3*), intention recognition is trivial[1]. Making *more profits* is the goal of the plan, so it is intended by the

[1] Note that our intention recognition method is generally applied to a plan library with multiple plans and sequences of actions. In this oversimplified example, intention recognition becomes trivial.

chairman (*Question 3*). *Environmental harm* is a side effect of goal attainment, so it is not intended by the chairman (*Question 4*).

In *Scenario 3*, the vice president's counter-proposal provides additional information (*E4*). More beliefs can be derived:

(B3) know(*CH*, alternative(new-program, alternative-program))
(from *E4*: counter-propose)
(B4) want(*VP*, do(*VP*, alternative-program))
(from *E4*: counter-propose)
(B5) ¬intend(*VP*, do(*VP*, new-program))
(from *E4*: counter-propose)
(B6) intend(*CH*, do(*VP*, new-program))
(from *E5*: order)
(B7) obligation(*VP*, do(*VP*, new-program), *CH*)
(from *E5*: order)
(B8) coerce(*CH*, *VP*, do(*VP*, new-program))
(from *B5*, *B7* and *E6*: accept)
(B9) coerce(*CH*, *VP*, achieve(environmental-harm))
(from *B2* and *B8*)
… …

Beliefs *B3*, *B4*, *B8* and *B9* give predictions to *Questions 1, 2, 3 & 4* in *Scenario 3*, respectively. Belief *B5* gives strong evidence of coercion.

There are several disagreements among the subjects in *Scenarios 2 & 3*. In *Question 4* of *Scenario 2*, one-third of the subjects think it the chairman's intention to harm the environment. Whether side-effects are intended is controversial in philosophy, and other empirical studies show similar results [16]. 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 evidence is weaker than 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. This suggests a limitation in our current model. In contrast, when asking the same question in *Scenario 3* (*Question 3*), almost all the subjects agreed that the vice president is coerced to start the new program.

In the first question of *Scenario 3*, some subjects think the chairman does not know the alternative program, though the vice president clearly informs 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, it is our mistake to omit this information in scenario design. Last, 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. Comparing the blame assignments in *Scenarios 2 & 3*, it shows that on one hand, the higher the degree of coercion, the less blame is assigned to the actor. This is 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*.

The accuracies of two tested rules are relatively lower than the others. One is the rule used in *Question 4* of *Scenario 2*. In our model, there are three evidence needed for the inference, *E2*, *E3* and *E4*. Almost all subjects chose evidence *E4*, but most ignored *E2* (except two subjects). One reason is that *E2* as knowledge (i.e., the new program helps increase profits), seems implied in *E4* (otherwise people will have difficulty understanding *E4*). Similarly, for the rule used in *Question 4* of *Scenario 3*, most subjects did not choose knowledge *E3*. However, we think this knowledge is necessary for the inference.

5.3.3 Scenario 4

In *Scenarios 2 & 3*, action and alternative are coerced by authority, whereas in *Scenario 4*, the vice president has some freedom of choice. While the high-level plan (i.e., starting the new program) is still coerced (*E4*), the agent can choose to execute either alternative (simple way or complex way in *E2*). As both ways will increase profits (*E3*), increasing profit is unavoidable in either way: the vice president is coerced to achieve this effect (*Question 1*). However, he is neither coerced to choose the simple way (*Question 2*), nor is he coerced to achieve the specific effect *environmental harm* that only occurs in the simple way (*Question 3*).

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 he must be willing to do so. People assigned more blame to the vice president, as he could have done otherwise. This result is consistent with psychological findings. However, people still assigned considerable blame to the chairman, though it was the vice president's choice to harm the environment.

The inference rules in *Question 1* and *Question 3* are based on evidence *E3*, *E4* and *E5*. Many subjects ignore knowledge *E3*. This lowers the accuracies of the two rules.

5.3.4 Discussion on limitations

Although results show general support for the model, they reveal some limitations of the approach. Subjects tended to assign shared blame to the individuals involved. Though our model is able to handle joint activity and multiagent plan, one limitation of the model is that it always blames one most deserving agent (or group of agents).

It is clear that people made assumptions about the scenarios that were not explicitly represented in the model. For example, people assumed the vice president had the goal of increasing profits even though this was not explicitly stated. This relates to the more general issue of ensuring the correspondence between the model's encoding of the scenario and the subjects' interpretation of the scenario. Currently, we construct this mapping by hand. This has the disadvantage that, as designers of the scenarios, we may unintentionally introduce discrepancies. Alternatives would be to explore ways to automatically generate descriptions from their representation in the model, or at least to use an independent set of coders to characterize the textual encoding.

It is well known that responsibility judgment is influenced by the observer's emotional state, interpersonal goals such as impression management [15], and dispositional differences such as personality. Further, our model of dialogue inference assumes that parties faithfully articulate their actions and beliefs, whereas people are notoriously biased when describing their involvement in creditworthy/blameworthy events. Although we have not accounted for these biases, our current model provides a framework for both generating and recognizing such influences [14].

6. CONCLUSION

A growing number of applications seek to incorporate automatic reasoning techniques into intelligent agents. In contrast to physical causal reasoning that underlies most intelligent systems, social causal reasoning emphasizes multiple causal dimensions, involves

epistemic variables, and distinguishes between physical cause, responsibility and blame. In this paper, we empirically evaluate a computational model of social causality and responsibility using human data. Results from our experimental studies are encouraging in general, though they also indicate some limitations and possible refinement to the computational model.

Our future work needs to better model shared responsibility and blame, and consider the influences of individual difference and the perceiver's subjective bias on the judgment process. We also need to improve experimental design, using objective device to minimize the discrepancies of information encoding.

7. ACKNOWLEDGEMENTS

This work was sponsored by the U. S. Army Research, Development, and Engineering Command (RDECOM). The first author benefits from the USC Knowledge Representation Group seminars led by Jerry Hobbs and Andrew Gordon. Thanks to Anya Okhmatovskaia for the helpful discussions. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. REFERENCES

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

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

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

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

[5] Experimental Philosophy Website (http://experimentalphilosophy.typepad.com/).

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

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

[8] Knobe, J. Intentional Action and Side-Effects in Ordinary Language. *Analysis*, 63:190-193, 2003.

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

[10] Mao, W., and Gratch, J. The Social Credit Assignment Problem (Extended Version). *USC/ICT Technical Report* (http://www.ict.usc.edu/publications/ICT-TR-02-2003.pdf), 2003.

[11] Mao, W., and Gratch, J. Social Judgment in Multiagent Interactions. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, 2004a.

[12] Mao, W., and Gratch, J. A Utility-Based Approach to Intention Recognition. *AAMAS 2004 Workshop on Agent Tracking: Modeling Other Agents from Observations*, 2004b.

[13] Mao, W., and Gratch, J. Social Causality and Responsibility: Modeling and Evaluation. *Proceedings of the Fifth International Conference on Intelligent Virtual Agents*, 2005.

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

[15] Mele, A. R. *Self-Deception Unmasked*. Princeton University Press, 2001.

[16] Nadelhoffer, T. On saving the Simple View. *Mind and Language*, forthcoming.

[17] Pearl, J. Reasoning with Cause and Effect. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

[18] Rietveld, T., and van Hout, R. *Statistical Techniques for the Study of Language and Language Behavior*. Mouton de Gruyter, 1993.

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

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

[21] Traum, D. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Thesis, University of Rochester, 1994.

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

[23] Weiner, B. 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, 2001.

[24] Zimmerman, M. J. *An Essay on Moral Responsibility*. Rowman & Littlefield, 1988.