Modeling Social Inference in Virtual Agents[1]

Wenji Mao

Key Lab of Complex Systems and Intelligence Science
Institute of Automation, Chinese Academy of Sciences
wenji.mao@ia.ac.cn

Jonathan Gratch

Institute for Creative Technologies
University of Southern California
gratch@ict.usc.edu

ABSTRACT Social judgment is a social inference process whereby an agent singles out individuals to blame or credit for multi-agent activities. Such inferences are a key aspect of social intelligence that underlie social planning, social learning, natural language pragmatics and computational models of emotion. With the advance of multi-agent interactive systems and the need of designing socially aware systems and interfaces to interact with people, it is increasingly important to model this human-centric form of social inference. Based on psychological *attribution theory,* this paper presents a general computational framework to automate social inference based on an agent's causal knowledge and observations of interaction.

Keywords *Social Causality*, *Responsibility Judgment*, *Attribution*, *Social Inference*, *Human-like Agents*

1 INTRODUCTION

A growing number of applications have sought to incorporate human-like intelligence into the design of artificial systems. Most intelligent systems incorporate planning and reasoning techniques designed to reason about physical causality. In contrast to how causality is used in the physical sciences, people instinctively seek out a human actor for their everyday judgments of responsibility, credit or blame. Such judgments are fundamental social explanations involving judgments not only of causality but individual responsibility, free will and mitigating circumstances [Shaver, 1985]. They 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.

[1] This work was performed in the Institute for Creative Technologies at the University of Southern California.

With the advance of multi-agent interactive systems, user-aware adaptive interfaces and systems that socially interact with people, it is increasingly important to model and reason about this human-centric form of social inference. Social inference and social judgment help an agent 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. They also facilitate human-human interaction by identifying the underlying cognitive process and principles of human judgments [Mao, 2006]. Social inference can inform the design of human-like agents, guide conversation strategies and facilitate the modeling and understanding of social emotions [Gratch *et al*, 2006]. In a multi-agent environment, social causal reasoning helps distribute responsibility in multi-agent organization [Jennings, 1992], automate after-action review of group performance [Gratch & Mao, 2003], and support social simulation for agent society [Conte & Paolucci, 2004].

Our goal is to develop a faithful computational framework for human-like virtual agents so as to drive realistic behavior generation [Swartout *et al*, 2006]. In virtual human applications, artificial agents not only have figures that resemble humans, they should also make sense of the perceived social events and exhibit human-like social reasoning skill. Psychological and philosophical studies agree on the broad features people use in their everyday judgments. 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, this paper presents a general computational approach to forming social judgment by inferring an agent's causal knowledge and observations of communication and task execution. We conduct empirical studies to validate our framework and extend the framework employing decision-theoretic approach.

2 FROM ATTRIBUTION THEORY TO COMPUTATIONAL APPROACH

Most contemporary psychological studies of responsibility and social judgment draw on attribution theory. Our work is based on the influential attribution theories of Shaver [1985] and Weiner [1995] on social causality and responsibility. In their theories, social causality and responsibility judgment rely on the assessment of several key factors (i.e., *attribution variables*), such as physical causality, foreseeability, intention and coercion (we adopt the terminology of Shaver in this paper). Below we summarize their theories.

The assessments of physical causality and coercion determine 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. *Coercion* refers to the situation as when some external force, such as a more powerful individual or a socially sanctioned authority, limits an agent's freedom of choice. In the absence of coercion, the actor whose action directly produces the outcome is regarded as responsible. However, in the presence of coercion, some or all of the responsibility may be deflected to the coercive force.

Intention and forseeability affect 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. 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. 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 is key in responsibility judgment. 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 assessed by the severity or positivity of the outcome as well as the degree of responsibility. The latter is based on the derived beliefs of attribution variables.

To model the process and inferences of social causality and responsibility judgment, we have constructed a computational framework that can automatically form judgment results of responsibility and blame/credit attribution based on the knowledge and observations of social agents. Two information sources contribute to the judgment 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, and social information of agents' roles and relationship.

Our computational framework takes the observed communicative events and executed actions of agents as inputs. *Causal knowledge* includes a domain action theory and a plan library. *Social*

information specifies social roles and the power relationship of agents. *Causal inference* derives beliefs from causal evidence. *Dialogue inference* derives beliefs from communicative acts. Both inferences make use of commonsense inference rules and generate beliefs of attribution variables. The beliefs then serve as the inputs of the attribution process (implemented as an algorithm). The *attribution algorithm* finally forms an overall judgment and assigns proper blame or credit to the responsible party.

3 REPRESENTING TASK AND SOCIAL INFORMATION

3.1 TASK REPRESENTATION

Causal knowledge is encoded via a hierarchical task representation. An *action* has a set of propositional preconditions and effects. Actions can be either *primitive* (i.e., directly executable by agents) or *abstract*. An abstract action may be decomposed in different ways and each decomposition constitutes an *alternative* way to realize an abstract action. A *plan* is a set of actions to achieve certain intended goal(s). The desirability of action effects (i.e., their positive/negative significance to an agent) is represented by utility values. 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.

3.2 REPRESENTING COMMUNICATION AND ATTRIBUTIONS

Communication between agents is a rich information source for social inference. We represent communicative events as *speech act* [Austin, 1962] sequence, and focus on the speech acts that help infer dialogue agents' desires, intentions, foreknowledge and choices in acting. In our framework, we include the acts such as inform, request, order, accept, reject and counter-propose, which are formally expressed as first-order predicate calculus. For example, inform(x, y, p, t) denotes that agent x informs agent y that proposition p at time t.

Attribution variables are represented in a similar way. Besides physical causality (in which human agency and action effects are encoded via task representation), foreseeability is represented using *know* and *bring about*. We use *intend* and *do* to represent act intention, and *intend* and *achieve* for outcome intention. Act coercion and outcome coercion are represented similarly. Two concepts are important in modeling coercion. One concept is social *obligation*. The other is

(*un*)*willingness*. For example, a command creates an obligation for the receiving agent to perform a certain act. But if the agent is willing to, this is a voluntary act rather than a coercive one.

4 REASONING ABOUT SOCIAL CAUSE AND RESPONSIBILITY

Social judgment needs to infer beliefs about attribution variables from observations of behavior. This inference is always from a perceiving agent's subjective perspective and evaluation of the significance of events is based on the perceiver's personal preference. As different perceivers may have different observations, different knowledge and preferences, they may form different beliefs and judge the same situation differently. We show how automatic methods of causal and dialogue reasoning can provide such a mechanism.

4.1 DIALOGUE INFERENCE

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. Assume communication between agents is grounded, and conversation conforms to Grice's maxims of Quality and Relevance [Grice, 1975]. Social information is also important, for example, an order can be successfully issued only to subordinate, but a request can be made of any agent. We design a small number of commonsense rules that allow a perceiving agent to derive beliefs based on dialogue evidence.

For example, *inform* (or *tell*) gives the evidence that the speaker knows the content *p* of the act. If grounded, the hearer is also believed to know *p*. A *request* shows what the speaker wants. An *order* (or a *command*) shows the speaker's intent. If requested or ordered by a superior, it creates a social obligation for the hearer to perform the content of the act. The hearer may *accept, reject* or *counter-propose* an order (or a request). Various inferences can be made depending on the previous acquired beliefs and the response of the hearer. For instance, if there is evidence of an agent's unwillingness (e.g., the agent does not intend), yet the agent accepts the obligation, there is evidence of coercion.

Rule *D10* [*unwilling-accept-obligation*]:

$\neg$intend(h, p, $t1$) $\wedge$ obligation(h, p, s, $t2$) $\wedge$ accept(h, p, $t3$) $\wedge$ $t1<t3$ $\wedge$ $t2<t3<t5$ $\wedge$ $\neg(\exists t4)(t3<t4<t5 \wedge \neg$coerce($s$, h, p, $t4$)) $\Rightarrow$ coerce(s, h, p, $t5$)

4.2 CAUSAL INFERENCE

Causal inference is a plan-based evaluation of agency, intention and coercion. To infer *agency*, a perceiver first identifies the performer of the action that directly produces the outcome. Other agents who assist the performer by enabling action preconditions are viewed as indirect agency.

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. In more general cases, when an action has multiple effects, in order to evaluate whether a specific outcome is intended or not, a perceiver may examine action alternatives the agent intends and does not intend, and comparative features of intended and unintended alternatives. For instance, if an agent intends an action A voluntarily and does intend its alternative B, and the effect set of B is a subset of that of A, it can be inferred that the agent must intend (at least) one effect that only occurs in A.

As plans provide context in evaluating intention, with association to the goals and reasons of an agent's behavior, if there is no clear belief about intention derived from causal and dialogue inferences, we employ a general plan-based algorithm to recognize intentions [Mao, 2006]. If a plan is intended by agents, then the actions and effects that are relevant to the 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 the goal attainment and thus are not intended by the agents. Intention recognition also helps infer *foreseeability*, as intention entails foreknowledge [Mao, 2006].

Causal inference helps infer *outcome coercion* from evidence of *act coercion*. For example, if an agent is coerced to execute a primitive action, the agent is also coerced to achieve all the action effects. However, this may not be the case for an abstract action. For instance, if the coerced abstract action has multiple decompositions, then the agent has options: only the effects appear in all alternatives are unavoidable, and thus these effects are coerced; Other effects that only appear in some (but not all) alternatives are not coerced. Besides, an agent can indirectly coerce other agent(s), e.g., by enabling action preconditions or blocking other alternatives. Each of these coercive situations follows different inference rules.

4.3 ATTRIBUTION ALGORITHM

The attribution process begins with some specific outcome that is under evaluation. The beliefs derived from the inferences above serve as its inputs. We have developed an algorithm to automate the judgment process [Mao, 2006]. The algorithm first applies dialogue and causal inferences. By default, the performer whose action directly causes the evaluated outcome is assigned to the primary responsible agent. The judgment process starts from the primitive action and works up the task hierarchy. If there is evidence of outcome coercion in certain layer, the authority is deemed responsible. The process proceeds until it reaches the root node of the hierarchy or outcome coercion is no longer true. We adopted a simple categorical model of responsibility assignment. If the outcome is intended, the degree of responsibility assigned is *high*. If the outcome is not intended, then the degree assigned is *low*. Otherwise, assign *medium* degree of responsibility. The *intensity* of credit or blame is computed by multiplying the degree of responsibility and outcome utility. The reader may refer to [Mao, 2006] for illustrations of how the algorithm and the model work in our virtual training application.

5 EVALUATION

To validate the proposed model, we first need to assess the consistency of model predictions of social cause and responsibility with human judgments, that is, given the same inputs, whether our model draws the same conclusion as people do. Second, rather than simply viewing the model as a black box, a more challenging task is to assess the consistency of the model's inferential mechanism underlying human attributions of responsibility and blame, that is, whether our model uses the same sources of evidence and generate the same intermediate beliefs as people do. The results for the first task were already reported in [Mao & Gratch, 2005], although here we seek to extend empirical findings to additional scenarios. Here we report our experimental study on the second task.

Chockler and Halpern [2004] proposed another computational approach to address social causality and responsibility judgment by extending causal models. We use four variants of the original firing squad scenario in the related work [Chockler & Halpern, 2004], to empirically compare our model with their model and two other models (i.e., simple cause model and simple authority model) [Mao & Gratch, 2005]. The results show that our model better approximates hu-

man attributions of responsibility and blame than these alternative models. As the related approaches are incapable of inferring the beliefs of internal variables (such as intention and coercion), for assessing the inference process and rules, we directly compare the predictions of our model with human data. We *claim* that our model predicts human judgments of social attributions and makes inferences consistent with what people do in their judgments.

5.1 METHOD

5.1.1 Participants and Procedure

There were *48* subjects participating in the experiments. They were computer science graduate students or staff at the University of Southern California. Their ages ranged from *20* to *35*. There were *30* male participants. Among them, *12* subjects completed the four-page surveys, and *36* completed two pages of the surveys. 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, intention, desire, obligation and coercion in the scenarios. In answering each question, the subjects were asked to mark the (multiple) lines in the scenario they based their answer on. At the end of each questionnaire, the subjects were asked to score how much blame the characters deserve in the scenario.

5.1.2 Materials

We adopt the "*company program*" scenario used in [Knobe, 2003]. This scenario has received much attention in recent folk psychology and experimental philosophy research. It states a social situation where the chairman of a company discusses a new program with the vice president. The vice president informs that the new program will help increase profits, but it will also harm the environment. However, the chairman only wants to make as much profits as he can. He commands the vice president to start the new program. The vice president follows the command, and the environment is harmed by the new program.

We design three variants of the original company program scenario and also the questionnaires following each scenario [Mao, 2006]. For the convenience of assessing inference rules,

descriptions of each scenario are organized into separately labeled statements of evidence (*E1-E7*).

5.1.3 Experimental Design

According to our computational approach, a good experimental design is to see how the model performs when evidence for the judgment is systematically varied. To this end, we take a description of a single social situation and systematically vary it, using the inference rules of our model as a guide. We encode information into each line of a scenario. The specific information includes speech act, causal knowledge, goal identification, physical action, the occurrence of effects, etc. We design questions to test beliefs about different variables. Each question corresponds to the firing of an inference rule.

In *Scenario 1*, we manipulate evidence related to agents' foreknowledge of the outcome (i.e., no foreknowledge). The vice president informs that the new program has no harm to the environment, and requests the chairman to start the new program. We design questions to test the inference rules for foreseeability, relation of intention and foreknowledge, connection of act and outcome intention, and speech acts request and inform. *Scenario 2* is the original scenario. It 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, having foreknowledge, 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 of statement to *Scenario 3* (and all the other lines remain the same as those in *Scenario 2*). Questions are designed to test the agent's willingness and coercions. 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 questions to test three important rules for coercion.

5.2 RESULTS

5.2.1 Assessing Inferred Beliefs

The experimental data show that for most questions, people agree with each other quite well. But certain disagreement exists on some of the questions. To measure the agreement between model prediction and people response, we use the *Kappa statistic*. We take the skewed distribution of categories (i.e., prevalence) into account and construct contingency tables for the calculation. The average *Kappa* agreement between the model and subjects is ***0.732*** [Mao, 2006], which indicates substantial agreement. The empirical results show good consistency between the model's derivation of intermediate beliefs and human data.

5.2.2 Assessing Inference Rules

Currently, we have 39 dialogue and causal inference rules in the model. This survey study covers 19 of them. As the answer to each question in the questionnaires corresponds to the firing of one inference rule, to assess the accuracies of evidence chosen by the rules, we compare the conditions of each rule with the evidence people use in forming each answer. We measure *accuracy* using confusion matrices. For every subject's answer to each question, we build a confusion matrix to compute the number of true positive (i.e., evidence both the rule and the subject use), true negative (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). The average accuracy of the rules in the model is ***0.85*** [Mao, 2006]. The empirical results show good consistency between the model's evidence choice for the inferences and human data.

5.3 DISCUSSIONS

In general, the evaluation results suggest that both the representation of the evidence and the underlying inference rules in our model are largely consistent with human data. However, they also reveal some limitations of the model. In *Scenario 1*, the accuracy of one rule inferring the vice president's desirability to start the new program is lower than the others. This is because, in addition to the evidence *E1* (speech act request), many people chose *E2* (i.e., the new program will help increase profits) as well. Post-experiment interviews uncovered that people made assumptions that were not explicitly expressed in the scenario. Many subjects had assumed that making

profits should be desirable to the vice president (because of his role), and therefore, he should want to start the new program to increase profits (supported by *E2*). For the same reason, in *Scenario 4*, some subjects think that the vice president is not coerced to increase profits, as they think it the vice president's job to increase profits. He must be willing to do so.

There are several disagreements among the subjects in *Scenarios 2* and *3*. 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 *Scenario 2*, some subjects do not think the vice president is coerced to start the new program by the chairman, as evidence is weaker than in *Scenario 3*. Half of them referred to the evidence *E5*, indicating that they expect the vice president to negotiate with the chairman rather than directly accept the command (*E5*). This result shows some limitation of our model in representing degree of beliefs. In contrast, when being asked the same question in *Scenario 3*, almost all the subjects agreed that the vice president was coerced. Comparing the blame assignments in *Scenarios 2* and *3*, it shows that the higher the degree of coercion, the less blame is assigned to the actor – a result consistent with psychological findings.

Subjects tended to assign *shared blame* to the individuals involved across four scenarios. Although our model supports joint activity and multi-agent plan, one limitation of the model is that it always assigns most of the blame to a single agent (or a group of agents), who has caused or coerced 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 *Scenario 1*). Some researchers' work is relevant to this. Norman and Reed [2000] discuss the issue of *task delegation*: When an agent decides to delegate tasks to others, the responsibility for the task is shared. Lickel [2003] investigates *collective responsibility*, in which blame is extended to others who are not behaviorally involved in the blameworthy event. Though our current approach has not addressed these issues, the model's representational and inferential mechanism has the potential to handle these extensions.

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 or blameworthy events [Nisbett & Wilson, 1977]. Although we have not accounted for these biases, our model provides a framework for both generating and recognizing such influences. Perhaps the most obvious limitation of the model is its inability to deal with *uncertainty*

inherent in observations and judgment process. As a result, the beliefs of variables are treated as binary in the model, either true or false. This is particularly problematic when it comes to infer the mental states of other parties (e.g., their intentions) from insufficient evidence. To address this limitation, we extend the model to incorporate probabilistic representation and decision-theoretic reasoning.

6 DECISION-THEORETIC EXTENSION

We extend the computational framework introduced earlier to incorporate probabilistic representation of actions and plans, and build the probabilistic reasoning mechanism to infer degrees of beliefs. To achieve this, we take a *decision-theoretic* approach that combines utilitarian preferences with probabilities of outcome occurrence. The evaluation of agents' behavior is based on the fundamental principle of "maximum expected utility" (*MEU*) underlying decision theory. Decision theory can be viewed as both a normative and a descriptive theory. As human intuitions in commonsense inference have both descriptive and prescriptive features, decision theory provides an excellent approximation to many judgments and decisions [Slovic *et al*, 1988].

6.1 PROBABILISTIC REPRESENTATION

In a probabilistic task representation, action preconditions and effects occur with probability. An action effect can be nondeterministic, denoted as *effect probability* (i.e., the likelihood of the occurrence of an action effect given the action is successfully executed). To represent the success and failure of action execution, an action has an *execution probability* (i.e., the likelihood of successful action execution given the preconditions are true). Representation of plans is similar to that in *Section 3*, except that we use (expected) plan utility to represent the overall benefit and disadvantage of a plan. The computation of plan utility considers both the desirability of plan outcomes and the likelihood of outcome achievements. The *degree* of belief model extends the classical truth values (true and false) to graded scales, with numerical values ranging from *0* to *1*.

6.2 DECISION-THEORETIC REASONING

Utility is used in two ways in our approach. One way is to represent the perceiving agent's preferences over states. Another way is to represent the presumed preferences of the observed agents. In the latter case, state preferences are used in recognizing agents' intentions.

Intention inference can be viewed as inferring the decision-making strategy of the observed agents. We take a decision theoretic approach and explicitly consider states and state desirability of agents. We assume an agent will adopt a plan (i.e., an intention) to maximize the expected utility of overall goal attainment. The computation of expected plan utility is similar to that in decision-theoretic planning (e.g. *DRIPS*), using the utilities of outcomes in the plan and the probabilities with which different outcomes occur. The calculation of outcome probability takes into account the uncertainty of action preconditions (i.e., state probabilities), uncertainty in action execution (i.e., execution probability), and nondeterministic action effects.

Intention recognition algorithm works on a possible plan library. The algorithm calculates the expected utility of each possible plan, and the one with the highest expected utility is current hypothesized one. Once current hypothesized plan is identified (with probability), we can further infer the intentional/unintentional actions/effects according to their relevance to goal attainment. The degree of intention (being true or false) equals to the probability of goal achievement in current hypothesized plan.

In the probabilistic context, coercion inference is similar to that described in *Section 4*, except that here we compare the expected utilities of actions and plans (instead of specific outcomes). For example, if an agent is coerced to execute an abstract action and the coerced abstract action has multiple decompositions, the evaluation of outcome coercion is based on the estimation of expected utilities of action alternatives. If the evaluated outcome appears in all alternatives (i.e., all the action alternatives have the same valence of expected utilities), then outcome coercion is true. If there is an action alternative with the valence of utility different from that of the evaluated outcome, then the agent has the freedom to choose at least one alternative to avoid the outcome. So outcome coercion is false. For details of model extension and illustrative example in our virtual training application, the reader may refer to [Mao, 2006].

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 software and 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 behavior judgment. We show how to automatically form such judgment based on an agent's causal knowledge and observations. We also provide empirical results to show the inferences derived by our model are largely consistent with human judgments. Our computational approach is domain-independent, and thus can be generally used to facilitate the design and modeling of human-like social behavior for artificial agents. Current work with our lab is exploiting this model in several ways, including improved models of social emotions (e.g., pride, anger and guilt), extending the capabilities of our agents to reason about the beliefs, intentions and judgments of other social actors (i.e., theory of mind), and informing the modeling of individual personality differences in how people make social attributions.

ACKNOWLEDGEMENTS

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the Air Force Office of Scientific Research under the grant #FA9550-06-1-0206. 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 #60621001 and #60573078, MOST grants #2006AA010106, #2006CB705500 and #2004CB318103, and CAS grants #2F05N01 and #2F07C01 in China.

REFERENCES

[1] Austin J (1962). How to Do Things with Words. Harvard University Press.

[2] Chockler H and Halpern JY (2004). Responsibility and Blame: A Structural-Model Approach. Journal of Artificial Intelligence Research, **22**:93-115.

[3] Conte R and Paolucci M (2004). Responsibility for Societies of Agents. Journal of Artificial Societies and Social Simulation, **7**(4).

[4] Gratch J and 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, pp.339-348.

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

[6] Grice HP (1975). Logic and Conversation. In Cole & Morgan (Eds.). Syntax and Semantics: Vol 3, Speech Acts, pp.41-58. Academic Press.

[7] Jennings NR (1992). On Being Responsible. In Werner & Demazeau (Eds.). Decentralized A.I., pp.93-102. North Holland Publishers.

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

[9] Lickel B, Schmader T and Hamilton DL (2003). A Case of Collective Responsibility: Who Else Was to Blame for the Columbine High School Shootings? Personality and Social Psychology Bulletin, **29**(2):194-204.

[10] Mao W (2006). Modeling Social Causality and Social Judgment in Multi-Agent Interactions. Ph.D. Dissertation. Computer Science Department, University of Southern California.

[11] Mao W and Gratch J (2005). Social Causality and Responsibility: Modeling and Evaluation. Proceedings of the Fifth International Conference on Intelligent Virtual Agents, pp.191-204.

[12] Nadelhoffer T (2006). On Trying to Save the Simple View. Mind and Language, **21**(5):565-586.

[13] Nisbett RE and Wilson TD (1977). Telling More than We Can Know: Verbal Reports on Mental Processes. Psychological Review, **84**(3):231-259.

[14] Norman TJ and Reed C (2001). Delegation and Responsibility. In Castelfranchi & Lesperance (Eds.). Intelligent Agents VII, pp.136-149. Springer-Verlag.

[15] Shaver KG (1985). The Attribution Theory of Blame: Causality, Responsibility and Blameworthiness. Springer-Verlag.

[16] Slovic P, Lichtenstein S and Fischhoff B (1988). Decision Making. In Atkinson, Herrnstein, Lindzey & Luce (Eds.). Stevens' Handbook of Experimental Psychology (Second Edition): Volume 2, Learning and Cognition, pp.673-738. John Wiley & Sons.

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

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

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