Modeling Social Causality and Responsibility Judgment in Multi-Agent Interactions: Extended Abstract

Wenji Mao
State Key Laboratory of Management and Control for Complex Systems, 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

Based on psychological attribution theory, this paper presents a domain-independent computational model to automate social causality and responsibility judgment according to an agent's causal knowledge and observations of interaction. The proposed model is also empirically validated via experimental study.

Introduction

Social causality refers to the inference an entity makes about the social behavior of other entities and self. Such inference differs dramatically from how traditional artificial intelligence methods (e.g., planning) reason about physical reality. Besides physical cause and effect, social causality includes reasoning about mental states (e.g., did the actor intend to cause the outcome? could she foresee the outcome?) and social power (e.g., did the actor have the freedom to act or was she coerced by circumstances or other individuals?). *Responsibility judgment* is the process whereby one forms judgment results about responsibility, credit or blame based on the inference of social causality. Social causality and responsibility judgment underlie how we act on and make sense of the social world around us. They lie at the heart of social intelligence.

Social causality and responsibility judgment are a key aspect of social intelligence, and a model for them facilitates the design and development of a variety of multiagent interactive systems. Social causal reasoning facilitates multiagent planning by augmenting classical planners with the ability to reason about which entities have the power to effect changes. It facilitates adaptive learning by appraising praiseworthy or blameworthy behavior, and reinforcing the praiseworthy. In modeling the communicative and social behavior of human-like agents, responsibility judgment helps inform models of social emotions (Gratch, Mao, and Marsella 2006). As people are usually adept at taking credit and deflecting blame in social dialogue, the information helps guide natural language conversation strategies (Martinovski et al. 2005).

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

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

Related Work

In modeling human social behavior, it is useful to distinguish between *normative*, *descriptive* and *legal* perspectives. Normative models attempt to prescribe how people *should* assign responsibility and blame/credit. Descriptive models characterize what people do *in practice*, which may differ considerably from normative prescriptions. Legal models refer to the formalized processes society uses for responsibility assignment, which can be seen as the amalgam of normative and practical considerations.

This paper is an extended abstract of the JAIR publication (Mao and Gratch 2012).

Normative Models

For the judgment of social causality, responsibility and blame/credit, research on normative models largely resides on moral philosophy where the aim is to identify *rational principles* to govern the assignment of social credit and blame. For example, Kant (1998) argued that, unlike what is often observed in practice, it would be rational to assign the same standards of responsibility regardless of the valence (i.e., praiseworthy or blameworthy) or severity of a social act. Within computer science and artificial intelligence, we are unaware of any other complete models based on the normative principles, with the exception of the computational work done by Chockler and Halpern (2004).

Legal Models

Legal models attempt to formalize responsibility judgment and inferences realized within judicial systems, typically with the aim of automating or verifying human legal judgments. This is a fertile research field at the intersection of artificial intelligence and law, and it has continuously been progressing since the development of early legal systems. There are fundamental differences between the judgments of normative and legal responsibility. Legal judgment largely depends on *specific* circumstances. That is why most legal reasoning systems are *case-based*, whereas evaluating moral responsibility identifies *general* theories that fall within the broad studies of cognitive functionalism.

In addition, researchers have proposed *logic-based* approaches that focus on general reasoning mechanism, typically defeasible inference using non-monotonic reasoning and defeasible argumentation (Hage 1997; Prakken 1997). The main efforts are on the representation of complex legal rules (e.g., contradictory, nonmonotonic and priority rules), inference with rules and exceptions, and handling conflict rules (Prakken and Sartor 2002). However, a layman's judgment of behavior in everyday situations is not quite the same as that made in the court. Not only does it occur in richer forms of social interaction, but it follows different set of rules.

Descriptive Models

Descriptive models attempt to characterize how people form social judgments in practice, which can differ from both the presumed normative principles and legal judgments. Descriptive models also differ in their criteria for validation. Whereas normative models are judged by their consistency with universal principles such as fairness and legal models are judged by their consistency with past legal decisions, descriptive models are assessed by their agreement with the judgments people form in their day-to-day lives. Research on descriptive models largely resides on *social psychology* (Heider 1958; Shaver 1985; Weiner 1995, 2001) and there is little work within artificial intelligence on attributing responsibility and blame/credit in a human-like fashion.

Computational Approaches

In AI and causality research, computational approaches were developed to address the problem by extending causal models (Halpern and Pearl 2001; Chockler and Halpern 2004). Halpern and Pearl (2001) presented 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 cause. A *causal model* (or a structural model) is a system of equations over a set of random variables (i.e., exogenous or endogenous variables). The values of exogenous variables are determined by factors outside the model. Each endogenous variable has exactly one *causal equation* that determines its value.

Causal inference is based on *counterfactual dependence* under some contingency. Chockler and Halpern (2004) further extended this notion of causality, to account for *degree of responsibility*. They provide a definition of degree of responsibility based on the consideration of contingencies. 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. Grounded on the *philosophical* principle (i.e., counterfactual reasoning), their extended definition of responsibility accounts better for multiple causes and the extent to which each cause contributes to the occurrence of a specific outcome. Another advantage of their model is that their definition of degree of blame takes an agent's epistemic state into consideration. However, they only consider one epistemic variable, that is, an agent's knowledge prior to action performance. Important concepts in moral responsibility, such as intention and freedom of choice are excluded from their definition.

Attribution Theory for Behavioral Judgment

Most contemporary psychological studies of social causality and responsibility judgment draw on attribution theory. Attribution research views that social perceivers make sense of the world by attributing behavior and events to their underlying causes. Two influential attributional models for social causality, responsibility and blame (or credit) are those proposed by Shaver (1985) and Weiner (1995), which identify the underlying key factors (i.e., attribution variables) people use in behavioral judgment.

The assessments of *physical causality* and *coercion* identify the responsible party. That is, 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, depending on the degree of coercion.

Intention and *foreseeability* determine the degree of responsibility. 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 foresee that the action brings about the outcome. The higher the degree of intention/foreseeability, the greater the responsibility assigned.

Weiner (2001) distinguished between act intentionality and outcome intent. An agent may intentionally perform an action (i.e., *act intention*), but may not intend all the action effects (i.e., *outcome intention*). It is outcome intention rather than act intention that are key in responsibility and behavior judgment. Similar difference exists in *outcome coercion* and *act coercion*. The result of the judgment process is the assignment of certain blame or credit to the responsible party. The *intensity* of blame or credit is determined by the severity or positivity of the outcome as well as the degree of responsibility.

Proposed Computational Model

Attribution theory identifies the general process and key variables people use in judging social behavior. However, this process and the variables are not directly applicable to computational systems, as they are described at an abstract conceptual level that is insufficiently precise from a computational perspective. To bridge the gap between conceptual descriptions of the theory and actual components in current intelligent systems, we need to develop the computational mechanisms that automatically convert the conceptual descriptions into a functionally workable model in use for intelligent systems.

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

Representations

Causal Knowledge

Causal knowledge is encoded via a hierarchical plan 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 multiple ways and each decomposition is one *choice* of executing the action. Different choices of action execution are *alternatives* each other. If an abstract action can be decomposed in multiple ways, it is a *decision node* (i.e., *or node*) and an agent must decide amongst the alternatives. Otherwise, it is a *non-decision node* (i.e., *and node*) and execution of the action is realized via executing all its *subactions*.

A *plan* is a set of actions to achieve certain intended *goal(s)*. *Outcomes* (we use them as exchangeable) are those desirable or undesirable action effects. The desirability of action effects is represented by utility values. To model the power relationships of agents, each action in a plan is associated with a *performer* and an agent who has *authority* over its execution.

Communicative Events

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

Attribution Variables

Causality refers to the relationship between cause and effect. In our approach, we encode causal knowledge about actions (i.e., human agency) and the effects they produce via plan representation. Act intention is represented using *intend* and *do*, outcome intention using *intend* and *achieve*, and the connection between act and outcome intentions using *intend* and *by*. We use *know* and *bring about* to represent foreseeability. Act coercion is represented using *coerce* and *do*, outcome coercion using *coerce* and *achieve*. In addition, we consider two important concepts in modeling *coercion*, social *obligation* and (*un*)*willingness*.

Notations

We use first-order logic as a formal tool to express 27 predicates and 15 functions in our model (Mao and Gratch 2012). The first ten predicates denote the features related to plan structure and action execution. Another six predicates represent communicative acts. Predicates *cause*, *assist-cause*, *know*, *want*, *obligation*, *intend* and *coerce* describe the epistemic variables (including attributions) used for inferring intermediate beliefs. The last four variables represent the power relationship and capabilities of agents. Besides predicates, the first seven functions denote the generic features in (hierarchical) plan representation. Another four functions describe *indefinite effect*, r*elevant action/effect* and *side effect* sets, and the last four functions represent the agents involved.

Reasoning about Social Causality

Social causality and responsibility judgment is always from a perceiving agent's subjective perspective. The perceiver uses her knowledge and observation of behavior to infer beliefs of social attributions. We design automatic methods of causal and dialogue reasoning to realize such a mechanism.

Dialogue Inference

Language communication between agents is a rich source of information for attributing behavior (Hilton 1990). We assume communication between agents is *grounded* (Traum 1994), and conforms to Grice's maxims of *Quali-*

ty and *Relevance* (Grice 1975). We identify a small number of commonsense rules that allow a perceiving agent to derive beliefs about attribution variables from social communication.

For example, a *request* shows what the speaker wants. An *order* shows what the speaker intends (*Rule D5*). If requested or ordered by a superior, it creates a social obligation for the hearer to perform the content of the act (*Rules D4&D6*). Note that *s*, *h*, *p* and *ti* denote speaker, hearer, proposition and time, respectively. To simplify logical forms, universal quantifiers are omitted in the rules. Also, to further simplify the expression, we introduce a predicate *etc*, which is similar to that used in (Hobbs et al. 1993) and stands for the absence of contradictory situations.

Rule *D4* [*superior-request*]:
request(s, h, p, $t1$) $\wedge$ superior(s, h) $\wedge$ $t1<t2$ $\wedge$ etc_4 $\Rightarrow$ obligation(h, p, s, $t2$)
Rule *D5* [*order*]:
order(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc_5 $\Rightarrow$ intend(s, p, $t2$)
Rule *D6* [*order*]:
order(s, h, p, $t1$) $\wedge$ $t1<t2$ $\wedge$ etc_6 $\Rightarrow$ obligation(h, p, s, $t2$)

The hearer may *accept, reject* or *counter-propose* an order (or request). Various inferences can be made depending on the response of the hearer and the social relationship between the speaker and the hearer. For instance, if the hearer accepts, and there is no obligation beforehand or the hearer is willing to, it can be inferred that the hearer intends (*Rules D7&D8*). In another case, if an agent is obviously unwilling to but accepts the obligation, there is *clear* evidence of coercion (*Rule D10*).

Rule *D7* [*accept*]:
$\neg$obligation(h, p, s, $t1$) $\wedge$ accept(h, p, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc_7 $\Rightarrow$ intend(h, p, $t3$)
Rule *D8* [*willing-accept*]:
want(h, p, $t1$) $\wedge$ accept(h, p, $t2$) $\wedge$ $t1<t2<t3$ $\wedge$ etc_8 $\Rightarrow$ intend(h, p, $t3$)
Rule *D10* [*unwilling-accept-obligation*]:
$\neg$intend(h, p, $t1$) $\wedge$ obligation(h, p, s, $t2$) $\wedge$ accept(h, p, $t3$) $\wedge$ $t1<t3$ $\wedge$ $t2<t3<t4$ $\wedge$ etc_{10} $\Rightarrow$ coerce(s, h, p, $t4$)

Causal Inference

Causal inference is a plan-based evaluation of agency, intention, foreknowledge and coercion based on the causal information provided by plan representation.

Given the domain theory *DT*, observed executed actions and an outcome *e*, the performer of an action *A* that directly causes *e* is the *causal agent*. Other performers of relevant actions to achieve *e* have *indirect agency*. In the absence of coercion, causal agent is deemed responsible for *e*, while other agents assist causing *e* should share responsibility with this causal agent.

Causal inference helps infer outcome intention from evidence of act intention. For example, when an action has multiple effects, in order to identify whether a specific outcome is intended or not, a perceiver may examine action *alternatives* the agent intends and does not intend, and compare the effects of intended and unintended alternatives. Intention inference also helps evaluate an agent's foreknowledge, as intention entails foreknowledge: if an agent intends an action *A* to achieve an effect *e* of *A*, then the agent must *know* that *A* brings about *e*.

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. If being coerced to execute an abstract action and the coerced action has multiple decompositions, then the subsequent actions are not coerced. Since the agent has options, only the effects that appear in all alternatives are unavoidable, and other effects that only appear in some (but not all) alternatives are avoidable, so they are not coerced. An agent can be indirectly coerced by enabling/disabling action preconditions, or blocking other action alternatives (Mao and Gratch 2012).

Algorithm and Evaluation

We have developed an algorithm to implement the attribution process (Mao and Gratch 2012). The algorithm first infers from dialogue evidence, and then it applies causal inference rules. During each loop, if outcome coercion is inferred, the authority is deemed responsible. Finally, the algorithm returns the responsible agents and assigns proper credit or blame to these agents.

To empirically evaluate the proposed model, we design and conduct two experimental studies. The experimental results show that our model predicts human judgments of social attributions and makes inferences consistent with what most people do in their judgments (Mao and Gratch 2012).

Conclusion

We model a key aspect of social intelligence in this paper, by formalizing the underlying social reasoning process in people's behavioral judgment. We show how AI knowledge representation and reasoning methods can be utilized to automate social inference and judgment process, and conduct human experiments to empirically evaluate the computational model. Thus the proposed model can be generically incorporated into an intelligent system to augment its social and cognitive functionality.

Acknowledgements

This work was supported in part by NNSFC grants #61175040, #71025001 and #91024030, AFOSR under grant FA9550-09-1-0507, and US Army RDECOM. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

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

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

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

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

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

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

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*.

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

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

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

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

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

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

Mao, W. and Gratch, J. 2012. Modeling Social Causality and Responsibility Judgment in Multi-Agent Interactions. *Journal of Artificial Intelligence Research*, 44:223-273.

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

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

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

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

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

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

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.

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

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

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