Appears in the proceedings of the 3rd International Conference on Autonomous Agents and Multi Agent Systems, New York, 2004

Social Judgment in Multiagent Interactions

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

Social judgment is a process of social explanation whereby one identifies which entities deserve credit or blame for multiagent activities. Such explanations are a key aspect of inference in a social environment and a model of this process can advance several design components of multi-agent systems. Social judgment underlies social planning, social learning, natural language pragmatics and computational model of emotion. Based on psychological attribution theory, this paper presents a computational approach to forming social judgment based on an agent's causal knowledge and communicative interactions with other agents.

1. Introduction

People rarely use simple causal interpretations when explaining social actions. In contrast to how causality is used in the physical sciences, people instinctively seek out individuals for their everyday judgments of credit or blame. Such judgments are a fundamental aspect of social intelligence. They involve evaluations of not only causality but also individual responsibility and free will [Shaver, 1985]. They imply how we act on and make sense of the social world and lie at the heart of social intelligence.

Social explanations make distinctions beyond traditional causal explanations, and social attributions are crucial for successful interactions with intelligent entities. With the advance of multi-agent interactive systems and the increasing sophistication of systems that socially interact with people, it is increasingly important to reason about this human-centric form of social inference. Social judgment can facilitate social planning by augmenting classical planners with the ability to reason about which entities have control to effect changes. It can facilitate social learning by evaluating behavior as creditworthy or blameworthy and reinforcing the creditworthy. In modeling the communicative and social behavior of human-like agents, social judgment helps inform which entities deserve credit or blame for specific outcomes. As people are usually adept at taking credit and deflecting blame in social situations, the information can help guide strategies of natural language conversations and inform models of emotion [Gratch and Marsella, 2004].

Our work is motivated by obvious defects in the multi-agent reasoning underlying a training application developed at our lab [Rickel *et al*., 2002]. For example, during a training exercise, a trainee (acting as the commander of a group of agents) ordered his second-in-command (the sergeant played by an autonomous agent) to adopt a course of actions that the agent considered highly undesirable. The trainee persisted with his decision even after being told of its undesirable consequence and of better alternatives. The command was such that it could not be executed directly, but rather the agent had to, in turn, order his own subordinates to perform the act. The current model assigns blame to the subordinates as they directly caused the action with the undesirable consequence. As a result, the agent expressed anger towards his subordinates, *even though he commanded them to perform the offensive act.* Human observers, instead, universally judge the situation and assign blame chiefly to the trainee, as the agent was clearly following orders and even attempted to negotiate for a different outcome. Such results indicate an impoverished capacity to judge credit or blame in social context.

People differ in how they evaluate a specific situation, but psychologists and philosophers agree on the broad dimensions people use for such judgments. This paper lays out a computational model of forming social judgment based on psychological attribution theory. Attribution theory identifies key variables used to explain agents' behavior, and the variable values are applied in the evaluation process to form the judgment. In the remainder of the paper, we first introduce attribution models for social judgment. Then based on the introduced models, we discuss the computational approach we employ to tackle the problem of social judgment.

2. Attribution Theory for Social Judgment

Social judgment has been studied extensively in moral philosophy [Williams, 1995], law [Hart, 1968], and social psychology [Weiner, 1995; Shaver, 1985]. Traditions differ to the extent that the models are prescriptive (i.e., what is the "ideal" criterion that people ought to conform in their judgments) or descriptive (i.e., what do people actually do in their judgments). Much of the work on AI has focused on identifying ideal principles of responsibility (e.g., legal codes or philosophical principles) and ideal mechanisms to reason about this, typically contradictory principles [McCarty, 1997; Chockler and Halpern, 2003]. As our primary goal is to inform the design of realistic virtual humans that mimic human behavior [Gratch *et al*, 2002], our focus is on descriptive models and we are particularly influenced by psychological *attribution theory*.

Attribution theory has been the favored psychological theory of folk explanations of behavior for decades [Malle, 2001]. The models of Weiner [1995] and Shaver [1985] are among the most influential ones in responsibility and blame attribution. In Shaver's model, the judgment of responsibility is a multi-step process initiated by events with positive or negative consequence. Dimensions of responsibility include causality, foreseeability, intention, coercion and excuse (currently excuse is not included in our current model). First one assesses causality, identifying the agent that caused the outcome (i.e., the causal agent). Then the process proceeds by assessing other key factors: Did the agent *foresee* the outcome; Was it the agent's *intention* to produce the outcome; Did the agent have *choice* or the agent was forced under *coercion* (e.g., acting under power or by certain social obligations)? As the last step of the process, proper credit or blame is assigned to the responsible party. In the above example, we may infer from the conversation that the trainee foresaw the consequence and coerced the sergeant to follow the undesirable course of actions. Baring the unknown mitigating factors, we would likely conclude that the trainee is primarily responsible for the outcome.

Social attributions involve evaluating consequences of events with personal significance to an agent. The evaluation is always from a perceiving agent's perspective and the significance of the consequences is based on individual perceiver's preferences. The perceiver uses her own knowledge about the observed agents and her observations to form beliefs about the observed agents. The attribution values acquired by the perceiver are then used in the attribution process to form an overall judgment[1]. Given the same situation, as different perceivers have different preferences, different knowledge and observations, they may form different beliefs and thus judge the same situation differently.

Nevertheless, the attribution process and inferences are general, and applied uniformly to different perceivers. Following Weiner [1995], we use coercion to determine responsible agents, and intention and foreseeability for assigning the intensity of credit or blame[2]. If an event brings about positive/negative consequence, and there is no clear evidence of coercion, then the *causal agent* is responsible for the consequence and credit or blame is assigned to this agent. Otherwise, the *coercer* is credit- or blameworthy. In a multi-agent setting, a performer often causes an outcome through the assistance of other agents or a coercer coerces an agent through the assistance of other agents. Therefore, we also need to consider indirect performers and coercers who are partially responsible.

In human social interactions, attribution variables are acquired from various sources: from observation of behavior, from statements made through natural language, from causal information and models built up through past interactions, stereotypes and culture norms. In this paper, we show how to derive the variable values by inferring natural language conversation and causal knowledge, and how the variables are utilized in the algorithm and process to form an overall judgment.

3. Representation

To inform social judgment, we need to represent the knowledge and inferential mechanism that impact the attribution process.

3.1 Plan Knowledge

Causal reasoning plays a central role in deriving attribution variables. In our approach, causal knowledge is represented via probabilistic plan representation. Each *action* consists of a set of propositional preconditions and effects. Actions can have non-deterministic effects (denoted as *effect_prob*) and/or conditional effects. To represent the success and failure of action execution, actions have execution probability (denoted as *execute_prob*). The likelihood of preconditions and effects is represented by probability values. The desirability of action effects (i.e., their positive/negative significance to an agent) is represented by utility values [Blythe, 1999].

In a hierarchical plan representation, an action can be *primitive* (i.e., an action that can be directly executed by an agent) or *abstract*. An abstract action may be decomposed in multiple ways and each decomposition consists of a sequence of primitive or abstract sub-actions. A *non-decision node* in plan structure is an action that can only

[1] A perceiver's knowledge may not necessarily reflect "the truth", and there might be errors in observations and judgment process.

[2] Note that these models differ in terminology. Here we adopt the terminology of Shaver.

be decomposed in one way. A *decision node*, on the other hand, can be decomposed in multiple ways and an agent must decide amongst the options. The options at a decision node are called *action alternatives* (with respect to the decision node). A primitive action is a non-decision node, while an abstract action can be a decision node or a non-decision node.

A *plan* is represented as an action sequence. Each plan has preconditions and outcomes, and is associated with an intended *goal*. When a plan contains abstract actions, each decomposition of the abstract actions yields a *fully expanded plan* (i.e., a primitive action sequence). There might be more than one fully expanded plan available to achieve a goal, and the options are called *plan alternatives* (with respect to the goal). The utility of a plan measures the overall benefit and disadvantage of the plan.

Consequences or outcomes (we use the terms as exchangeable) are represented as primitive action effects with non-zero utilities. In a hierarchical plan representation, *consequences* of an abstract action are determined by its descendents as follows: Consequences of a non-decision node are the aggregation of the consequences of its descendents. Consequences of a decision node are the set of its common consequences (i.e. consequences occur among all the action alternatives). Consequences of a plan are the aggregation of the consequences of the actions that constitute the plan.

In addition, each action in a plan is associated with a *performer* (i.e., the agent performing the action) and an agent who has *authority* over its execution. The performer cannot execute the action until authorization is given by the authority. This represents the hierarchical organizational structure of social agents.

3.2 Attribution Variables

Now we revisit the key conceptual variables underlying attribution theory.

Causality refers to the connection between actions and the effects they produce. Causal information is encoded via *plan representation*. In our approach, plan representation can be hierarchical or non-hierarchical. Interdependencies between actions are represented as a set of causal links and threat relations.

Foreseeability refers to an agent's foreknowledge about actions and consequences. If an agent knows that an action likely brings about certain consequence before action execution, then the agent foresees the consequence of the action. We use *know* with *bring about* to represent foreseeability.

Intention is generally conceived as a commitment to work toward certain act or outcome. Most theories argue that outcome intention (i.e., intention to bring about an outcome) rather than act intention (i.e., intention to perform an action) is key in determining accountability. Moreover, intended outcomes usually deserve much elevated accountability judgments [Weiner, 2001]. We use *intend* with *do* to represent act intention and *intend* with *achieve* for outcome intention. We extend the concept of intention in [Bratman, 1987; Grosz and Kraus, 1996], to consider indirect situations in multiagent interactions. For example, an agent intends an action or a consequence, but may not be the actor that actually performs the action (i.e., by intending another agent to act or achieve the consequence). Another situation is that an agent intends to act but is coerced to do so (see *axiom 1* below).

Coercion is to persuade an agent forcefully to act or achieve a proposition that the agent is unwilling to. An agent may be coerced to act (i.e., act coercion) yet not be coerced to achieve any outcome of the action (i.e., outcome coercion), depending on whether the agent has choices in achieving different outcomes amongst alternatives. It is outcome coercion that actually affects our judgment of behavior, and is used to determine the responsible agents. We use *coerced* with *do* to represent act coercion and *coerced* with *achieve* for outcome coercion.

4. Inferences

To infer attribution variables, we extract evidence from agents' interactions and knowledge states of agents. Two important sources of evidence contribute to the inference process. One source is the actions performed by the observed agents (including physical acts and communicative acts). The other is the causal evidence about the observed agents, represented as plan knowledge. We introduce commonsense heuristics that allow an agent to make inference based on this evidence.

There are interrelations between the attribution variables. We identify them as the properties of the variables and express them as axioms. These axioms can be used as supplementary inference rules.

4.1 Axioms

The variables x, y and z denote agents (x and y should be different agents). Let A be an action and p be a proposition. The following *axioms* hold from a rational agent's perspective.

(1) $\exists y(\mathit{coerced}(x, A, y)) \Rightarrow \mathit{intend}(x, A)$
(2) $\mathit{intend}(x, \mathit{do}(z, A)) \wedge \neg(\exists y)(\mathit{coerced}(x, A, y)) \Rightarrow \exists p(p \in \mathit{consequence}(A) \wedge \mathit{intend}(x, \mathit{achieve}(z, p)))$
(3) $\mathit{intend}(x, \mathit{achieve}(z, p)) \Rightarrow \exists A(p \in \mathit{consequence}(A) \wedge \mathit{intend}(x, \mathrm{do}(z, A)))$
(4) $\mathit{intend}(x, \mathit{do}(z, A)) \wedge p \in \mathit{consequence}(A) \wedge \mathit{intend}(x, \mathit{achieve}(z, p)) \Rightarrow \mathit{know}(x, \mathit{bring\text{-}about}(A, p))$

The *first* axiom shows that act coercion entails act intention. It means that if an agent is coerced an action A by another agent, then the coerced agent intends A. The second and the third axioms show the relations between act intention and outcome intention. The *second* one means that if an agent intends an action A (either by performing the action or by intending another agent to perform), and the agent is not coerced to do so (i.e., a voluntary act), then the same agent must intend at least one consequence of A. The *third* means that if an agent intends an outcome p, the same agent must intend at least one action that leads to p. The *last* one shows the relation between intention and foreseeability. It means that if an agent intends an action A to achieve a consequence p of A, the same agent must know that A brings about p.

4.2 Inferring Communication Events

Conversation communication is a rich source of information [Cohen *et al*, 1990]. 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 [Austin, 1962] and beliefs acquired previously.

Assume conversations between agents are *grounded* [Traum and Allen, 1994] and they conform to Grice's maxims of *Quality*[3] and *Relevance*[4] [Grice, 1975]. Social information (agents' roles, relationship, etc) is also necessary to interpret the speech acts. For example, an order can only be issued to a subordinate and have as its effect a social obligation for the subordinate to perform the content of the act.

For our purpose, we analyze negotiation in task-oriented dialogue [Traum *et al*, 2003] and focus on the speech acts that help infer dialog agents' desires, intentions, foreknowledge and choices in acting. We have designed commonsense rules that allow a perceiving agent to infer from dialog patterns. These rules are generally designed. They can be combined and applied flexibly to variable multi-party conversations (e.g., hybrid human-agent team).

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 the speaker's desire (or want). An order shows the speaker's intent. The hearer may *accept, reject* or *counter-propose* the order/request. Various inferences can be made depending on the response and the power relationship between the speaker and the hearer. For instance, if the hearer accepts what the speaker wants or intends (an agent can accept by actually *trying* the act), it can be inferred that the hearer intends; if the hearer accepts what the superior intends, and the hearer is not believed to want or intend beforehand, there is evidence of coercion.

For the complete version of inference rules, the reader may refer to [Mao and Gratch, 2003].

4.3 Causal Inference

Plan representation and plans provide further evidence for inferring agency, intention and coercion, in both direct and indirect cases.

4.3.1 Direct and Indirect Agency. The performer of the action that directly causes a specific outcome is the *causal agent* for the outcome. In multi-agent plan execution, the preconditions of an action might be established by the activities of other agents. These other agents are viewed as indirect agency that helps causing the outcome. Given an executed action set, observed action effects and a specific outcome p, the following actions and effects are *relevant* to achieving p:

- Action A that causes p is relevant.
- The actions and effects that establish a precondition of a relevant action to achieve p are relevant.
- If p is enabled by the consequent of a conditional effect of A, the actions and effects that establish the antecedent of the conditional effect are relevant.
- If a precondition of a relevant action is enabled by the consequent of a conditional effect of an action, the actions and effects that establish the antecedent of the conditional effect is relevant.

In the absence of coercion, the causal agent for p is the primary responsible agent. Other performers of relevant actions to achieve p are the secondary responsible agents.

4.3.2 Direct and Indirect Coercion. Causal agent may perform an action or achieve a specific outcome under coercion. The coercer can be another party ranking higher in power relation, a social obligation, etc. One can infer coercion by examining plan structure and alternatives, and the activities of other agents.

If an agent is coerced to execute a primitive action, the agent is also coerced to achieve the action effects. In a hierarchical plan representation, if being coerced to execute an abstract action and the action is a non-decision node in plan structure, then the agent is also coerced to achieve the outcomes of subsequent actions, because there is no other choice. If the coerced action is a decision node, the agent has choices. Even if the agent is coerced the action, it does not follow that a specific outcome of the action is coerced. In a decision node, if an outcome is common among all the action alternatives, then it is unavoidable: outcome coercion is true. Otherwise, the agent has options to choose an alternative action to avoid the outcome, so outcome coercion is false.

[3] Quality Maxim: One ought to provide true information in conversation.
[4] Relevance Maxim: One's contribution to conversation ought to be pertinent in context.

If an agent is coerced to achieve a goal and there is no plan alternative (i.e., only one plan available to achieve the outcome), then the plan is coerced: the agents are coerced to execute the only plan and achieve all the outcomes of the plan. If an agent is coerced to achieve a goal and plan alternatives are available, then the evaluation process needs to compute utilities of plan alternatives. If there is a plan alternative with a different utility value (e.g., current plan has a negative utility value but a plan alternative has a positive value), then the agents have options to choose an alternative plan to avoid the outcome. So the plan is not coerced and outcome coercion is false. If other agents block all the plan alternatives with different utilities, these other agents help coercing the plan as well as the outcomes of the plan.

When outcome coercion is true, the responsible agents are redirected to the coercers. If coercing the evaluated outcome, then the coercers are the primary responsible agents. If coercing the relevant effects in order to achieve the outcome, the coercers are the secondary responsible agents.

4.3.3 Intention Recognition. Outcome intention can be partially inferred from evidence of act intention and comparative features of the consequence set of action alternatives. According to *Axiom 2*, if an agent intends a voluntary action *A*, the agent must intend at least one consequence of *A*. When an action has multiple consequences, a perceiver may examine action alternatives the agent intends and does not intend, and compare the consequence set of the intended and unintended alternatives. If an agent intends an action *A* voluntarily and does not intend its alternative *B*, it can be inferred that the agent intends at least one consequence that only occurs in *A*. If there is only one such consequence, infer that the agent intends the consequence.

If there is no clear belief about intention derived from dialogue inference, another source of evidence is the possible goals and preferences of the observed agents. Intention is detected via general plan/goal recognition technique. As utilities of states are available in many real-world applications [Blythe, 1999] as well as in our own, comparing with the typical plan recognition approaches [e.g., Schmidt *et al*, 1978; Allen and Perrault, 1980; Kautz, 1991; Charniak and Goldman, 1993], we take states and state desirability into consideration. We view plan recognition as inferring the decision-making strategy of other agents and assume that a rational agent will adopt a plan that maximizes the expected utility.

The computation of expected plan utility is similar to that in decision-theoretic planning (e.g. DRIPS, [Haddawy and Suwandi, 1994]), using the utilities of outcomes and the probabilities with which different outcomes occur. However, in our approach, we use the observations as evidence to incrementally update state probabilities and the probabilities of action execution, and compute an exact utility value rather than a range of utility values as in decision-theoretic planning.

We use plan utility in two ways in our approach. Here we use the expected utility of a plan as a criterion for disambiguation in intention recognition. State utilities represent the observed agents' preferences. The other way of using plan utility is to represent the overall benefit and disadvantage of a plan. Since the attribution process is from an observing agent's perspective, in the latter way, state utilities represent the observer's preferences.

Given the observed executed actions, the plan recognizer selects a candidate plan in plan library with the highest expected utility. As current plan is identified (with probability), act/outcome intention can be evaluated by checking whether the action/outcome are relevant to the goal attainment. The detailed formulae for computing action/outcome probability and plan utility can be found in [Mao and Gratch, 2004].

4.4 Evaluation Algorithm

We have developed an algorithm for evaluating the responsible agents for a specific outcome *p* (*A* is the action causing *p*). Initially, by default, the algorithm assigns the performer of each relevant action to its coercer (*step 1&2*). Then it searches dialogue history and infers dialogue evidence (*step 3*). If a goal is coerced (*step 4*), the algorithm computes utilities of plans and infers plan alternatives (*step 5*). If a plan is coerced (*step 6*), then each relevant action in the plan is coerced by the coercers of the goal/plan (*step 7*). If a relevant action is coerced (*step 8.1*), infer action alternatives (*step 8.2*). If a relevant action effect is coerced (*step 8.3*), assign the superior to the coercer (*step 8.4*). As coercion may occur in more than one level of plan hierarchy, the superior here may not be the direct authority of the performer. Finally, the algorithm assigns the coercers to the responsible agents (*step 8*).

Algorithm (*p*, *utility-function*):

1. FOR each $B \in$ *relevant-action*(*p*)
2. *coercer*(*B*)=*performer*(*B*)
   END-FOR
3. Search dialog history and apply dialog inference rules
4. IF a *goal* is coerced
5. Compute plan utilities and apply plan inference rules
6. IF a *plan* is coerced
7. FOR each $B \in$ *relevant-action*(*p*)
   7.1 IF $B \in$ *plan*
   7.2 *coercer*(*B*)=*coercer*(*goal*)
   END-IF
   END-FOR
   END-IF

8. ELSE
   FOR each $B \in$ *relevant-action*(p)
   8.1 IF B is coerced
   8.2 Apply action inference rules
   8.3 IF $e \in$ *relevant-effect*(p)$\cap$*effect*(B) is coerced
   8.4 *coercer*(B)=*superior*(*performer*(B))
   END-IF
   END-IF
   END-FOR
7. *P-responsible*(p)=*coercer*(A)

$$S\text{-}responsible(p)=\Big(\bigcup_{C\in relevant-action(p)} coercer(C)\Big)-P\text{-}responsible(p)$$

END-IF

5. Illustration

Now we return to the example introduced earlier in the paper. Several social actors are involved in the example, the student, the sergeant and squad leaders. The student is a human trainee, acting as a superior of the sergeant. Squad leaders act as subordinates of the sergeant. Conversations between agents are represented via speech acts and a dialog history is accessible in the system.

In the scenario, the student's mission is to "*helping-eagle-1-6*", which is a desirable team goal for the troop. Two plans are available to achieve this goal in the plan library, namely, P_1 and P_2. Plan P_1 is composed of actions "*assemble*", "*one-squad-forward*" and "*remaining-squads-forward*". Action "*remaining-squads-forward*" achieves the goal "*helping-eagle-1-6*" (with effect probability 0.75). Plan P_2 consists of actions "*assemble*", "*two-squads-forward*" and "*remaining-squads-forward*", in which "*two-squads-forward*" achieves "*helping-eagle-1-6*" (with effect probability 0.8), but also brings about the outcome "*unit-fractured*". Besides, "*one-squad-forward*" and "*remaining-squads-forward*" compose the abstract action "*sending-one-squad-forward*", and "*two-squads-forward*" and "*remaining-squads-forward*" compose "*sending-two-squads-forward*". The performer and authority of each action, state probabilities and utilities are shown in the figure. The execution probability of each action is assigned 0.95.

The dialogue history includes the following acts, ordered by the time the speakers addressed them ($t1<t2<\ldots<t6$. *std*, *sgt* and *sld* stand for the student, the sergeant and squad leaders, respectively).

(1) *order*(*std*, *do*(*sgt*, *sending-two-sqds-fwd*), *t1*)
(2) *tell*(*sgt*, *std*, *bring-about*(*sending-two-sqds-fwd*, *unit-fractured*), *t2*)
(3) *counter-propose*(*sgt*,, *do*(*sgt*, *sending-two-sqds-fwd*), *do*(*sgt*, *sending-one-sqd-fwd*), *t3*)
(4) *order*(*std*, *do*(*sgt*, *sending-two-sqds-fwd*), *t4*)
(5) *accept*(*sgt*, *do*(*sgt*, *sending-two-sqds-fwd*), *t5*)
(6) *order*(*sgt*, *do*(*sld*, *1st-and-4th-sqds-fwd*), *t6*)
(7) *try*(*sld*, *do*(*sld*, *1st-and-4th-sqds-fwd*), *t7*)
… …

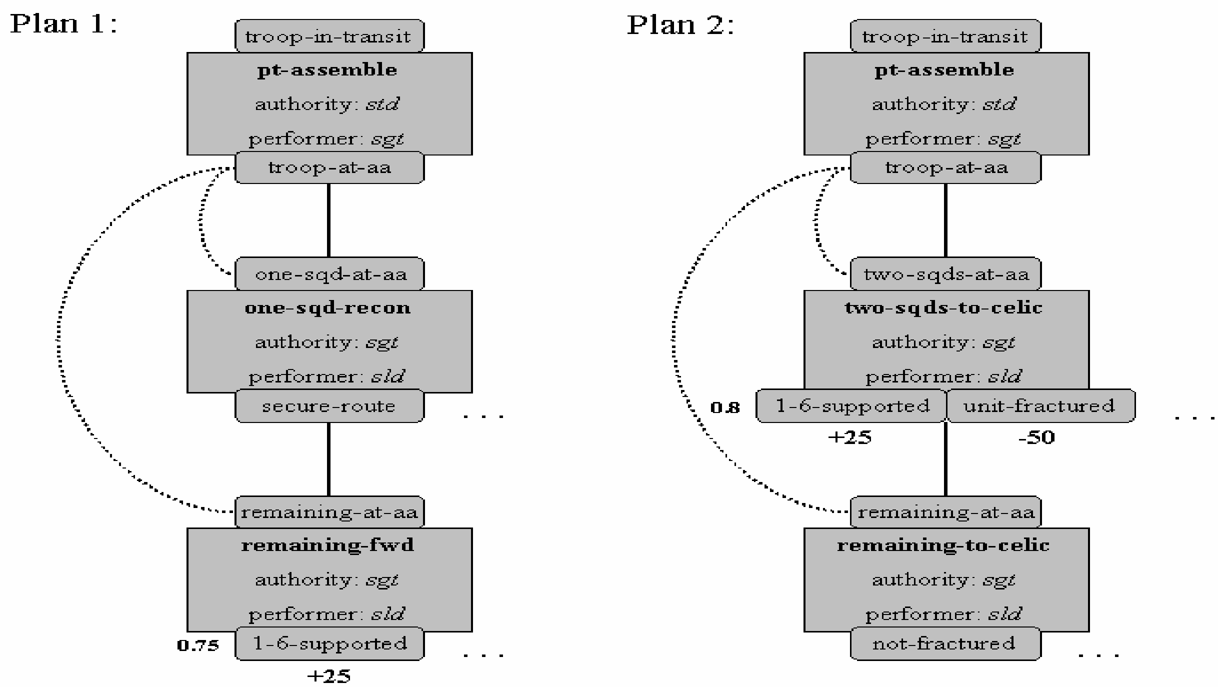
Take the sergeant's perspective as an example. The sergeant perceives the above conversation between the agents and action execution. Assume the sergeant assigns negative utility to *unit fractured* and this consequence serves as input to the evaluation algorithm. We illustrate how to find the responsible agent given the sergeant's causal knowledge and observations.

The observed action sequence of the troop, *assemble* and *1st-and-4th-sqds-fwd* (an instance of *two-sqds-fwd*) support plan P_2 (utility(p_1)=16.9; utility(p_2)=20). So plan P_2 is the current hypothesized plan (with probability). As *1st-and-4th-sqds-fwd* directly causes *unit-fractured*, the performer *sld* is the *causal agent* for the outcome. As *assemble* establishes the precondition of *1st-and-4th-sqds-fwd* and *sgt* is the performer, *sgt* is the indirect agency for the evaluated outcome.

Step 1-2: 1st-and-4th-sqds-fwd and *assemble* are the relevant actions to achieve the outcome *unit-fractured*. Initially, assign the performer of each relevant action to its coercer.

Step 3: From the observed speech acts in communication, the sergeant can derive a number of beliefs:

(1) *intend*(*std*, *do*(*sgt*, *sending-two-sqds-fwd*))
    (*Act* 1 "*order*")
(2) *know*(*sgt*, *bring-about*(*sending-two-sqds-fwd*, *unit-fractured*))
    (*Act* 2 "*tell*")
(3) *know*(*std*, *bring-about*(*sending-two-sqds-fwd*, *unit-fractured*))
    (*Act* 2 "*tell*")
(4) *know*(*sgt*, *alternative*(*sending-one-sqd-fwd*, *sending-two-sqds-fwd*))
    (*Act* 3 "*counter-propose*")
(5) *know*(*std*, *alternative*(*sending-one-sqd-fwd*, *sending-two-sqds-fwd*))
    (*Act* 3 "*counter-propose*")
(6) *want*(*sgt*, *sending-one-sqd-fwd*)
    (*Act* 3 "*counter-propose*")
(7) ¬*want*(*sgt*, *sending-two-sqds-fwd*)
    (*Act* 3 "*counter-propose*")
(8) ¬*intend*(*std*, *do*(*sgt*, *sending-one-sqd-fwd*))
    (*Belief* 5 and *Act* 4 "*order*")
(9) *coerced*(*sgt*, *sending-two-sqds-fwd*, *sld*)
    (*Belief* 1, *Act* 5 "*accept*" and "*superior*")
(10) *intend*(*sgt*, *do*(*sld*, *1st-and-4th-sqds-fwd*))
    (*Act* 6 "*order*")
(11) *coerced*(*sld*, *1st-and-4th-sqds-fwd*, *sgt*)
    (*Belief* 10, *Act* 7 "*try*" and "*superior*")



Step 4: The student had the obligation of fulfilling his mission. The student is coerced by the goal of *helping eagle 1-6*.

Steps 5-6: There are plan alternatives P_1 and P_2 to achieve the goal. By computing the utilities of P_1 and P_2 (utility(p_1)=17.8; utility(p_2)=-30), the sergeant knows that there are alternative plans to help eagle 1-6, and the plan alternative has a different utility value. There are no other agents' activities to block the plan, so *std* was not coerced to execute the plan. Nor was *std* coerced to fracture the unit.

Step 8: Since the primitive action *1st-and-4th-sqds-fwd* was coerced and *sld* was the performer, *sld* was coerced to achieve the outcomes *helping-1-6-supported* and *unit-fractured* by *sgt*. Since *sgt* was coerced abstract action *sending-two-sqds-forward* and the action is a non-decision node, *sgt* was coerced *helping-1-6-supported* and *unit-fractured* by *std*. So *std* is the coercer of *1st-and-4th-sqds-fwd*. Assign *std* to the *primary* responsible agent, and *sgt* to the *secondary* responsible agent.

From the results of dialogue inference, the sergeant also learns that the *std* intended sending two squads forward and did not intend sending one squad forward. Since the consequence set of *sending one squad forward* is a subset of that of *sending two squads forward*, by applying an inference rule, the sergeant believes that *std* intended *unit-fractured*. Since the student *foresaw* and *intended* the outcome, the student is to *blame* for *unit-fractured* with *high* intensity.

6. Conclusion

Social judgments are a fundamental aspect of social intelligence that involve evaluation of causality and responsibility, and facilitates social planning, social learning, natural language pragmatics and computational model of emotion. With the advance of multi-agent systems, interactive environments and the modeling of human-like agents, it is increasingly important to model and reason about this uniquely human-centric form of social inference. Based on the psychological attribution theory, this paper presents a computational approach to the problem. Our work relies on commonsense heuristics of human inference from conversation communication and causal representation of agents. Our treatments are domain-independent and thus can be used as a general approach to the problem.

Acknowledgements

This paper was developed with funds of the Department of the Army under contract number DAAD 19-99-D-0046. Thank David Traum, Jerry Hobbs, Stacy Marsella and Andrew Gordon for the valuable discussions and comments. Any opinions, findings and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

References

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

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

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

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

E. Charniak and R. Goldman. A Bayesian Model of Plan Recognition. *Artificial Intelligence*, 64(1):53-79, 1993.

H. Chockler and J. Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003.

P. N. Cohen, J. L. Morgan and M. E. Pollack (Eds). *Intentions in Communication*. The MIT Press, 1990.

J. Gratch and S. Marsella. A Domain-Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research,* in press, 2004.

J. Gratch, J. Rickel, E. Andre, N. Badler, J. Cassell and E. Petajan. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, 17(4):54-63, 2002.

H. P. Grice. Logic and Conversation. In: P. Cole and J. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*, pp.41-58. Academic Press, 1975.

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

P. Haddawy and M. Suwandi. Decision-Theoretic Refinement Planning Using Inheritance Abstraction. *Proceedings of the Second International Conference on Artificial Intelligence Planning*, 1994.

H. L. A. Hart. *Punishment and Responsibility*. Oxford University Press, 1968.

H. A. Kautz. A Formal Theory of Plan Recognition and its Implementation. In: J. F. Allen, H. A. Kautz, R. N. Pelavin and J. D. Tenenberg (Eds.), *Reasoning about Plans*, pages 69-125. Morgan Kaufmann Publishers, 1991.

B. F. Malle. Folk Explanations of Intentional Action. In: B. F. Malle, L. J. Moses and D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, pp. 265-286. The MIT Press, 2001.

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

W. Mao and J. Gratch. Decision-Theoretic Approach to Plan Recognition. *ICT Technical Report* (http://www.ict.usc.edu/publications/ICT-TR-01-2004. pdf), 2004.

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

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

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum and W. Swartout. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems*, 17(4):32-38, 2002.

C. F. Schmidt, N. S. Sridharan and J. L. Goodson. The Plan Recognition Problem: An Intersection of Psychology and AI. *Artificial Intelligence*, 11(1-2):45-83, 1978.

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

D. Traum & J. F. Allen. Discourse Obligations in Dialogue Processing. Paper presented at *the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

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

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

B. Weiner. 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.

B. Williams. *Making Sense of Humanity and Other Philosophical Papers*. Cambridge University Press, 1995.