Appears in the 4th International Working Conference on Intelligent Virtual Agents, Kloster Irsee , Germany 2003

The Social Credit Assignment Problem

Wenji Mao Jonathan Gratch

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

Abstract. *Social credit assignment* is a process of social judgment whereby one singles out individuals to blame or credit for multi-agent activities. Such judgments are a key aspect of social intelligence and underlie social planning, social learning, natural language pragmatics and computational models of emotion. Based on psychological *attribution theory*, this paper presents a preliminary computational approach to forming such judgments based on an agent's causal knowledge and conversation interactions.

1 Introduction

After the Northridge earthquake led to 58 deaths in Los Angeles, people questioned who was *answerable*: building regulators or insufficient preparedness on the part of city officials? It seems odd or unsatisfying to blame the earthquake itself, since this is largely outside human control. In contrast to how causality is used in the physical sciences, people instinctively seek out a human actor for their everyday judgments of credit or blame. Such attributions are fundamental *social* explanations involving judgments not only of causality but individual responsibility, free will and mitigating circumstances [Shaver, 1985]. These explanations underlie how we act on and make sense of the social world: 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.

With the advance of multi-agent systems, user interfaces, and human-like agents, it is increasingly important to reason about this uniquely human-centric form of social inference. This paper lays out a preliminary model of social credit assignment based on psychological attribution theory. We see several immediate applications of this model. It can inform social explanations by augmenting traditional explanations with attributions of social judgment (e.g., explaining to a student which actors deserve credit in a multi-agent training simulation. It can inform social planning by augmenting traditional causal planners with the ability to reason about which actors have authority to effect change. It can inform social learning by distinguishing praiseworthy behavior from blameworthy one and reinforcing the praiseworthy. It can inform theories of natural language as much of human conversation centers around strategies for taking credit or deflecting blame. Finally, it is key for understanding human emotion, as social emotions such as pride, anger and guilt turn on the assessment of credit or blameworthiness [Gratch, 2000].

To be concrete, consider an example from a leadership trainer we are developing [Rickel *et al*., 2002]. The trainee is in command of an infantry platoon, *eagle 2-6*, in peacekeeping operations near the Bosnian city of Celic. His mission is to reinforce another unit, *eagle 1-6*. In route, one of his vehicles seriously injures a civilian and he must balance whether to continue the mission or render aid. Many decisions and outcomes are possible. In our example, he splits his forces, ordering his sergeant to send half of his squads to aid eagle 1-6. His sergeant responds that this is a bad idea; it will allocate too few forces to either goal, and instead, one squad should be sent ahead to scout the route. The trainee overrules this recommendation. In the end, the trainee finds he has insufficient resources to render aid in a timely manner. The central question addressed here is to assess who, if anyone deserves blame for this unfortunate outcome, to what extent to blame the responsible party, and how to avoid naïve attributions, such as blaming the squad leaders that actually implement the orders.

People differ in whom they praise or blame, but psychologists and philosophers agree on the broad features underlying such judgments. Did someone *cause* the outcome? Did he *intend* the act? Did he *know* the consequence? Did he have *choice* or was he *coerced* by another agent? In the example, we may infer from the conversation that the trainee coerced the sergeant to follow an undesirable choice. We can further surmise that the trainee was forewarned of the consequences. Baring unknown mitigating factors (e.g., the sergeant always gives bad advice), we would conclude that the trainee is to blame for the delay. This example shows that proper assignment of credit or blame in a social setting must not only consider the actions (both physical acts and speech acts) and knowledge states of different actors, but also need to utilize information available to reason about key attributions that contribute to the judgment process.

2 Attribution Theory for Social Judgment

The assignment of social credit or blame has been studied extensively in philosophy, law, and social psychology. As our primary goal is to inform the design of realistic virtual humans that mimic human communicative and social behavior [Gratch *et al.*, 2002], our work differs from prior computational work in emphasizing the social nature of such judgments and by focusing on descriptive rather than proscriptive models (i.e., what people do rather than what they should do). In contrast, much of the work on AI has focused on trying to identify "ideal" principles of responsibility (e.g., the legal code or philosophical principles) and ideal mechanisms to reason about these, typically contradictory principles (e.g., non-monotonic or case-based reasoning) [McCarty, 1997].

Our work is based on *attribution theory*, specifically the work of Weiner [1995] and Shaver [1985], as it is readily adapted to AI methods. In these models, the assignment of credit or blame is a multi-step process initiated by events with positive or negative consequences. First one assesses *causality*, distinguishing between personal versus impersonal causality (i.e., is causal agent a person or a force of nature). If personal, the judgment proceeds by assessing key factors: was it the actor's *intention* to produce the outcome; did the actor *foresee* its occurrence; was the actor forced under *coercion* (e.g., was the actor acting under orders)? As the last step of the process, proper degree of credit or blame is assigned to the responsible agent (note that these theories differ in terminology; here we adopt the terminology of Shaver). Causality and intention map to standard concepts in agent-based systems, particularly frameworks that explicitly represent beliefs, desires and intentions [Bratman, 1987; Grosz and Kraus, 1996]. Coercion requires representation of social relationships and understanding of the extent to which it limits one's range of options. For example, one may be ordered to carry out a task but to satisfy the order, there may be alternatives that vary in blame or creditworthiness.

In modeling realistic human behavior, we cannot assume that a perceiving agent has privileged access to the mental states of other agents (e.g., intention is private to an agent), so deriving attribution variables can be nontrivial. In human social interactions, such variables are gleaned from a variety of sources: from observation of behavior, from statements made through natural language, from knowledge and models built up through past interactions, stereotypes and cultural norms. We show how to infer such information by analyzing natural language and causal evidence, making use of agents' knowledge of actions and consequences as well as commonsense intuition.

3 From Theory to Computational Approach

To inform social judgments, we need to represent knowledge states of agents and core conceptual variables underlying attribution theory. We also need to discuss how the representational primitives are applied in the attribution process.

3.1 Representation

An action consists of a set of preconditions, effects and steps. An action can be primitive (i.e., directly executed by an agent) or abstract. An abstract action may be decomposed hierarchically in multiple ways and each decomposition consists of a sequence of primitive or abstract sub-actions. The desirability of action effects (i.e., its positive/negative significance to an agent) is represented by utility values [Blythe, 1999].

A *non-decision node* is an abstract action that can only 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 the *choices* of the action node, and the choices are *alternatives* each other. Clearly, a primitive action is a non-decision node, while an abstract action can be either a non-decision node or a decision node.

Consequences or outcomes (we use the terms as exchangeable) are represented as a set of primitive action effects. The *consequence set* of an action in a plan hierarchy is determined by its descendents as follows: Consequences of primitive actions are those effects with non-zero utility. For non-decision nodes, the consequence set is the aggregation of the consequences of its descendents. For decision nodes, we differentiate two kinds of consequences. If a consequence of a decision node occurs among all its choices, we call the consequence a *common consequence* of the decision node; otherwise the consequence is a *non-common consequence* of the node. The consequence set of a decision node is defined as the set of its common consequences.

In addition, each action step is associated with a *performer* (i.e., the agent that performs 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

Causality: refers to the connection between actions and the effects they produce. In our approach, causal knowledge is encoded via the hierarchical task representation. Interdependencies between actions are represented as a set of causal links and threat relations. Each causal link specifies that an effect of an action achieves a particular goal that is a precondition of another action. Threat relations specify that an effect of an action threatens a causal link by making the goal unachievable before it is needed.

*Forseeability***:** refers to an agent's foreknowledge about actions and consequences. Currently we use *know* and *bring-about* to express foreseeability. If an agent knows an action brings about a consequence before its execution, then the agent foresees that the action brings about the consequence.

*Intention***:** is generally conceived as a commitment to work toward a certain act or outcome. Intending an act (act intention) is distinguished from intending an outcome of an act (outcome intention). Most theories argue that outcome intention rather than act intention is key in determining accountability and intended outcome usually deserves more elevated accountability judgments. Follow [Grosz and Kraus, 1996], we use *intend-to* and *intend-that* to denote act intention and outcome intention.

*Coercion***:** As in intentions, an agent may be coerced to act (act coercion) yet not coerced to achieve any outcome of the action (outcome coercion). We use *coerced-to* and *coerced-that* to denote act coercion and outcome coercion. In the case of outcome coercion, the responsible agent for a specific outcome is the performer or the authority of an action, but the action may not be the primitive one that directly leads to the outcome.

3.3 The Attribution Process

Social credit assignment focuses on consequences with personal significance to an agent. This evaluation is always from the perspective of a perceiving agent (e.g., an actor, an authority, a bystander, etc). As different perceivers have different preferences, different observations, and different knowledge and beliefs, different perceivers canform different judgments of the same situation. For example, an agent may think itself is not blameworthy, but a perceiver thinks the agent is. Nevertheless, the attribution process is general, and applied uniformly to different perceivers. If an action performed by an agent brings about *positive/negative* consequence, and the agent is not coerced to achieve the consequence, then credit/blame is assigned to the *performer* of the action. Otherwise, assign credit/blame to the *authority*. If the authority is also coerced, the process needs to be traced further to find the *responsible* agent for the consequence.

Following Weiner, we use coercion to determine the responsible agent, and intention and foreseeability in assigning the intensity of credit/blame. We adopt a simple categorical model of intensity assignment, though one could readily extend the model to a numeric value by incorporating probabilistic rules of inference. If the responsible agent intends the consequence while acting, the intensity assigned is *high*. If the responsible agent does not foresee the consequence, the intensity is *low*.

4 Inference from Communication and Plans

Judgments of causality, foreseeability, intentionality and coercion are informed by dialogue and causal evidence. Many theories have formally addressed subsets of this judgment task. For example, [Sadek, 1990] addresses the relationship between dialogue and inferences of belief and intention. These theories have not tended to consider coercion. Rather than trying to synthesize and extend such theories, we introduce small number of commonsense rules that, via a justification-based truth maintenance system (JTMS), allow agents to make inferences based on this evidence.

4.1 Inferring from Conversational Dialogue

Dialogue between agents is a rich source of information for the derivation of attribution variables. In a conversational dialogue, a speaker (S) and a hearer (H) take turns alternatively. When a speech act is performed, a perceiving agent observes the conversation and makes inferences based on its beliefs. As the conversation proceeds, the perceiver forms new beliefs and updates inferences accordingly.

Assume conversations between agents are *grounded*. According to Grice's maxims [Grice, 1975], we assume agents communicate *sincerely* and *relevantly*. Background information (agents' social roles, relationship, etc) is also important.

x and *y* are different agents. *A* and *B* are actions. *p* is a proposition and *t* is time. We focus on the following speech acts:

inform(*x, y, p, t*): *x* informs y that *p* at *t*.
order(*x, y, A, t*): *x* orders *y* to act *A* at *t*.
request(*x, y, A, t*): *x* requests *y* to act *A* at *t*.
accept(*x, A, t*): *x* accepts to act *A* at *t*.
reject(*x, A, t*): *x* rejects to act *A* at *t*.
counterpropose(*x,, A, B, y, t*): *x* counters *A* and proposes *B* to *y* at *t*.

Let *z* be a perceiving agent. If at time *t1*, S *inform*s H that *p*, then after *t1*, *z* can infer that S knows *p* as long as there is no intervening contradictory belief.

An *order* (or a *request*) gives evidence of S's *desire* (or *want*) to let H act. An order also gives evidence that S intends to act.

H may *accept, reject* or *counterpropose* an order/request. If S wants H to act and H *accept*s, it can be inferred that H intends to act. An agent can accept via speech or by action execution. If H accepts an act wanted by a superior, there is evidence of coercion.

$$\text{believe}(z, \text{want}(x, \text{do}(y, A)), t1) \wedge \text{accept}(y, A, t2) \wedge t1<t2<t4 \wedge \neg(\exists t3)(t2<t3<t4 \wedge \text{believe}(z, \neg\text{intend-to}(y, A), t3)) \Rightarrow \text{believe}(z, \text{intend-to}(y, A), t4)$$

$$\text{believe}(z, \text{want}(x, \text{do}(y, A)), t1) \wedge \text{accept}(y, A, t2) \wedge \text{superior}(x, y) \wedge t1<t2<t4 \wedge \neg(\exists t3)(t2<t3<t4 \wedge \text{believe}(z, \neg\text{coerced-to}(y, A, x), t3)) \Rightarrow \text{believe}(z, \text{coerced-to}(y, A, x), t4)$$

If S wants H to act and H *reject*s, infer that H does not intend to act. If H *counterpropose*s act *B* instead of *A*, both S and H are believed to know that *A* and *B* are alternatives. If S has known that two actions are alternatives and orders one of them, infer that S intends to the chosen one instead of the alternative.

As these rules are general, they can be combined flexibly and applied to variable dialogue sequences of multiple participants. For the complete version of inference rules, the reader may refer to [Mao and Gratch, 2003].

4.2 Inferring from Plans

Conversation communication provides information about agents' intentions and choices in acting, that is, *intend-to* and *coerced-to*. To derive *intend-that* and *coerced-that* for judgments, we need to solve the problem of inferring outcome intention and outcome coercion from act intention and act coercion.

Different agent will have different plans and preferences, though the structure of plans can be described using general terms such as action types, effects and alternatives. We adopt a domain-independent hierarchical task formulism that differentiates action types, explicitly represents consequences of alternatives, and separate common consequences of an action from its non-common ones.

Inferring Outcome Intention from Act Intention

If an agent intends a voluntary act (i.e., not coerced to do so), the agent must intend to achieve (at least) one consequence of the action. If a voluntary, intended action has only one consequence, then the agent is believed to intend the consequence. In more general cases, when an action has multiple consequences, in order to identify whether a specific outcome is intended or not, a perceiver may examine alternatives the agent intends and does not intend, and compare the consequences of intended and unintended alternatives.

If an agent intends an action *A* voluntarily and does not intend alternative *B*, we can infer that the agent either intends (at least) one consequence unique to *A* or does not intend (at least) one consequence unique to *B*, or both. If the consequence set of *A* is subset of that of *B*, the rule can be simplified. As there is no consequence of *A* not occurring in the consequence set of *B*, we can infer that the agent does not intend (at least) one consequence unique to option *B*. In particular, if there is only one consequence *p* of *B* that does not occur in the consequence set of *A*, infer that the agent does not intend *p*. On the other hand, given the same context that an agent intends an action *A* and does not intend its alternative *B*, if consequence set of *B* is subset of that of *A*, infer that the agent intends (at least) one consequence that only occurs in *A*. If there is only one consequence *p* of *A* that does not occur in the consequence set of *B*, the agent must intend *p*.

Inferring Outcome Coercion from Act Coercion

In a non-decision node, if an agent is coerced to act, the agent is also coerced to achieve the consequence of subsequent actions, for the agent has no other choice.

In a decision node, however, an agent must decide amongst multiple alternatives. Even if an agent is coerced to act, it does not follow that the agent is coerced to achieve a specific consequence of these alternatives. To infer *coerced-that* from *coerced-to* in a decision node, we examine the choices at a decision node. If an outcome is a common consequence of every alternative, then it is unavoidable: *coerced-that* is true. Otherwise, the agent has the option to choose an alternative that avoids the consequence: *coerced-that* is false. Our definition of consequence set ensures the consistency when these rules are applied to the actions at different levels of plan structure.

Back-Tracing Algorithm

We have developed a back-tracing algorithm for evaluating the responsible agent for a specific consequence [Mao and Gratch, 2003]. The algorithm starts with the primitive action that directly causes the specific consequence and works up the plan hierarchy. In each pass of the main loop, the algorithm applies inference rules to infer attribution variables. If there is evidence that the performer is coerced to act, the algorithm applies inference rules to assess outcome coercion. If there is outcome coercion, the authority is deemed responsible. If current action is not the root node in plan structure and outcome coercion is true, the algorithm proceeds up the plan hierarchy.

After the execution of the algorithm, the responsible agent for the outcome is determined. Meanwhile, the algorithm may also acquire values for act intention and foreknowledge. The rules for inferring outcome intention then can be applied to determine the responsible agent's intention in achieving the evaluated consequence.

Events may lead to more than one desirable/undesirable consequence. For multiple consequences, we can apply the algorithm the same way, evaluating one consequence each time during its execution. Then, to form an overall judgment, the results can be aggregated and grouped by responsible agents.

5 Illustrative Example

We are developing this work in the context of the Mission Rehearsal Exercise (MRE) leadership trainer [Rickel *et al*., 2002]. In that system, there are three social actors, the *student* (*std*), the *sergeant* (*sgt*) and the *squad leader* (*sld*), who work as a team in task performance. *std* acts as an authority over *sgt* and *sld* as a subordinate of *sgt*. Conversations between agents are represented via speech acts and a dialogue history.

Take *sgt*'s perspective as an example, we illustrate part of the task structure and evaluate one of the negative consequences. In the plan, *one squad forward* (*one-sqd-fwd*) and *two squads forward* (*two-sqds-fwd*) are two choices of abstract action *support eagle 1-6* (*support-1-6*). *One-sqd-fwd* is composed of primitive actions *4th squad recon forward* and *remaining squads forward* (*remaining-fwd*). *Two-sqds-fwd* consists of primitive actions *1st and 4th squads to celic* (*1st-and-4th-to-celic*) and *2nd and 3rd squads to celic*. Two effects, (*eagle*) *1-6 supported* and *unit fractured*, are salient to *sgt*. Both *remaining--fwd* and *1st-and-4th-to-celic* have the effect *1-6-supported*, which is a desirable team goal. Besides, *1st-and-4th-to-celic* has the side effect *unit-fractured*. Assume *unit-fractured* is undesirable to *sgt* and this negative consequence serves as input to the back-tracing algorithm. We illustrate how to find the blameworthy agent.

The algorithm starts from *1st-and-4th-to-celic*, of which *unit-fractured* is an effect. The action was executed by *sld*. Dialogue history shows that *sgt* ordered *sld* to act. Initially, *coerced-to*(*sld, 1st-and-4th-to-celic, sgt*) and *coerced-that*(*sld, unit-fractured, sgt*) are unknown. By default, the algorithm assigns *sld* to the responsible agent.

Since *sgt* ordered *sld*, apply an inference rule. The algorithm infers that *sgt* believes he wants *sld* to act. Since *sld* accepted by action execution, and *sgt* is the superior, apply another inference rule. The algorithm infers that *sgt* believes he coerced *sld* to act. As *coerced-to* is true and the primitive action is a *non-decision* node, infer that *sgt* believes he coerced *sld* to fracture the unit. Now that *coerced-that* is true, assign *sgt* to the responsible agent. So *sgt* believes he is responsible for *unit-fractured*. Since parent node is not the *root* of the plan and *coerced-that* is true, the algorithm enters the next loop.

The action is *two-sqds-fwd*, performed by *sgt*. A variety of beliefs can be inferred from commonsense rules:

believe(*sgt*, want(*std*, do(*sgt*, *two-sqds-fwd*)))
believe(*sgt*, know(*std*, alternative(*one-sqd-fwd*, *two-sqds-fwd*)))
believe(*sgt*, intend-to(*std*, *two-sqds-fwd*))
believe(*sgt*, ¬intend-to(*std*, *one-sqd-fwd*))
believe(*sgt*, coerced-to(*sgt*, *two-sqds-fwd*, *std*))
believe(*sgt*, coerced-that(*sgt*, *unit-fractured*, *std*))

As *coerced-that* is true, the algorithm assigns *std* to the responsible agent.

The action is *support-1-6*, performed by *std*. There is no clear evidence of coercion. The algorithm terminates. Since the consequence set of *one-sqd-fwd* is a subset of that of *two-sqds-fwd*, the algorithm infers that *sgt* believes *std* intended *unit-fractured*. So the student is to blame for *unit-fractured* with high intensity.

6 Summary and Future Work

Based on psychological attribution theory, this paper presents a preliminary computational approach to social credit assignment. The problem is central in social psychology and social cognition. With the development of human-like agent systems, it is increasingly important for computer-based systems to model this human-centric form of social inference. Our work attempts to help bridge between psychological accounts and computational models by means of AI methods. Rather than impose arbitrary rules on judgment process, 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.

This work is still in its early stages. The current implementation has focused on simple commonsense rules in contrast to the more rigorous, often non-monotonic theories typically explored in models of beliefs and intentions. Our sense is these rules are sufficient for our practical applications, more efficient, though less general than these more formal methods. Our future work must explore more deeply the relationship between these approaches. The model must also be extended before it can be fully integrated in our existing applications. We must incorporate probabilistic reasoning to deal with uncertainty in observations and judgment process. For modeling more complex multi-agent teamwork, we need to consider joint responsibility and sharing responsibility among teammates (the current model assumes one agent has sole responsibility). Some inference rules are too restrictive and need to make better use of plan knowledge, particularly considering how preconditions and effects indirectly limit one's choices in acting. As our task representation has already encoded information about action preconditions and effects, this should be a natural extension of our existing methods.

Acknowledgement

This paper was developed with funds of the U.S. Department of the Army under contract number DAAD 19-99-D-0046. We thank our colleagues on the MRE project, in particular David Traum and Andrew Gordon, and the anonymous reviewers for their valuable comments. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the United States Department of the Army.

References

1. J. Blythe. Decision-Theoretic Planning. *AI Magazine*, 20(2):37-54, 1999.
2. M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, 1987.
3. J. Gratch. Emile: Marshall Passions in Training and Education. In: *Proceedings of the 4th International conference on Autonomous Agents*, Barcelona, 2000.
4. J. Gratch, J. Rickel, E. André, N. Badler, J. Cassell and E. Petajan. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, 17(4), pp. 54-63, 2002.
5. H. P. Grice. Logic and Conversation. In: P. Cole and J. Morgan (Eds.), *Syntax and Semantics: Vol 3 Speech Acts*. Academic Press, 1975.
6. B. Grosz and S. Kraus. Collaborative Plans for Complex Group Action. *Artificial Intelligence*, 86(2): 269-357, 1996.
7. M. Mao & J. Gratch. "The Social Credit Assignment Problem (Extended Version), *ICT Technical Report*, ICT-TR-02-2003, 2003.
8. L. McCarty. Some Arguments about Legal Arguments. In: *Proceedings of 6th International Conference on Artificial Intelligence and Law*, Melbourne, 1997.
9. J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum and B. Swartout. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems*, 17(4), pp.32-38, 2002.
10. Sadek, M.D., "Logical task modeliling for man-machine dialogue," in *Proceedings of AAAI*, 1990.
11. K. G. Shaver. *The Attribution Theory of Blame: Causality, Responsibility and Blameworthiness*. Springer-Verlag, 1985.
12. B. Weiner. *The Judgment of Responsibility*. Guilford Press, 1995.