Plenty of Blame to Go Around:
A Qualitative Approach to Attribution of Moral Responsibility

Emmett Tomai **Ken Forbus**

Qualitative Reasoning Group
Northwestern University
2133 Sheridan Rd, Evanston, IL 60208
etomai@northwestern.edu

Abstract

We present a computational model of blame attribution. Recently Mao and Gratch, following Attribution theory, created a computational model that assigned blame to an agent for a negative occurrence. Their model made categorical judgments, and could only assign blame to a single agent. Our model extends this work, using QP theory to provide a continuous model for the parameters involved in attribution and directly capturing the constraints postulated by Attribution theory. This allows our model to infer relative amounts of blame in a situation in a manner that is consistently overall with relative amounts of blame attributed in a psychological experiment.

Who is to Blame?

Bad things happen, and an important capability of social agents is to understand who is responsible. From the affairs of nations to personal misfortunes, accountability is an important part of how we understand the world around us. But how does a person go from perceiving such a situation to making a judgment of blame? This question has been the topic of much research in social psychology. Driven by the need to create social agents that can interact with people for a variety of purposes (tutoring, entertainment, assistants), creating computational models to capture such judgments is receiving increased attention. Without an understanding of blame assignment, an agent cannot properly infer the implications of social interactions.

This paper describes how Qualitative Process theory [Forbus 1984] can be used in such modeling. We briefly summarize the elements of *Attribution theory* that address blame judgments. We then discuss the Mao and Gratch computational model [Mao & Gratch 2005][Mao 2006]. We present an alternative model for attribution of blame based on QP theory, which we claim better represents the underlying theory as well as human data. We present an evaluation of our model using data collected by Mao, showing that our model captures that data better, and makes additional predictions.

Attribution Theory

Attribution theory is an area of research in Social Psychology based on the founding work of Heider [Heider 1958] and advanced by Kelley [Kelley 1973] and Jones [Jones & Davis 1965]. Its goal is to identify the conditions that will lead a perceiver, through an *attribution process*, to attribute some behavior, event or outcome to an internal disposition of the agent involved, as opposed to an environmental condition. Attribution is, therefore, a judgment embedded in the point of view of the perceiver and subject to the epistemic state of that perceiver.

Further work in Attribution theory has directly addressed the question of the attribution of blame [Shaver 1985, Weiner 1995]. Our model is based primarily on the work of Shaver who makes the distinction between *cause*, *responsibility* and *blameworthiness*. For a given negative outcome, cause is defined as being an insufficient but necessary part of a condition which is itself unnecessary but sufficient for that result. Only causes which represent human agency are of interest to the theory. It is certainly possible that people attribute responsibility and even blame to animals or the inanimate, but in doing so they would have to give that target additional human qualities, to the extent that it would be no different than pretending it was a human agent from the point of view of the theory. Responsibility is a broad term with several senses; the one of interest in this process is referred to by Shaver as "moral accountability", distinct from legal responsibility, the responsibilities of a formal office or mental/emotional capacity (e.g. "He was not responsible for his actions"). Blame is a moral condemnation that follows from responsibility for a morally reprehensible outcome but may be mitigated by *justification* or *excuse*.

Shaver's attribution process begins with an outcome that has been judged negative and evaluates an involved agent for attribution of responsibility against five *dimensions of responsibility*, which are: *causality*, *intentionality*, *coercion*, *appreciation*, and *foreknowledge*. Shaver's process is sequential in its evaluation. We discuss the role of each dimension in the attribution process in turn.

Schultz and Schleifer [Schultz & Schleifer 1983] argue that a judgment of responsibility presupposes a judgment of cause. While Shaver points out that responsibility judgment may be driven purely by a desire for answerability, without a causal connection, his model of attribution adheres to the principle that there must be some judgment of cause for any responsibility to be attributed.

Shaver describes intention as a scale of deliberateness with intentional at one end and involuntary at the other. He describes it as the central concern in attribution of responsibility, and claims that a judgment of intention should result in the strongest judgment of responsibility. There are, however, exceptions to be found in the judgments of coercion and appreciation.

Coercion captures the force exerted by another agent which limits the available choices, from a social standpoint, for the agent in question. This could be through some direct threat or via an authority relationship. In Shaver's model coercion comes into play only once intentionality has been established – it therefore covers only that influence which leads to intentional obedience. An agent who is coerced is assigned less responsibility than one who acts intentionally in the absence of coercion.

In the appreciation dimension the perceiver judges whether the agent in question has the capacity to understand that the outcome in question is morally wrong. If the agent does not have such capacity, then they still bear some responsibility, but are held exempt from blame.

Foreknowledge is defined as the extent to which the agent was aware that a particular action would result in a particular outcome, prior to execution. As with all aspects of this process, it is the perceiver's judgment of the knowledge the agent possessed that is evaluated. Foreknowledge may also be the perceiver's judgment of what knowledge the agent should have had. In the absence of intention, foreknowledge becomes the driving question for responsibility. Shaver attributes less responsibility to an agent who should have known than to one that did know, and less to either than to an agent in the intentional cases above.

Once an agent has been judged responsible, blame follows unless there is a successful intervention by justification or excuse. Justification does not deny responsibility but presents a reason why responsibility for a negative outcome should not carry the negative attribution of blame. This would be the case where someone shot someone else dead, but did it in self-defense. Excuses deny responsibility by appealing the judgments of the dimensions. Examples are "I didn't do it", "I didn't know", "I didn't mean it", "He made me do it" and "She doesn't know it's wrong". Successful intervention by an excuse alters the assignment of responsibility.

Mao and Gratch Computational Model

Mao [Mao 2006], in collaboration with Gratch [Mao & Gratch, 2005] developed a computational model of responsibility assignment which models the judgments of *attribution variables* based on the dimensions of causality, intentionality, coercion and foreknowledge, and the attribution of blame[1] following from those judgments. It does not deal with justifications and excuses, thus blame follows directly from responsibility.

In Mao's model, actions are encoded via hierarchical plans. Non-primitive actions can have multiple decompositions, representing alternative ways the action can be achieved. Actions have propositional preconditions and effects, as well as slots to indicate the agent that could or did perform the action and the agent under whose authority the action falls. Communicative events are modeled as a sequence of *speech acts* [Austin 1962; Searle 1969] representing informing, requesting, and negotiations.

Mao describes a set of inference rules that takes these representations and assigns values to attribution variables that capture how each involved agent is judged with respect to each negative outcome. Causality is ascertained by performance of the primitive action that resulted in the outcome. For intention, a significant distinction is made between act and outcome intention, following from [Weiner 2001]. It is assumed that an agent intends any action that he or she performs or orders performed. However, act intention implies intention of at least one outcome, not all of the outcomes. When an action has multiple possible decompositions and the performing agent was allowed to choose the decomposition, outcome intention moves from the ordering agent to the performing agent if not all decompositions led to the negative outcome. Coercion is inferred from order negotiation; one agent ordering another to perform an action shows act coercion and possibly outcome coercion, depending on how much choice the performing agent had in carrying out their orders. The rules for determining outcome coercion follow the same logic as for determining outcome intention, with the additional constraint that an agent with prior intention is not coerced by being ordered to do what they already intended. Foreknowledge is strongly implied by communication of knowledge of the outcome to another agent before the action is executed. It is also implied by intention, as one cannot intend what one is unaware of.

Mao's work is an important step towards modeling blame attribution. However, there are three limitations we address here. First, as [Mao 2006] observes, it uses Boolean values for attribution variables, whereas Attribution theory describes the dimensions of responsibility in terms of scalar values. Second, all blame is assigned to a single agent (or group of agents in a joint action). This is inconsistent with the human data in Mao's own experiment. Third, the degree of blame assigned by the system is limited to a value of *high* for intentional action and a value of *low* in the absence of intention. These assignments do not match up with the human data.

[1] Social psychological research cited in [Mao 2006] indicates that there are differences in the processes used for responsibility for positive events and negative events, hence the exclusive focus on negative events here.

Qualitative Model of Attribution

We claim that these limitations can be addressed by an encoding of Attribution theory using the principles of Qualitative Process (QP) theory [Forbus 1984]. We claim that this model makes more informative distinctions between blame assignments within and across scenarios.

While physical domains have been a major focus of QR research, an increasing number of researchers have found these techniques useful in fields where theories are expressed in continuous parameters more generally, including organization theory (cf. [Kamps & Peli, 1995]), economics (cf. [Steinmann, 1997]), and political reasoning [Forbus & Kuehne 2005]. Qualitative reasoning, we believe, provides an especially appropriate level of representation for reasoning about social causality. Theories typically are expressed in terms of continuous parameters, such as "amount of intention" and "degree of foreknowledge", but there tend to not be principled ways to move to quantitative models and numerical values for such parameters. In those circumstances, qualitative modeling becomes the most rigorous way to proceed, and ordinal fits with human data becomes the most robust measure.

Our model takes as input the same attribution variables generated by Mao's planning and dialogue inference systems; we tackle neither of those issues, since it is not clear that QR has much to say about them. We extend the inferences on those variables to allow qualitative rather than just Boolean values, and support assignment of responsibility to multiple recipients. Finally, we model Shaver's attribution process to judge responsibility based on those values. Because we omit justification and excuse at this time, we speak in terms of responsibility rather than blame.

In our model, judgments of responsibility, as well as the attribution variables for intentionality, coercion and foreknowledge, are represented by nonnegative continuous parameters. Judgments of causality remain Boolean as that is the extent of their impact in Shaver's model of attribution. A value of zero is a lower limit point indicating the absence of responsibility, intentionality, coercion or foreknowledge in the judgment of the perceiver. Foreknowledge is a function of time: it is the knowledge about the outcome of an action held over an interval prior to, during and after the action. We represent both foreknowledge that the perceiver believes the agent had (epistemic) and should have had (expected). Intention is also evaluated with respect to time as it may be judged to vary over time. We use Allen's interval calculus relations *contains* and *overlaps* in our inference rules [Allen 1983].

For a given scenario, where Mao's system asserts a value of true for intention, coercion or foreknowledge, we assert a value greater than zero. Where epistemic foreknowledge is inferred in Mao's system by communication of the knowledge, we assert equality to an upper limit point of certainty.

In attribution theory, intention does not refer to desire. That is to say, an agent who points a gun and pulls the trigger may or may not have wanted that person to die, but they certainly intended for their action to produce that outcome. Even with that distinction, there is much philosophical discussion on the meaning of intention. According to Shaver, a judgment of intention presupposes epistemic foreknowledge, but not the other way around [Shaver 1985]. On the other hand, Bratman argues that epistemic foreknowledge, and the degree to which it is certain, combined with action must imply intention [Bratman 1990]. Acknowledging these differences in opinion, our model makes the weaker inference that when an agent is certain of an outcome and performs or authorizes the action, it implies only some non-zero level of intention.

Attribution of responsibility from the attribution variables begins with an assessment of eligibility. The agent that performed the action that caused the outcome is eligible of course. Where an agent is in a position of authority over the action that caused the outcome, that agent is also eligible. In both cases, the agent is *responsible by action.* In the case of coercion, the coercing agent is *responsible by coercion* and is also eligible for responsibility for the outcome. Note that R2 and R3 are mutually exclusive – an authority who coerces is responsible by coercion, not by action. These rules are as follows:

```
R1: causes(?action, ?outcome) ∧
    performedBy(?action, ?agent)
⇒ responsibleByActionFor(?agent, ?action, ?outcome)

R2: causes(?action, ?outcome) ∧
    authorizedBy(?action, ?agent) ∧
    performedBy(?action, ?coerced) ∧
    CoercionFn(?agent, ?coerced, ?outcome) = 0
⇒ responsibleByActionFor(?agent, ?action, ?outcome)

R3: causes(?coercion-action,
           CoercionFn(?agent, ?coerced, ?outcome) > 0) ∧
    performedBy(?coercion-action, ?agent)
⇒ resposibleByCoercionFor(?agent, ?coercion-action,
                          ?outcome)
```

Given our omission of the more special-case dimension of appreciation, Shaver's attribution process displays four distinct modes of judgment: causal without foreknowledge, causal without intent, intentional but coerced and intentional in the absence of coercion. Responsibility is strictly increasing across those modes, in that order. Within each state, responsibility is qualitatively proportional to a different attribution variable. These modes translate into six model fragments or views in our model.

The first two modes translate directly into two views. The third and fourth modes each translate into two views based on whether the agent being considered is responsible by action or coercion. In the former case, intention and foreknowledge are measured at the time of the causal action. In the latter case, they are measured at the time of the coercing action. The six views are as follows:

```
View: CausalWithoutForeknowledge
Conditions:
 responsibleByActionFor(?agent, ?action, ?outcome) ∧
 ¬∃?s1(KnowledgeFn(?agent,
                   causes(?action, ?outcome),
                   ?s1) > 0 ∧
       contains(?s1, ?action)) ∧
 ¬∃?s2(IntentionFn(?agent, ?outcome, ?s2) > 0 ∧
       contains(?s2, ?action)) ∧
 Knowledge-ExpectedFn(?agent, causes(?action, ?outcome),
                      ?s3) > 0 ∧
 contains(?s3, ?action)
Consequences:
 ResponsibilityFn(?agent, ?outcome) ∝Q+
   Knowledge-ExpectedFn(?agent,
                        causes(?action, ?outcome),?s3)

View: CausalWithoutIntent
Conditions:
 responsibleByActionFor(?agent, ?action, ?outcome) ∧
 KnowledgeFn(?agent, causes(?action, ?outcome),
             ?s1) > 0 ∧
 contains(?s1, ?action)) ∧
 ¬∃?s2(IntentionFn(?agent, ?outcome, ?s2) > 0 ∧
       contains(?s2, ?action))
Consequences:
 ResponsibilityFn(?agent, ?outcome) ∝Q+
   KnowledgeFn(?agent, causes(?action, ?outcome), ?s1)

View: IntentionalButCoerced
Conditions:
  responsibleByActionFor(?agent, ?action, ?outcome) ∧
  IntentionFn(?agent, ?outcome, ?s1) > 0 ∧
  contains(?s1, ?action))) ∧
  CoercionFn(?coercer, ?agent, ?outcome) > 0
Consequences:
  ResponsibilityFn(?agent, ?outcome) ∝Q+
    CoercionFn(?coercer, ?agent, ?outcome)

View: IntentionalByCoercionButCoerced
Conditions:
  responsibleByCoercionFor(?agent, ?coercion-action,
                           ?outcome) ∧
  IntentionFn(?agent, ?outcome, ?s1) > 0 ∧
  contains(?s1, ?coercion-action))) ∧
  CoercionFn(?coercer, ?agent, ?outcome) > 0
Consequences:
  ResponsibilityFn(?agent, ?outcome) ∝Q+
    CoercionFn(?coercer, ?agent, ?outcome)

View: Intentional
Conditions:
  responsibleByActionFor(?agent, ?action, ?outcome) ∧
  IntentionFn(?agent, ?outcome, ?s1) > 0 ∧
  contains(?s1, ?action))) ∧
  ¬∃?coercer(CoercionFn(?coercer, ?agent, ?outcome)
             > 0)
Consequences:
  ResponsibilityFn(?agent, ?outcome) ∝Q+
    IntentionFn(?agent, ?outcome, ?s1)

View: IntentionalByCoercion
Conditions:
  responsibleByCoercionFor(?agent, ?coercion-action,
                           ?outcome) ∧
  IntentionFn(?agent, ?outcome, ?s1) > 0 ∧
  contains(?s1, ?coercion-action))) ∧
  ¬∃?coercer(CoercionFn(?coercer, ?agent, ?outcome)
             > 0)
Consequences:
  ResponsibilityFn(?agent, ?outcome) ∝Q+
    IntentionFn(?agent, ?outcome, ?s1)
```

Given a scenario with a negative outcome and some number of agents, our system first infers which agents bear some level of responsibility for the outcome. For each agent in that set, it infers what mode of judgment to use and the qualitative proportionality that constrains the amount of responsibility attributed. Given a number of such scenarios, our system is able to infer ordinal constraints on responsibility for all pairs of agents both within and across the scenarios. Clearly for situations where two responsibility judgments being considered fall into different judgment modes, the inference is straightforward. For judgments within the same mode, we can infer relative amounts of responsibility when ordinal relationships between the control parameters are known.

In Mao's inference system, evidence of intention prior to coercion determines the strength of the coercion. If the agent in question intended the action or outcome prior to being ordered to do it, then there is no coercion. If the agent did not intend it, then there is strong coercion. If the agent's prior intent is unknown, then there is weak coercion. The strong/weak distinction is not used in Mao's attribution process, but is targeted towards a probabilistic extension of the system. We modify this rule to infer ordinal constraints on the amount of coercion as follows:

```
R4: causes(?coercion-action1,
           CoercionFn(?coercer1, ?agent1, ?outcome1)
                > 0) ∧
    IntentionFn(?agent1, ?outcome1, ?s1) = 0 ∧
    overlaps(?s1, ?coercion-action1) ∧
    causes(?coercion-action2,
           CoercionFn(?coercer2, ?agent2, ?outcome2)
              > 0) ∧
    ¬∃?s2(IntentionFn(?agent2, ?outcome2, ?s2) = 0 ∧
          contains(?s2, ?coercion-action2))
⇒ CoercionFn(?coercer1, ?agent1, ?outcome1) >
    CoercionFn(?coercer2, ?agent2, ?outcome2)
```

In the dimension of causality, Shaver argues that omission is just as blameworthy as commission. In our model we extend this allowance to the dimension of coercion. As stated in rule R2, an agent who is in a position of authority over a causal action is extended eligibility for responsibility. If the authority is aware of the possibility of a negative outcome from the underling's actions, yet does not coerce his or her underling away from that outcome, then he or she is guilty of abdicating authority. Under these circumstances the authority is subject to the same evaluation of intention as the underling. However, if the authority is unaware of the underling's actual intention to cause that outcome, then his or her outcome intention is constrained to be less than the intention of the underling. These rules are as follows:

```
R5: causes(?action, ?outcome) ∧
    performedBy(?action, ?underling) ∧
    authorizedBy(?action, ?authority) ∧
    CoercionFn(?authority, ?underling, ?outcome) = 0 ∧
    KnowledgeFn(?authority, causes(?action, ?outcome),
                ?s1) > 0 ∧
    contains(?s1, ?action)
⇒ abdicatedAuthority(?authority, ?underling,
                     ?action, ?outcome)
```

```
R6: abdicatedAuthority(?authority, ?underling,
                       ?action, ?outcome) ∧
    IntentionFn(?underling, ?outcome, ?s1) > 0 ∧
    contains(?s1, ?action) ∧
    IntentionFn(?authority, ?outcome, ?s2) > 0 ∧
    contains(?s2, ?action) ∧
    ¬∃?s4(KnowledgeFn(?authority,
                      (IntentionFn(?underling, ?outcome,
                                   ?s3) > 0 ∧
                       contains(?s3, ?action))
                      ?s4) ∧
          contains(?s4, ?action))
⇒ IntentionFn(?underling, ?outcome, ?s1) >
     IntentionFn(?authority, ?outcome, ?s2)
```

Finally, the outcome intention of an agent who chooses not to coerce, even one in authority, must be considered less than that of an agent who chooses to coerce. Again as follows:

```
R7: IntentionFn(?agent1, ?outcome1, ?s1) > 0 ∧
    contains(?s1, ?action1) ∧
    CoercionFn(?agent1, ?coerced1, ?outcome1) > 0 ∧
    causes(?action2, ?outcome2) ∧
    abdicatedAuthority(?agent2, ?underling,
                       ?action2, ?outcome2) ∧
    IntentionFn(?agent2, ?outcome2, ?s1) > 0 ∧
    contains(?s2, ?action2)
⇒ IntentionFn(?agent1, ?outcome1, ?s1) >
     IntentionFn(?agent2, ?outcome2, ?s2)
```

Evaluation

Mao presents an evaluation of her system against human data collected in a survey of 30 respondents. The survey presented four scenarios, variations starting with the "*company program"* scenario used by Knobe [Knobe 2003], replicated below. The scenarios involve two agents, a chairman and a vice president, and a negative outcome of environmental harm. Each scenario was followed by a set of Yes/No questions intended to validate the judgments of intermediate variables, including the attribution variables, and a final question asking the respondent to score the blame each agent deserved on a scale of 1-6. Due to space limitations, we refer the reader to [Mao 2006] for details on the data collection process.

Corporate Program Scenarios

Scenario 1. The vice president of Beta Corporation goes to the chairman of the board and requests, "Can we start a new program?" The vice president continues, "The new program will help us increase profits, and according to our investigation report, it has no harm to the environment." The chairman answers, "Very well." The vice president executes the new program. However, the environment is harmed by the new program.

Scenario 2. The chairman of Beta Corporation is discussing a new program with the vice president of the corporation. The vice president says, "The new program will help us increase profits, but according to our investigation report, it will also harm the environment." The chairman answers, "I only want to make as much profit as I can. Start the new program!" The vice president says, "Ok," and executes the new program. The environment is harmed by the new program.

Scenario 3. The chairman of Beta Corporation is discussing a new program with the vice president of the corporation. The vice president says, "The new program will help us increase profits, but according to our investigation report, it will also harm the environment. Instead, we should run an alternative program, that will gain us fewer profits than this new program, but it has no harm to the environment." The chairman answers, "I only want to make as much profit as I can. Start the new program!" The vice president says, "Ok," and executes the new program. The environment is harmed by the new program.

Scenario 4. The chairman of Beta Corporation is discussing a new program with the vice president of the corporation. The vice president says, "There are two ways to run this new program, a simple way and a complex way. Both will equally help us increase profits, but according to our investigation report, the simple way will also harm the environment." The chairman answers, "I only want to make as much profit as I can. Start the new program either way!" The vice president says, "Ok," and chooses the simple way to execute the new program. The environment is harmed.

	Human Data		**Mao Model**		
	Chair	VP	Chair	VP	Degree
Scenario1	3.00	3.73		Y	Low
Scenario2	5.63	3.77	Y		Low
Scenario3	5.63	3.23	Y		Low
Scenario4	4.13	5.20		Y	High

Table 1. Blame attribution results

Table 1 shows for each scenario the average blame attributed to each agent by the survey respondents, the single choice of the blameworthy agent made by Mao's system and the degree of responsibility for that agent asserted by Mao's system. In each scenario, Mao's model correctly selects the agent who receives the higher degree of blame, but with the incorrect implication that the other agent involved is free of responsibility. The degree of responsibility assertions made by Mao's model do not match the human data.

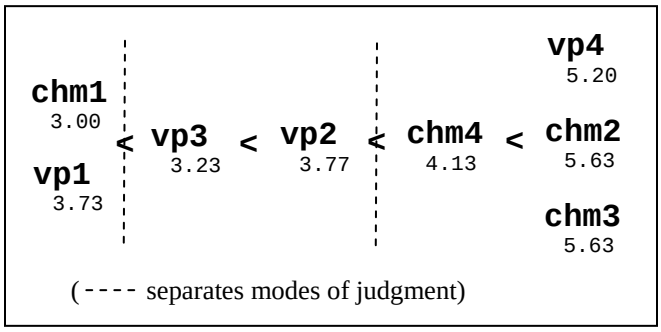

Figure 1. Ordinal constraints on responsibility and average participant attribution numbers

Figure 1 shows the ordinal constraints inferred by our model on the amount of responsibility for the agents in the four scenarios, together with labels indicating the average blame attributed to each by the survey respondents.

The eight agents being considered fall into three of the four modes of judgment from Shaver's attribution theory. The ordering of those modes establishes the ordinal relations between all pairs of agents in different modes. The chairman and vice president in scenario 1 both fall into the `CausalWithoutForeknowledge` view which is part of the causal without foreknowledge mode of judgment. Within this view, the responsibility of each agent is qualitatively proportional to only the amount of expected foreknowledge each agent is judged to have. As there are no inferred constraints on their expected foreknowledge, they remain unordered. The vice president in scenario 2 and the vice president in scenario 3 fall into the `IntentionalButCoerced` view which is part of the intentional but coerced mode of judgment. Their respective degree of responsibility is qualitatively proportional to the amount of coercion judged to have been applied. There is no indication of the outcome intention of the vice president in scenario 2 prior to the coercion action, while the vice president in scenario 3 clearly shows lack of outcome intention prior to being coerced. The vice president in scenario 3 is therefore judged to have a higher degree of coercion by rule R4 and thus a lower degree of responsibility. The chairman and vice president in scenario 4 fall into the `Intentional` view while the chairmen from scenarios 2 and 3 fall into the `IntentionalByCoercion` view, both of which are part of the intentional mode of judgment. Their degrees of responsibility are qualitatively proportional to their outcome intention. The chairman in scenario 4 abdicated his authority to the vice president as captured by rule R5. However, because he did not coerce the outcome nor did he have prior knowledge of the vice president's intention, he is constrained to have a lower degree of intention than the other three by rules R6 and R7. This results in a lower judgment of responsibility while the other three remain unordered.

In 23 of the 28 possible comparisons between agents our system correctly infers which agent should receive more blame. In 4 of the remaining comparisons, our system establishes a constraint between the degree of responsibility and the value of an attribution variable for each agent, but cannot infer an ordinal relation between those control variables. In the remaining case our model is inconsistent; comparing the vice president in scenario 1 with the vice president in scenario 3, the survey respondents attributed less blame to the agent who had foreknowledge but was coerced than to the agent with no foreknowledge at all. Interestingly, this constitutes a violation of the strict ordering of the modes of judgment assumed in Shaver's model. The vice president in scenario 1 has no foreknowledge of the environmental harm and thus no intention to cause it, placing the judgment of his responsibility in the causal without foreknowledge mode. On the other hand, the vice president in scenario 3 has foreknowledge but is strongly coerced, placing the judgment of his responsibility in the intentional but coerced mode. This overlap between these two states indicates that, under some circumstances, an agent acting under coercion with full foreknowledge of the consequences may be counted less responsible than one who simply does not know the outcome. The first scenario was worded differently than the others, in that the vice president is presented as initiating a new program that the chairman had no prior knowledge of. In the other scenarios the program is assumed to already be known to both participants. Further, the vice president in scenario 3 is explicitly shown to have expended some amount of effort to avoid the outcome. We suspect that these differences introduce a variable of personal desire to run the program or not on the part of the vice presidents, which is distinct from intention and not accounted for in the current model nor in the underlying theory.

Based on the four cases where our model infers a constraint with a free variable, we can make predictions about additional constraints in the attribution variables. Given that participants attributed a higher degree of blame to the vice president in scenario 1 over the chairman, our model predicts that they would also indicate that the vice president was more responsible than the chairman to know that environmental harm would result from the new program. Likewise, as the respondents attributed equal blame to the chairmen in scenarios 2 and 3, our model predicts that they would judge the outcome intention of the chairmen as being equal as well. This is consistent with the implicit claim in attribution theory that, while coercion mitigates the responsibility of the coerced, it has no such effect on the responsibility of the coercer, who is judged on his intention instead. Finally, as the respondents attributed less blame to the vice president in scenario 4 than to the chairmen in scenarios 2 and 3, our model predicts that they would judge the outcome intention of that vice president to be less than the outcome intention of each chairman, respectively.

Discussion

We have shown that QP theory can be used to formally encode a model for attributing responsibility for negative outcomes, based on Attribution theory. Our model explains the corporate scenario data better than Mao's model does, due to our use of qualitative representations instead of categorical, Boolean values. While a purely qualitative model would not be sufficient for all purposes – for example, deciding whether or not someone was blameworthy enough to report an action – our evaluation suggests that qualitative modeling captures an important level of reasoning about social situations. Even when numerical models are desired, working out qualitative models first could provide constraints on more detailed models. And, as our demonstration of the violation of an assumption of Shaver's theory indicates, formally

encoding qualitative models and examining ordinal fits with human data could provide social scientists with a new set of tools for exploring the consequences of their theories.

As part of our ongoing work on narrative understanding, we intend to incorporate this model into our Explanation Agent natural language understanding system [Kuehne 2004]. This work represents part of a larger effort to model and reason about moral decisions presented in folktales and the explanatory stories that people tell. In that context, we plan to expand the factors that go into judging attribution variables beyond plan recognition and order negotiation speech acts and do further evaluation of the validity of those judgments and the predictions made by this model regarding the attribution of blame.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research. We thank Jon Gratch for useful conversations.

References

Allen, J. F. and D. Waltz. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26, 832-843.

Austin, J. 1975. *How to Do Things with Words*. Harvard University Press.

Bratman, M. 1990. What is intention? In P. Cohen, J. Morgan & M. Pollack eds., *Intentions in Communication*. MIT Press.

Forbus, K. 1984. Qualitative Process Theory. *Artificial Intelligence*, 24, 85-168.

Forbus, Kenneth D. & Sven Kuehne. 2005. Towards a qualitative model of everyday political reasoning. *Proceedings of the Nineteenth International Qualitative Reasoning Workshop*.

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

E. E. Jones and K. E. Davis. 1965. From Acts to Dispositions: The Attribution Process in Person Perception. In L. Berkowitz ed. *Advances in Experimental Social Psychology* (Vol.2), pp. 219-266. Academic Press.

Kamps, J. and Peli, G. 1995. Qualitative reasoning beyond the physics domain: The density dependency theory of organizational ecology. *Proceedings of the Ninth International Qualitative Reasoning Workshop.*

H. H. Kelley. 1973. The Processes of Causal Attribution. *American Psychologist*, 28:107-128.

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

Kuehne, S. E. 2004. On the Representation of Physical Quantities in Natural Language Text. *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*, Chicago, Illinois, USA, August.

Mao, W. 2006. *Modeling Social Causality and Social Judgment in Multi-Agent Interactions.* Doctoral dissertation, University of Southern California, Los Angeles, California.

Mao, W. and Gratch, J. 2005. Social Causality and Responsibility: Modeling and Evaluation. *Fifth International Conference on Interactive Virtual Agents*.

Shultz, T.R. & Schleifer, M. 1983. Towards a refinement of attribution concepts. In J. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 37-62). London: Academic.

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

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

Steinmann, C. 1997. Qualitative reasoning on economic models. *Proceedings of the Eleventh International Qualitative Reasoning Workshop.*

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

B. Weiner. 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. MIT Press.