A Theory of Apologies
(preliminary draft – comments welcome)

Benjamin Ho[*]
Stanford Graduate School of Business

Revised: March 2005
First Draft: May 2003

ABSTRACT

Apologies are a previously unstudied social institution integral in the maintenance of relationships within society. Their application ranges from corporate culture to political systems to legal settings and beyond. This paper formulates a game theoretic signaling model using rational agents that serves as a framework for understanding apologies and their use. The base model uses costly signaling, but further models use cheap talk, where signaling is achieved by a multi-dimensional type-space based on status or empathy. A positive explanation of the fundamental attribution error from social psychology is also provided.

[*] Thanks to Edward Lazear for excellent advising, as well as Douglas Bernheim for comments.

1 Introduction

> Paul had a problem. Several weeks ago he made a date to play tennis with his friend Amy at noon on Saturday. Paul arrived at the courts right on time, but Amy shows up one hour late. Paul is angry but after Amy apologies profusely, Paul readily forgives her and they both enjoy the tennis match. They make a date for the following week, whereupon Amy is again late, again apologizes, and again is forgiven. On the third week, Amy is late once more, and at this point Paul is fed up. Talk is cheap, why do apologies have any meaning?[1]

A common opening for papers about apologies across the social sciences is to note the dearth of articles about apologies in their respective field. I will begin no differently. A search of Econ-lit produces no results on the subject of apologies, yet this social institution has wide ranging significance. Beyond the use of apologies in daily interpersonal interactions, apologies appear in organizational design, political reputations, legal litigation, international relations, corporate governance, and beyond. In an interconnected world where economic actors are all embedded in a network of relationships (Granovetter 1985; Sen 1977; etc.), apologies act to restore frayed connections.

My approach is to develop a rational choice framework to understand apologies in a principal-agent context. I avoid behavioral/psychological assumptions and primarily limit my analysis to preference for consumption, as this allows applicability to potentially highly rational actors such as politicians and governments. I argue that apologies exist to maintain relationships. However, if apologies help relations, talk is cheap, why is it that not everybody apologizes?

This paper focuses on interactions where exogenous factors makes payments infeasible, whether, for example, because of legal reasons—the inability for a politician to bribe the electorate for example—or because of social norms—gifts between friends tend to be limited to symbolic gestures. For simplicity, I consider one-sided interactions, where both the principal's and the agent's payoff

[1] The story is true, but the names and details have been changed to protect the innocent.

are dependent on the hidden actions the agent takes. One can think of a large number of symmetric interacting actors where random tasks arrive to randomly assigned principals who must seek out an agent for its successful completion.

Agents come with different types. The principal would prefer to partner with an agent suitable to the task but is uncertain of the agent's type. Suitability is defined as the agent that maximizes the principal's expected utility. The principal's beliefs regarding the agent's type are indicative of the strength of the relationship. After the agent has chosen her action and the outcome of the task is observed, the agent can choose to apologize, signaling her type for the purposes of securing future interactions.

In the language of psychology, the apology shifts the principal's *attribution* of the cause of the task's outcome, blaming the failure on random chance, i.e. the situation, rather than the agent's type, i.e. her disposition.

In the example above, Paul[2] is randomly chosen as the principal and assigned the task to find somebody to meet to play tennis. Paul chooses Amy to be the agent based on his beliefs regarding her punctuality, i.e. her type. During the first two apologies, Paul is willing to attribute the lateness to random justifiable events, for example traffic. After the third time, Paul concludes that he was wrong with his earlier attributions and potentially ends the relationship.

In this paper, I establish a general framework for thinking about apologies applicable to a broad class of games, and then propose and examine several game theoretic mechanisms that describe the apology interaction. I begin with the simplest model, and suppose that an apology is a costly signal where "good" types find it easier to apologize, in the spirit of Spence (1973; etc.). I then build on that framework and consider a typology of apology mechanisms that allows me to endogenize the cost. By introducing a second dimension of type to preference alignment, call it control or status, I offer a cheap talk model of

[2] I am well aware that I am confounding gender roles by making the principal a "he" and the agent a "she" but it was a true story, and I believe that true gender equality means it should not matter.

apologies. I also briefly outline two alternatives based on social contracts, and finally empathy.

In the sparse psychology and sociology literature, as well as through introspection, various properties of apologies can be extracted and various questions that can be asked in a rational framework can be raised. Why do people apologize? When do people apologize? Why are they accepted? When are they accepted? What is the cost of an apology? What is the benefit? Why does the impact of apologies deteriorate with time? Importantly, if talk is cheap, then why is it that people do not always apologize?

The fundamental goal of this project is to provide a framework for a rational explanation that explains the workings of a social institution that is presumed to be based on emotion. With an understanding of how the mechanism could work given rational actors, then we can begin to understand how the mechanism arose evolutionarily, and how and why we pass on notions of guilt and remorse to our children. The ancillary goal is that in the process, I formally define common notions such as mistake, sympathy, empathy, situation, disposition, and intention in ways that make them tractable by game theory.

The purpose of the model is that it can be used to help understand important policy questions. For example, the recent emergence of "I'm sorry" laws in California, Massachusetts, Texas, Florida and other states that make apologies inadmissible in courts. I also examine prominent political apologies. The popular press does not understand why politicians never apologize. Additionally, I consider the implications of apologies on organizational design. By making formal terminology and systems that have previously only been informally discussed, I seek to bring clarity to these issues.

The paper proceeds as follows: Section two presents a few important concepts for thinking about apologies, and how those concepts will be used in this paper. Section three contains a review of the broader social science literature on apologies, followed by a review of the economic tools used in this paper, and ends with a review of the psychological theory of attribution that is foundational in informing the model. Section four presents the basic model of apologies with exogenous costs. Section five provides several alterative specifications that endogenizes the cost of apology. The first generalizes this model to a contracting framework. The other two allow multi-dimensional type, the added dimension representing first empathy and then competence. The paper concludes with an analysis of several cases and a check of the empirical evidence.

2 Apology Concepts

Everyone has some notion of what it means to apologize. In order to crystallize ideas before proceeding, I lay out the framework that I will use for the remainder of the paper. An apology occurs between two actors, a principal and an agent. One can think of the principal and the agent as being selected at random from a larger community where the principal needs a task accomplished and solicits the agent. For simplicity, presume there are two types of agents (though the results largely generalize), one of which has preferences better aligned with the principal's. Formally, let the agent's type have increasing differences with respect to the principal's expected payoff in the agent's utility function. In a political game, this would be closeness of ideal point. In a standard principal-agent framework, this would be the productivity of the agent. In a simple divide the dollar game, this would be a Fehr-Schmidt (1999) or Becker (1976) style altruism parameter. Social psychologists would call this type, the agent's *disposition*, and I will speak of good dispositions and bad dispositions. The American Heritage Dictionary (2000) defines *sympathy* as "A relationship or an affinity between people or things in which whatever affects one correspondingly affects the other." Thus, the type θ is also a measure of sympathy.

The broad sketch of the game proceeds as follows: nature draws the agent's type, θ, the agent chooses an action x, there is randomness imposed by the environment in the form of another draw by nature, ω, and the outcome, $y(x,\omega)$, is mutually observed by both principal and agent. An apology can be put forward or not, and the agent receives continuation payoff, $v(-)$.

(Insert Figure 1 about here)

The model is essentially a signaling game appended to a moral hazard problem. The action of the apology mechanism occurs primarily with the signaling, and thus for much of the paper, I will consider a reduced form of the game that takes the agent's action as exogenously fixed. However, the action is important because the opportunity must be provided for the agent to make a mistake. The American Heritage Dictionary (2000) defines an apology as "An acknowledgment expressing regret or asking pardon for a fault or offense." Regret or fault requires a mistake. A mistake for homo economicus is the difference between what is ex post optimal and what is ex ante optimal. The environment variable, ω, can represent either new information about the world, or new thinking and better reflection by the agent.[3] It is important for mistakes to be modeled using a random move by nature, because alternatives such as off equilibrium play or trembles presumes that mistakes occur with near zero probability.

Thus, using this framework, an apology by the agent is an effort to signal that her original choice of x represented good intentions. Effectively she would have chosen a different x had she known the realization of the state of the world ω. Social psychologists call this state, the *situation*.

The principal can either attribute a particular outcome, y to the agent's disposition, a low θ, or to the agent's situation, a bad draw of ω. The resolution of this dilemma is the subject of attribution theory (Ross, 1977; etc.) in psychology whose findings are elaborated upon in the next section. An apology is an attempt by the agent to shift the principal's attribution from disposition to situation.

The precise specification for an apology differs in each of the four treatments offered here. However, the commonality is that an apology entails a

[3] Another interpretation suggested by Bernheim and Rangel (2004) is that ω is a mood, a temporary shock to preferences that are not indicative of typical preferences. The setup

message, $a \in \{0,1\}$, where a one indicates an apology. One might argue that a richer message space may offer different results, but especially for my cheap talk models, as Crawford-Sobel (1982) shows, in a sender-receiver game with imperfectly aligned preferences, only a finite number of different messages are possible, and there are many equilibria. In this paper, I consider the equilibrium where there are two messages. I also do not distinguish between a non-apology, and inaction. I interpret inaction to be a message of $a = 0$.

Another key component of an apology is its value in the restoration of relationships (Tavuchis, 1991; Ohbuchi et al., 1989; McCullough et al., 1998; etc.). The larger ecology of apology is beyond the scope of this paper, but if we assume that actors attempt to maintain relationships with those of higher (more aligned) type, then the principal's beliefs regarding the agent's type can be seen as a measure of the strength of the relationship.

Finally, as this paper considers situations where payments between actors are disallowed, $v(-)$ is a payoff based on the probability the principal decides to continue the relationship, i.e. forgive. Alternatively, it is based on the probability that the principal stays with the same agent for this task, rather than switch to a different one. At times, it will be useful to think of $v(-)$ as the continuation value of a stationary game where the above framework is a stage game. This notion of forgiveness is very reduced form. A more in-depth study of forgiveness in left for future research.

3 Literature Review

3.1 Applications

To emphasize the import of apologies in many economic and political interactions, I highlight six applications in particular.

would be slightly different than that which is presented here, but the analysis would largely be the same.

3.1.1 Interpersonal Relationships:

Apologies are a common occurrence in everyday life, particularly in the maintenance of friendships. Empathy is a particularly important component of these apologies. Several questions will be addressed regarding cultural differences in apologies—Asians apologize more than Americans (Takaku et al., 2001; etc.)—or gender differences—women apologize more (Gallup, 1989 in Tavuchis, 1991).

What psychology literature exists is focused on experimentally validating stylized facts about the apology process. An apology by the agent reduces the anger the principal feels toward the agent as well as the principal's desire to punish (Ohbuchi et al., 1989). In tasks where the agent is less responsible or where the offense is less severe, the apology is rejected less often (Bennett and Earwalker, 2001). Apologies are almost always accepted (Mullet et al., 1998; Bennett and Dewberry, 1994). Forgiveness occurs more often in closer relationships (McCullough et al, 1998).

The results of all the psychological experiments I found were consistent with the findings in the models presented here.

3.1.2 Organizations:

The prevalence of apologies in various organizational settings is indicative of differences in task assignment, in risk taking, in turnover, etc. A coherent model of apologies offers insight into cultural differences in organizational design.

Lee and Tiedens (2001a) find that within an organization, when individuals of high status—those with control—make excuses for their behavior, they suffer a decrease in status.

3.1.3 Corporate Governance:

CEOs are expected to be responsible to shareholders. When performance is low or scandal arises, should an apology be expected? Does an apology carry any weight? Lee and Tiedens (2004) find that kinds of attributions for past performance found in company annual reports—effectively kinds of apology—can predict a firm’s stock prices one year out.

The failure of business ethics due to the economist’s profit maximization assumption has received recent attention (Ghosal, 2005). Apologies can be used to reflect a firm’s ethical domain.

3.1.4 Politics:

There is a stylized fact that politicians never apologize. Consider, Bush on Iraq, Clinton on Lewinsky and Berlusconi on Germany. Lee and Tiedens (2001) conducts an experiment where she constructs two videos, one where Clinton appears to apologize regarding the Lewinsky affair, and one where he appears angry. Subjects who saw apologetic Clinton liked him more, while subjects who saw the angry Clinton liked him less and complained about how Clinton never apologized. However, on questions of leadership, questions of ability, and importantly, questions of whether you would re-elect, the angry Clinton did better. Their result is robust to choice of politician and the crime.

3.1.5 International Relations:

An apology by a government is important either between the government and its people (e.g. South African apartheid, Japanese interment, or the United States civil war), or between governments in international relations (the difference between Germany and Japan’s response to World War II).

3.1.6 Law:

In recent decades in the US, apologies have become increasingly important in litigation damages. California, Florida and other states have recently passed laws to prevent apologies from being considered as evidence in order to encourage their use. Apologies are especially relevant in medical malpractice, as a vicious circle has arisen. Doctors are afraid to apologize because of the large risk of lawsuits. Patients are more likely to sue due to anger for not receiving an apology. (Taft, 2000; Cohen, 2000; Latif, 2001; Cohen, 2002)

3.2 Mechanisms

The psychology literature provides many properties of the apology mechanism, but is largely deficient at systematically describing the process. One common reason given for why people apologize is “negative affect alleviation.” People feel guilty, and an apology removes the guilt (??,??). This view dates to a Freudian model of human behavior where humans have stocks of emotion—in this case guilt—which when accumulated, causes distress until emptied.

Tavuchis (1991) has an extensive sociological treatment of apologies, in which he sees apologies as a complex social system designed to maintain relationships and to establish membership in community. Tavuchis describes an apology as a kind of social exchange, a device that somehow restores social order paradoxically without altering the thing which is being apologized for. Nothing material has been exchanged, yet the relationship has changed. An apology is painful, and the pain is created by the social system of shame that accompanies it.

These psychological and sociological models of behavior, along with the notion of behavioral scripts or heuristics pushed by Kahneman and Tversky (1974; etc.) and others while not consistent with economic actors, may be accurate. However, the model I propose is more teleological. In the mode of Frank (1989), culture and evolution has provided devices such as guilt, remorse and shame, in order to facilitate the apology mechanism for the purpose of maintaining the social links that society depends upon. Thus, I retain the assumption of rational actors for modeling purposes.

Attribution theory from social psychology is more useful for illuminating how apologies work, though it is still incomplete. The main question addressed by attribution theory concerns a situation when an outcome is observed—Amy is late for a tennis match—that has two possible causes: dispositional—Amy is lazy and inconsiderate—or situational—Amy was held up by unexpected traffic. Van den Steen (2004) provides a rational explanation of the self-perception bias of attribution theory which has observed that individuals tend to attribute their own successes to their disposition while attributing their failures to the situation

(Bradley, 1978; Zucherman, 1979). This paper will address the fundamental attribution error where individuals tend to attribute too much blame to the disposition of the actor and not enough to the situation (Heider, 1958; Ross, 1977; Jones and Nisbett 1972). Further research refines this notion. The ultimate attribution error (Pettigrew, 1979; Bodenhausen, 1988) argues that attributions are made that assume the worst about an agent who is in the principal's outgroup.

Weiner et al. (1987) applies attribution theory to the process of apologies, and in experiments, finds that apologies that attribute bad outcomes to external uncontrollable factors, ω, are more likely to increase liking than apologies that attribute bad outcomes to internal controllable factors, x. This simple insight demonstrates the basis of the models in this paper. Effectively, I argue that an apology by an agent shifts the principal's attribution of the cause of the bad outcome to external factors.

3.3 Economics

The closest economic literature gets to apologies, are the repeated game strategies such as Tit-for-Tat studied by Axelrod (1984), etc., or the repeated game punishment scenarios of Green and Porter (1984), etc. In this paper, I focus more on the apology mechanism itself, a self-contained event that does not depend on punishment or penitence. This paper also focuses primarily on rational mechanisms, largely avoiding behavioral assumptions.

The four models presented here elucidate four different mechanisms by which apologies can operate. They are intended to be a typology of possible apologies, and useful in illuminating how apologies function.

The base case is the simplest where an apology is a costly signal of type. The cost can be either a tangible cost such as legal sanction, or a more behavioral notion such as shame.

In the remaining models, an apology is cheap.

The second model generalizes the base case by considering the apology mechanism as a social contract, and considering the properties of such a contract that will ensure truthful revelation.

The third and fourth models allow apologies to be cheap by adding an additional dimension of type. In the third model, actors are characterized as having empathy or not. Empathy—the understanding of another's situation or feeling (American-Heritage, 2000)—is modeled as having knowledge by the actor of the principal's payoff in the actor's information set.

In the fourth model, actors are characterized both by a preference alignment dimension, θ, but also by a degree of control or competence dimension, η. An apology shifts attribution from one characteristic to the other. Incentive compatibility is maintained by the principal's choice of future tasks. The fourth model can be seen as an endogenization of the cost function used in the base case.

4 Base Model

4.1 Setup

The timing of the base model proceeds as in Figure 1. The utility of the Agent is given by a linear combination of the agent's utility from the inputs and outputs, the cost of apologizing, and the agent's discounted payoff:

$$U_A(x,\omega,\theta) = u(x, y(x,\omega),\theta) - c(\omega)1_{a(\omega)} + \delta\, v(b(a(\omega), y(x,\omega))) \tag{1}$$

Nature sets the agent's disposition, $\theta \in \{\theta^B, \theta^G\}$, who then chooses an action, $x \in X$, to maximize her expected utility over possible realizations of the situation, $\omega \in \Omega$, where $\omega \sim F(-)$. After observing the outcome, $y(x,\omega)$, she chooses to apologize or not, $a(\omega) \in \{0,1\}$, and incurs the cost, $c(\omega)$, only if $a = 1$.

δ is the discount factor and reflects either time preferences or the likelihood of repeated interaction, thus δ goes down as the number of other players increases.

The principal observes only the outcome and whether the apology was tendered or not, and from that updates his beliefs, $b(a, y)$, where p is his prior:

(2) $$b(a, y) = \Pr(\theta = \theta^G \mid a, y, p)$$

One can think of the principal maximizing his utility in each stage game with the function given by:

(3) $$U_P(y) = \sum_t \delta^t y_t$$

The agent's type θ is a measure of preference alignment, in the following sense. If we let $E_\omega y(x,\omega)$ be invertible in x, then we can rewrite $Eu(x, y,\theta)$ as $\tilde{u}(Ey,\theta)$. By preference alignment, we mean $\tilde{u}(Ey,\theta)$ has increasing differences in θ and $\mathrm{E}y$. That is to say, higher agent types either value the principal's utility more, or it is easier for higher agent types to provide principal's with higher utility.

This specification for the utility function and production technology is somewhat awkward, but many common moral hazard situations are captured. For example, the standard moral hazard with high and low productivity where x represents effort:

(4) $$u(x, y,\theta) = -\frac{x^2}{\theta}$$ $$y(x,\omega) = \begin{cases} 1, & x + \omega > K \\ 0, & x + \omega < K \end{cases}$$

A second example is a political game, where there is a uni-dimensional policy space, x represents the agent's choice of policy, the principal has an ideal point of zero, while the agents have ideal points away from zero, and $1/\theta$

represents the agent’s ideal point. This functional form can also model the Paul and Amy interaction, where x is the choice of departure time, and ω is the amount of traffic:

$$\begin{aligned} &u(x,y,\theta) = -(x+\omega-\frac{1}{\theta})^2 \\ &y(x,\omega) = \begin{cases} 1, & x+\omega < K \\ 0, & x+\omega > K \end{cases} \end{aligned} \tag{5}$$

A third example might have θ as an altruism parameter, and the task is some noisy division of a pie, where the agent’s choice x is how much of the pie the agent keeps for herself.

$$\begin{aligned} &u(x,y,\theta) = \theta y + x \\ &y(x,\omega) = \begin{cases} 1, & x+\omega < K \\ 0, & x+\omega > K \end{cases} \end{aligned} \tag{6}$$

Given that $\tilde{u}(Ey,\theta)$ has increasing differences, then, $U_A(x,y,\theta)$ also has increasing differences, and so the θ^G agent always chooses $x = x_G^*$ that yields a higher Ey for the principal, and the θ^B agent always chooses $x = x_B^*$ that yields a worse outcome for the principal. Note that increasing differences depends on the lack of dependence of $v(-)$ on type. This independence is assumed to simplify the analysis and clarify the signaling mechanism. However, when $v(-)$ is taken as a continuation value later on, dependence on θ will be necessary.

The production technology, $y(x,\omega) \in \{0,1\}$ is uniquely determined by the agent’s action and nature’s draw of the situation. To simplify analysis, define:

$$\begin{aligned} \Omega^G &= \{\omega : y(x_G^*,\omega) = 1\} \\ \Omega^B &= \{\omega : y(x_B^*,\omega) = 1\} \end{aligned} \tag{7}$$

Then, assume that the set of states $\Omega \setminus \Omega^G$ where the good type fails is a subset of the set of states where the bad type fails, $\Omega \setminus \Omega^B$. Essentially, some states are just universally harder to succeed in than in other states. This assumption is actually stronger than necessary, but it is intuitive.

Note, in this analysis presented, variables such as type and output are all dichotomous, though these can be generalized to continuous variables without any substantive change in the results.

The principal's only action is to decide whether to continue or end the relationship. As in the career concerns models of Dewatripont et al. (1999), there exists a market of outside options that forces the princpial's "payoff" to be commensurate with his beliefs regarding the agent's type. The principal's action is effectively constrained by a renegotiation proofness condition and outside options.

In terms of the model at hand, the principal decides whether to continue based on the chance to interact instead with an outside option: a competing agent's who's prior is drawn from some distribution. The principal continues the relationship if his beliefs about the agent's type is higher than his prior regarding the outside option's type. This strategy establishes an exogenous probability of whether the relationship continues as a function of the principal's beliefs regarding the agent's type. This probability multiplied by the value of the relationship to the agent justifies the exogenously defined $v(b(-))$. The only restriction on $v(-)$ is that it be monotonically increasing in $b(-)$.

4.2 Cost Function

This model restricts the dependence of the cost function to only the situation, ϖ. Dependence on type or intention could easily be added, but only complicates the issue without adding intuition. In this model, the cost of signaling is tied more directly to the agent's action rather than the agent's type as in traditional signaling models.

Additionally, the cost of apologizing is given exogenously instead of being chosen by the agent. One might think it more natural for the cost to be chosen by the agent, instead of being prescribed. For example, the agent could choose the size of the box of chocolates, or a drug company with a defective product might choose how long to hide the information before apologizing. However, since I am interested in situations where pooling occurs, I would focus on equilibrium where both types choose the same cost. In which case, one could think of the exogenously given cost function, $c(-)$ as the internal unobserved cost required to achieve the equilibrium externally observed cost.

The cost function thus defined can have multiple interpretations:

4.2.1 Tangible Cost

A tangible cost is the simplest burn money scenario such as gifts of roses or some other symbolic gift or gesture. Amy stands outside in the rain until Paul forgives her.

Alternatively, there can be legal sanction. An apology after a car accident increases the probability of losing a lawsuit.

4.2.2 Ambiguity Resolution

It is possible that the outcome is imperfectly observed, or observed but not contractible, thus an apology is an admission of fault and resolves the ambiguity. This still depends on their being some exogenous cost that is a function of the certainty that a bad outcome was attained. An apology by a politician regarding a policy failure confirms that the outcome really is a failure. An apology by a defendant confirming his guilt establishes proof beyond a reasonable doubt.

4.2.3 Lying

The word apology derives from the Greek for story. An apology is an example of account giving or excuse giving. The cost of lying is the probability the lie will be found out. Amy claims she was late due to traffic, but there is some chance Paul will hear a traffic report and catch her lie. Or alternatively, many argue that humans are evolved to be poor liars and dishonesty can be detected by facial cues (Ekman, 1969; Frank, 1989). In either case, the possibility exists that a lie will be found out, and the cost represents the punishment associated with lying, determined socially outside of this model.

4.2.4 Status

A common reason given that it is difficult to apologize is that an apology entails a loss of status. This will be modeled more explicitly, effectively endogenizing the cost, in Section 1.

4.2.5 Psychic Costs

The costs can be entirely behavioral, whether through norms of shame that surround the very act of giving an apology, mitigated by alleviation of guilt. Tavuchis (1991) argues that society makes it painful to apologize through norms of social disapprobation, because the pain is what gives apologies meaning. The model here gives testament to this view of apologies. Again, although there is considerable psychological and sociological evidence that such mechanisms are at work, this paper focuses on rational explanations.

4.3 Analysis

To simplify notation, recall p is the principal's prior and $b(-)$ is the principal's posterior belief that the agent is a good type after observing the outcome and the apology (or lack thereof). Define:

$$\begin{aligned} b_g &= b(a=0,y=1)=\Pr(\theta=\theta^G \mid y=1) \\ b_1 &= b(a=1,y=0)=\Pr(\theta=\theta^G \mid y=0,a=1) \\ b_0 &= b(a=0,y=0)=\Pr(\theta=\theta^G \mid y=0,a=0) \end{aligned} \tag{8}$$

In the case of success, $y=1$,[4] we do not define a variable for beliefs in case of apology, because we will show that agents never apologize in case of success.

[4] This analysis could be extended to allow for a continuous range of y, rather than dichotomous. The result would have some y^* above which both agents never apologizes (i.e. success), and below which they sometimes do (i.e. failure). Further, the principal has more information on which to condition his beliefs. Qualitatively, however, things look the same.

By definition of conditional probability,

(9) $$p = \Pr[\theta^G \mid y=1]\Pr[y=1] + \Pr[\theta^G \mid y=0]\Pr[y=0]$$

Solving the model by backward induction, the agent only apologizes if

(10) $$v(b_1) - v(b_0) > c(\omega)$$

The left hand side is independent of ω, so there is some cut point, c^* where the equation is satisfied. Define Ω^A to be the set of states of ω where $c(\omega) \leq c^*$.

By Bayes' Rule:

(11) $$\begin{aligned} b_1 &= \frac{F(\Omega^A \setminus \Omega^G)p}{F(\Omega^A \setminus \Omega^G)p + F(\Omega^A \setminus \Omega^B)(1-p)} \\ b_0 &= \frac{F((\Omega \setminus \Omega^G) \cap (\Omega \setminus \Omega^A))p}{F((\Omega \setminus \Omega^G) \cap (\Omega \setminus \Omega^A))p + F((\Omega \setminus \Omega^B) \cap (\Omega \setminus \Omega^A))(1-p)} \end{aligned}$$

It is possible for corner solutions to arise where there is pooling and Ω^A is empty. The rest of the analysis assumes an interior solution.

Note that Ω^A non-empty and since $c(-)$ is assumed positive we get

(12) $$b_1 > b_0$$

Which implies that good types always apologize more than bad types.

(13) $$\Pr[a=1 \mid \theta = \theta^G] \geq \Pr[a=1 \mid \theta = \theta^B]$$

By the assumption that $\Omega \setminus \Omega^G \subset \Omega \setminus \Omega^B$, the set of states of the world where the good type apologies, $\Omega \setminus \Omega^G \cap \Omega^A$, is weakly smaller than the set of states where the bad type apologizes, $\Omega \setminus \Omega^B \cap \Omega^A$. Even if both types always apologize, then at best $b_1 = p$. If there were any states where an apology would be too expensive for one type but not the other, it would only be in cases where the bad type failed when the good type would have succeeded. Thus $b_1 \leq p$.

Combining equations 9 and 12 allows us to write:

$$b_g > p \geq b_1 > b_0 \tag{14}$$

Successes strengthen relationships. In case of failure, apologies remedy the relationship and restore the relationship but only imperfectly so. If we treat $v(-)$ as a continuation value and have the players myopically repeat the game, an apology after one failure may restore the relationship sufficiently to continue the game, but a succession of failures can leave b low enough for the principal to end it. Hence, Paul's frustration after Amy's third episode of tardiness.

Taking the limit in equation 11 as p goes to zero or to one, means the priors begin to dominate the posteriors, which means the differences in beliefs $b_1 - b_0$ goes to zero. Thus

$$\lim_{p \to 0} F(\Omega^A) = \lim_{p \to 1} F(\Omega^A) = 0 \tag{15}$$

This implies that when the principal is fairly confident in the agent's type, the prevalence and impact of an apology goes to zero. Thus, the impact of an apology is maximized when there is uncertainty about the agent's type. When the principal is fairly certain of the agent's type, then the impact of an apology is minimal, and thus the cost is not worthwhile. When the principal has decided that the agent is no good, an apology will not change his mind.

Similarly, this result also implies that "Love means never having to say you're sorry" (Segal, *Love Story*, 1972). This second implication, however, is not robust to all specifications. In fact, in the other models, love implies you always apologize. However, this inconsistency reflects the inherent controversy of the movie quote. Linguist Deborah Tannen (1996), for example, argues that Segal's quote is exactly wrong. The other models will better address her concerns.

The results are summarized in the following graphs:

(Insert Figure 2 about here)

4.4 Attribution Theory

With this base model in hand, we can begin to consider the theory's implications in regard to attribution theory. The principal upon observing the outcome can make two possible inferences, either the bad outcome was due to a bad type, the actor's disposition, or a bad draw of nature, the actor's situation.

A simple by-product of the nature of the model is that in the absence of an apology, the principal assumes the cause was situation. Essentially, our principal commits the fundamental attribution error. The commission of this error is somewhat a product of the model, but it is an essential product.

The intuition is that if some costly mechanism exists to change someone's beliefs, or attributions, then the default must be to think the worst of the other person, otherwise they would have no incentive to correct the misattribution. Thus when observing an agent fail at an action, we always assume the actor was at fault, because had we assumed it was the environment, the agent would have no incentive to correct it, and the environment is by definition incapable. This interpretation of attribution is consistent with the extensive evidence though it applies only to failures. You would expect the opposite biases in attribution when it comes to successes. Most of the experimental evidence has focused on the negative. Furthermore, studies involving the ultimate attribution error and hostile attribution bias confirm this asymmetry (Pettigrew, 1979; Bodenhausen, 1988; etc.).

Though this model does not address self-perception bias—people attribute their own successes to their disposition, and their own failure to the situation—the same intuition applies. Incentives need to be provided to others to correct the misperceptions, and if no correction is forthcoming, then we are right to adjust out attribution accordingly.

Psychologists have been good at identifying sources of bias in human decision making but have largely ignored the question of why these biases exist. The model presented in this paper demonstrates that the bias exhibited in the fundamental attribution error is not a mistake, but a rational response given a fuller understanding of the social interaction. The model also allows predictions as to when the fundamental attribution error is more likely to be observed. In addition to the asymmetry between positive and negative outcomes, the error is most likely to occur when uncertainty about the other's type is highest. The error is also likely to be more pronounced when the cost of apology is high, such as when individual status is more important that community relationships. Since Asian cultures tend to exhibit far more apologies than Western cultures, one would expect Westerners to commit the fundamental attribution error more often. Psychological research bares these predictions out (many cites here, ????).

It should be admitted that in laboratory settings, subjects should be aware that no apology is possible, and so a rational actor should not be subject to such effects. However, one can presume that these beliefs arise from heuristics developed for the real world, and thus are short circuited in laboratory settings.

5 Cheap Apologies

The model presented in the previous section is effective at explaining many of characteristics observed in apology interactions. However, its reliance on an externally given cost function leaves the modeling unsatisfying. In this section, I return to the question originally posed: if talk is cheap, why do apologies have any meaning? From here on out, an apology will simply be a message, and its meaning will be derived in equilibrium. I begin with two alternative mechanisms based on social contracts and empathy. I then finish with a status model that endogenizes the cost in the base model using multi-dimensional type.

A note on continuation values: For the remainder of the paper, it will be convenient to think about $v(-)$ as a continuation value, and thus it is necessary to introduce the dependence on type, thus we have $v(b,\theta)$. The main problem here is that this dependence may induce strange behavior in that it may encourage good types to purposefully fail at the task in order to signal her type. However, recall that $v(-)$ is discounted by δ. In all future examples, assume δ sufficiently small so that increasing differences still holds, so that the good type will always produce better results for the principal. That is, presume either sufficiently impatient actors, or, better yet, presume a large enough community that ensures frequency of interaction to be sufficiently rare. Note, that this assumption only impacts the moral hazard part of the model. The apology signaling mechanism payoffs occur all in the same period, and thus are unaffected.

5.1 Status

A common reason that is given as to why apologies are difficult is because an apology entails loss of status. One could argue that the reason women apologize more than men is because evolutionary pressure has made status more important for men, because men need status to compete for mates. Thus, since status matters relatively less for women, they can apologize more. The same can be said for Asian cultures relative to Western ones. If Asian cultures value group preference alignment more than individual ability, then apologies will be more prevalent in Asian cultures. Social psychologists have also found that culture matters in attribution error (Markus and Kitiyama, 1991; Nisbett, 1993). For example, Iyengar and Lepper (1999) find that Asians are less likely to commit the fundamental attribution error providing further evidence of the connection between attribution and apology. This section will formalize the notion of status and by doing so, provide a more compelling reason why cheap talk apologies are effective.

Tiedens (2001) experimentally demonstrates that even though politicians gain approval and liking by apologizing, a apology causes the politician to lose status as measured by respect or willingness to re-elect. Up until now, the point of an apology was to shift the principal’s attribution of the cause of a bad outcome from an internal controllable quality of the actor, to an external uncontrollable quality of the environment. Lee and Tiedens (2001a) find that even a successful shift of attribution may not be a good thing for the agent, if the agent is expected to have control over the environment. They find no such downside to apologies when the agent is not expected to have control.

5.1.1 Setup

To model this quality of control, consider another dimension of type for control called $\eta \in \{\eta^H, \eta^L\}$ which represents either high control or low control. For the rest of this section, I will assume a specific functional form for simplicity. Let the alignment type be

$$\begin{aligned} \theta^G &= 1 \\ \theta^B &= -\infty \end{aligned} \tag{16}$$

and let the control type be

$$\begin{aligned} \eta^H &= 0 \\ \eta^L &= \infty \end{aligned} \tag{17}$$

Restrict the action space so that $x \in \{0,1\}$ which represents either positive effort, $x = 1$, or no effort, $x = 0$.

Let the utility function[5] be given as

$$U_A(x, y, \theta, \eta) = \theta y - \eta x + \delta v(b^\theta, b^\eta, \theta, \eta) \tag{18}$$

Finally, let the production technology be simply

$$y(x, \omega) = x \tag{19}$$

Note that for the purposes of this example, I ignore randomness in situation. I do this in order to limit attribution to two possibilities again, thus simplifying analysis. Aligned agents want the same outcome as the principal, while misaligned agents will do anything to avoid it. However, only high control agents are capable of producing.

For this model, I will treat $v(-)$ as a continuation value and give the principal a more sophisticated strategy. There could be multiple stages, but I will restrict to two periods for illustration purposes. In the second period, there is no

[5] So θ and η cannot technically take on infinite values as they occasionally need to be multiplied by zero. Assume very large numbers instead.

$v(-)$. The principal, now, instead of deciding whether to continue or discontinue the relationship, chooses instead which task to offer.

As one further addition, instead of assuming that tasks are identical, I instead assume that there is a sequence of possible tasks. For each task, i, the agent has a pair, (θ_i, η_i). For simplicity, let the prior for all θ be p and let the prior for all η be q. The principal does not know the values of the agent's types, only the priors. The agent does know the correlation between any two tasks, however:

(20)
$$\begin{aligned}\rho_{i,j} &= Corr[\theta_i, \theta_j] \\ \phi_{i,j} &= Corr[\eta_i, \eta_j]\end{aligned}$$

For example, after playing tennis with Amy, Paul does not know if she will be late for lunch, but he does know that chronic lateness for tennis is highly correlated with chronic lateness for lunch.

5.1.2 Analysis

Given these extreme parameters, first period output is given by

θ	η	y
G	H	*1*
G	L	*0*
B	H	*0*
B	L	*0*

Output is only produced if both preferences are aligned and the agent is in control. Given these payoffs, a simple equilibrium strategy for the principal is to assign tasks based on the following rule:

- If $a = 1$ choose a task with $\rho = 1, \phi = 0$
- If $a = 0$ choose a task with $\rho = 0, \phi = 1$

In words, if the agent apologizes, offer her a task that is perfectly correlated along the alignment dimension, if the agent does not apologize offer a task that is perfectly correlated along the control dimension. In Tiedens' (2001) political example, if the politician apologizes, offer him a task in the next period based on character, such as dating your daughter. If the politician does not apologize, offer him a task in the next period based on control or ability or competence, such as running the country.

The agent only receives positive utility if both her preferences are aligned, and her control is high. The second period payoffs for an agent are:

$$U_A = \Pr(\theta_2 = \theta^G \mid \theta_1)\Pr(\eta_2 = \eta^H \mid \eta_1) \tag{21}$$

Thus, given the above strategy of the principal, the second round payoffs for an agent who failed are

θ	η	$E_{\mid a=1}y$	$E_{\mid a=0}y$
G	L	*q*	*0*
B	H	*0*	*p*
B	L	*0*	*0*

Thus assuming indifference is properly resolved, good types will always apologize, while bad types will not. More generally, for any arbitrary correlations $(\rho_{i,j}, \phi_{i,j})$ for the next task, the payoffs to the agent will be

θ	η	Ey
G	H	$[p+\rho(1-p)][q+\phi(1-q)]$
G	L	$[p+\rho(1-p)][q(1-\phi)]$
B	H	$[p(1-\rho)][q+\phi(1-q)]$
B	L	$[p(1-\rho)][q(1-\phi)]$

Thus, the principal can maintain this cheap talk equilibrium so long as he offers two tasks that satisfies the proper incentive constraints. Assume the principal focuses on choosing tasks that separate the (G, L) types from the (B, H) types. Whether the (B, L) types apologize will depend on the tasks available.

One can think of a larger ecology of actors, where each period, a randomly chosen principal has a task that requires an agent. If there are a sufficient number of diverse agents the principal can choose from, then the principal can commit to assigning that task only to an agent where the correlations satisfy the IC constraints from their previous interaction.

Other things to consider is that if there were added an additional exogenous cost of apology, such as the legal sanctions from before, then this would shift the set of tasks that satisfy the IC constraints. Laws that change these costs could again impact this interaction.

Additionally, it should be noted that apologies will generally increase liking, as θ's are updated upward, while non-apologies will do the opposite. Also, any particular action will in fact shift beliefs about all other tasks via the correlation matrix, thus complicating analysis if extended to more than two periods, though not substantively so.

Thus, returning to the aforementioned gender and cultural differences, men apologize less because they value tasks that depend on control more than women. In Asian cultures, tasks are primarily based on preference alignment, and individual initiative is deemphasized, thus making apologies more common.

5.2 Alternative I: Social Contracts

There are two alternatives to adding a dimension of control to type. The first depends on allowing the principal to commit to changing the payoffs in future periods.

Generalizing the model from section 4, it is possible to effectively subsume the cost of apologies into the continuation value, and think about specifying $v(-)$ as some sort of social contract that maintains the apology mechanism. Then, what is necessary is to have a payoff function, $v(-)$, such that the following incentive compatibility constraints are satisfied:

$$\begin{aligned} v(b_{a=1} = 1, \theta^G) &\geq v(b_{a=1} = 0, \theta^G) \\ v(b_{a=1} = 0, \theta^B) &\geq v(b_{a=1} = 1, \theta^B) \end{aligned} \tag{22}$$

One such way to obtain such a payoff function in the first stage is to have a two-period game with appropriate payoffs in the second stage. I call this a “fool me once, shame on you, fool me twice, shame on me” contract. The intuition is that the principal would like to know the private information of the agent, but the agent has incentive to misrepresent her type, so an apology is a claim to be a good type. However, if the agent claims she is a good type, the principal will demand much more out of the agent, and tolerate failure far less, whereas if the agent does not apologize, the principal will be more forgiving of failure. The following is an illustrative example.

(Insert Figure 3 about here)

Assume for simplicity that preference alignment is strong enough that second stage probability of success is fixed at s^G for the good type and s^B for the bad type. Then to satisfy the IC conditions in the first stage, we need:

$$\begin{aligned} s^G(v_{a=1,y=1} - v_{a=0,y=1}) &\geq (1 - s^G)(v_{a=0,y=0} - v_{a=0,y=0}) \\ s^B(v_{a=1,y=1} - v_{a=0,y=1}) &\leq (1 - s^B)(v_{a=0,y=0} - v_{a=0,y=0}) \end{aligned} \tag{23}$$

Effectively, since $s^G > s^B$, the marginal benefit of success in the second stage in the case of an apology in the first, must be higher than the marginal benefit in case of no apology in the first.

There are issues of renegotiation proof-ness that will not be dealt with here.

5.3 Alternative II: Empathy

In this section, I consider again the possibility of another dimension of type, but instead of control, I consider *empathy*, which will be a reflection of the information structure of the game.

Though the primary purpose of an apology, and the primary dictionary definition, is in relation to a fault or offense, the notion of apologies is often conflated with a general sense of empathy, or awareness of the other’s emotional state: “I am sorry to hear that your grandmother died.” Specifically, this section tries to understand partial apologies, those apologies that do not come with an admission of guilt.

Alternatively, empathy can be interpreted as awareness of the agent of what the principal considers appropriate rules of conduct. The apology act could be an implicit part of some bargain over appropriate norms of behavior.

To capture this interaction, in addition to a preference alignment type, θ, let there be an empathy type, $\tau \in \{0,1\}$, where empathic and non-empathic types differ only in their information sets; non-empathic types do not observe the principal's payoff, y.

Again, empathy could be made a continuous variable by specifying information sets over states of the world, ω, rather than outcomes, y, with agents that have greater empathy having a finer partition. However, I again favor simplicity.

To make empathy relevant, let there be a positive correlation, ψ, between the two types, either because the empathic types are more effective at producing given their better understanding of the principal, or for some external common reason such as better upbringing.

In such a game with cheap apologies, consider an equilibrium where empathic types always apologize in case of failure, and never apologize in case of success, and non-empathic types never apologize.

Given such an equilibrium, the principal's beliefs of the agent's empathy will be (with prior on $\tau = q$):

$$\begin{aligned}
&\Pr[\tau = 1 \mid a = 1, y = 1] = 0 \text{ (off - equilibrium)}\\
&\Pr[\tau = 1 \mid a = 0, y = 0] = 0 \text{ (off - equilibrium)}\\
&\Pr[\tau = 1 \mid a = 1, y = 0] = 1\\
&\Pr[\tau = 1 \mid a = 0, y = 1] = \Pr[\tau = 1] = q
\end{aligned} \tag{24}$$

An inappropriate apology automatically gives the non-empathic types away. An appropriate apology proves empathy, and conveys information on preference alignment via the correlation with theta. Then, the principal's updated beliefs regarding the agent's preference alignment, θ, in such a game are:[6]

(25)

$$b(a=0,y=1)=\frac{F(\Omega^G)p}{F(\Omega^G)p+F(\Omega^B)(1-p)}$$
$$b(a=1,y=0)=\frac{F(\Omega\setminus\Omega^G)\Pr[\tau=1\,|\,\theta=\theta^G]p}{F(\Omega\setminus\Omega^G)\Pr[\tau=1\,|\,\theta=\theta^G]p+F(\Omega\setminus\Omega^B)\Pr[\tau=1\,|\,\theta=\theta^B](1-p)}$$
$$b(a=0,y=0)=\frac{F(\Omega\setminus\Omega^G)\Pr[\tau=0\,|\,\theta=\theta^G]p}{F(\Omega\setminus\Omega^G)\Pr[\tau=0\,|\,\theta=\theta^G]p+F(\Omega\setminus\Omega^B)\Pr[\tau=0\,|\,\theta=\theta^B](1-p)}$$

Using the positive correlation between τ and θ we know that

(26)
$$\begin{aligned}&\Pr[\tau=1\,|\,\theta=\theta^G]>1/2>\Pr[\tau=0\,|\,\theta=\theta^G]\\&\Pr[\tau=0\,|\,\theta=\theta^B]>1/2>\Pr[\tau=1\,|\,\theta=\theta^B]\end{aligned}$$

so we can get:

(27)
$$b(a=1,y=0)>b(a=0,y=0)$$

This equilibrium strategy is optimal for the empathic types, since in the event of failure, the principal believes that the agent is a high type more if an apology is given. The equilibrium strategy is optimal for the non-empathic types so long as:

(28)
$$\begin{aligned}&F(\Omega^i)v(b(0,1))+F(\Omega\setminus\Omega^i)v(b(0,0))\\&\qquad>F(\Omega^i)v(b(1,1))+F(\Omega\setminus\Omega^i)v(b(1,0))\end{aligned}$$

[6] These expressions can be simplified using the fact that the conditional probability for two correlated Bernoulli variables with the same prior is given by
$$\Pr(\theta=\theta^H\,|\,\tau=1)=p+\psi(1-p)$$
$$\Pr(\theta=\theta^H\,|\,\tau=1)=p(1-\psi)$$

Or re-arrange to get the probability of success times the marginal benefit of apologizing in case of success must be greater than the probability of failure times the marginal benefit of apologizing in case of failure:

$$F(\Omega^i)[v(b(0,1)) - v(b(1,1))] > F(\Omega \setminus \Omega^i)[v(b(1,0)) - v(b(0,0))] \tag{29}$$

Essentially, if we assume that in most situations, the agent is successful and thus an apology is unwarranted, then the non-empathic agent finds it optimal to never apologize to avoid revealing her lack of empathy.

It is useful to note that once the apology is made, in a simplified model without any further noise and empathy is stable across time, the principal learns for sure that the agent is empathic. Thus over repeated interactions, the principal quickly becomes aware that an agent is empathic. However, repeated failures would still lead the principal to conclude the agent is a bad type, and thus the principal will end the relationship anyway. This rapid devaluation of the perfunctory apology corresponds to the real world observation that often such apologies seem meaningless. Once empathy has been established, further apologies have little impact on the principal's beliefs regarding the agent's type. The relatively minor impact of apologies in this scenario accords with the assertion that these apologies—apologies without admission of fault—are only partial apologies. However, if ever an agent fails to offer even a partial apology when it is expected, judgments can shift quickly in this off the equilibrium path scenario.

Another interpretation of this model is in the situation where an apology is tendered before the principal is even aware of the mistake. In addition to the tangible costs of ambiguity resolution (see Section 4.2.2), an apology also demonstrates awareness than a transgression occurred and an apology is warranted.

Alternatively, one might think of a world where different standards of behavior are possible. In one culture, being an hour late is unacceptable while for another, being an hour late is a virtue. An apology can be thought of as an acknowledgement by the agent that she violated a norm according to the standards of the principal. An apology indicates a shared agreement of the norms of behavior, or at the very least, an awareness by the agent of what the principal considers are the norms of behavior.

Incidentally, this model applies equally well for other perfunctory pleasantries such as “thank you” or “congratulations.”

6 Discussion

In this section, I consider the implications of the models presented on two applications, the use of apologies in legal settings, and the implications of apologies in regard to corporate culture.

6.1 Apologies and the Law

One of the few areas of scholarly research that examines intensively the question of apologies is in the area of law. Apologies have an important impact on the outcome of cases. Unsolicited apologies can have an impact on conviction rates, as well as sentence and judgment sizes (Rehn and Beatty, 1996). Furthermore, even court ordered apologies appear to mitigate punishment (Latif, 2003). Yet as a result, many legal scholars (Cohen, 2002; Taft, 2000) worry that by using apologies the courts are interfering with a “natural” process of social remediation. This concern has prompted lawmakers to consider legislation that exempts apologies from use in the courts. In this section, I examine the implications of the model applied in the legal system.

The model provides several implications on the impact of apologies as part of legal settlements. Consider sentences that mandate apologies. If there is a fixed amount of punishment associated with any given crime, then one would expect mandated apologies to be accompanied by lower sentences and smaller fines.

The model also predicts that apologies have higher impact when there is more uncertainty regarding the agent's type. If one were to estimate a probit analysis of an individual's guilt based on the observable characteristics of the case, then apologies should have a larger effect on sentences and fines for cases with greater uncertainty. Flanagan (????) finds that the cases that do go to trial are the most uncertain.

In order to assess the impact of an apology on legal outcomes, a central tension between culpability, x, and character, θ, must be resolved. A fair presumption might be that convictions and findings are based on culpability. However, the size of the sentence or the award is based on character (????????).

Consider cases where an apology is voluntarily tendered. If an apology is considered a partial admission of guilt, then one would expect voluntary apologies to be associated with higher conviction rates, but lower sentences, presuming the court would give leniency for good intentions.

Rehm and Beatty (1996) suggest one way to separate the culpability dimension from the moral character dimension by considering attorney discipline cases where the court decision should be based only on character. Using case studies, they argue that in such cases, apologies always help.

The consequences of apology are of special concern in the area of medical malpractice. Doctors are typically told to avoid admissions of fault, and apologies in particular, because of the risk of lawsuits (Pinkus, 2000; Novack et al., 1989). Yet patient lawsuits in the event of error are rare, and only occur when the patient has a poor relationship with the doctor (Gallagher et al., 2003; Huycke and Huycke, 1994). A lack of apology often is the reason why a lawsuit is even filed (Hickson et al., 1992; May and Stengal, 1990). Thus the lack of apologies often results in a vicious cycle that recent "I'm sorry" legislation hopes to break (Cohen, 2003). States ranging from Massachusetts, Texas and California have passed or are considering legislation that exempts expressions of remorse from being used as evidence in civil cases.

A quick implication of this legislation is that in states where "I'm sorry laws" have been implemented, a voluntary apology should have no impact on evidence, but potentially gives information about intentions. If intentions did matter in determining conviction, then voluntary apologies should then be associated with lower conviction rates, relative to voluntary apologies in other states.

The "I'm sorry" laws can be modeled using the base model as an exogenous shift in the cost function. If we presume that the natural process of remediation is based on psychic costs imposed by social norms, then the total cost of apology will be legal costs plus psychic costs. Many of the legal reforms proposed will reduce the cost of apologies in an effort to increase their use. Though the exact effect is dependent on the shape of the reduction, it is useful to consider the case where the cost is just a constant function: $c(\omega) = \kappa$.

Then, a decrease in the cost of apologies, κ , will increase the likelihood of apology, but it will also decrease the benefit, $b_1 - b_0$, that an apology induces. Apologies will have less impact, potentially counteracting the reform. The attenuated impact of an apology is reflected in moral arguments made by Taft (2000) that apologies have been subverted by the legal system. Taft argues that if laws specially exempting apologies from the legal system, apologies would lose part of their moral weight that comes from the associated responsibility that an apology entails.

Furthermore, a decrease in the cost of apology will reduce the costs associated with failure, thereby decreasing the incentives in the action stage, and increasing the equilibrium probability of failure. This increase in moral hazard is echoed by Cohen (2002) who worries that the predicted decrease in law suits files will have a detrimental impact. Already, very few cases of medical malpractice come to trial (Huycke and Huycke, 1994). One could argue that since these lawsuits are essential for restorative justice and efficient monitoring, welfare would be enhanced if there were more lawsuits, not fewer.

Another quick testable implications, in states where such laws have been passed, apologies should be more common, but at the same time less effective. Thus the fraction of patients that pursue lawsuits given an apology should increase.

One point of note is that except for Connecticut and Hawaii, the “I’m sorry” legislation in most states exempt only partial apologies, that is apologies that contain no admission of guilty (Cohen, 2002). Thus, one might say that the empathy model (Section 5.3) might be more applicable.

6.2 Apologies in corporate culture

Another natural venue where apologies are likely to be observed is within an organization. Corporate culture is question oft posed in economics but has yet to be fully answered (Kreps, 1986; Lazear, 1989; etc.). One way to tackle corporate culture is by observing that in certain organizations, apologies are common, while in other organizations, apologies are uncommon. A natural question to ask is what are the correlates with the prevalence of apologies.

This section is going to make the argument that the root of corporate culture is fundamentally economic. The basic argument is that the optimal production technology for a given firm is given exogenously. The production technology implies a given pay structure. The pay structure attracts a particular workforce composition. The pay structure and the workforce composition both lead to firms to adopt what the organizational behavior literature calls either a masculine or feminine corporate culture.

The model provides two reasons why apologies would be common in a given organization. The first is that the cost of apology is low: either the norms specify only minor social disapprobation associated with apologies, or status is not valued. The second reason for common apologies is that the benefits of maintaining relationships are high. Both reasons can also be understood in the context of the status model in an environment where activities that depend on preference alignment are more prevalent that activities that depend on individual competency.

6.2.1 A gender based typology

Consider two firms, firm M and firm F. Firm M has a production technology that stresses individual achievement. Either output is easily contractible or observable, or there are few complementarities in output between workers. Examples would be a sales department based on commission, or manufacturing jobs that typically pay on piece rate. Compensation in firm M is associated with individual achievement while relationships are devalued. Firm M is likely to attract employees that undervalue relationships but value personal control (low θ, but high η). Firm F, on the other hand, has an optimal production technology that emphasizes production in teams. Individual output is unobservable or uncontractable and/or there are complementarities in production. Examples might be marketing departments where creativity is emphasized, or workplaces that emphasize subjective performance evaluations. Firm F is likely to attract employees that have a high value of relationships relative to individual control (high θ, relative to η).

The model predicts that in firm F, apologies should be observed more often than in firm M for two reasons. First, since firm F is likely to attract more high θ individuals, the model predicts that high θ agents apologize more than low θ agents. This is the selection effect. The second reason that firm F will have more apologies is because a high θ is more highly compensated. Thus the incentives to signal high θ are higher. In the base model, this comes from a higher marginal return to a principal's belief that the agent is of high type, i.e. high dv/db. In the status model, this comes from higher returns or higher likelihood of liking (θ) based tasks relative to status/control (η) based tasks.

Since firm F will have workers with higher θ, more sympathy, by our definition of θ, then we would expect a more nurturing environment. Furthermore, the proximate cause of apology is associated with guilt, and thus we would expect guilt to be more prevalent. Conversely, firm M would have a culture that places a greater emphasis on pride and individual achievement.

Additionally, as in Lee and Tiedens (2001), the response to failure is more likely to be anger rather than contrition.

Past organization researches have characterized this typology of corporate culture as being either feminine—where apologies are common—or masculine—where apologies are not (????, ????; working on the refs here).

6.2.2 Apologies and authority

If we considers an extension to the model, we could make additional predictions about organizations and project choice. Consider now the question of organizational design, and consider the apology interaction to be a second stage sub-game, where in the first stage, the agent is given the authority to choose a task. Tasks differ in terms of difficulty in terms of the equilibrium probability of success or failure. We are interested here in the interaction between the cost and the project chosen. If the cost of failure is too high, even a good agent may choose a project that has low risk, i.e. a low chance of failure. An organization where risky projects are optimal may encourage a culture of easy apologies, at the expense of losing the potential gains from sorting that a high cost of apology provides.

Thus, the model predicts that organizations that have team-oriented production functions or risky project choices are more likely to have apologetic culture, while organizations that have individual oriented production functions or safer project choices are more likely to have unapologetic cultures.

6.2.3 External perception of corporate culture

The Kreps (1986) notion of corporate culture is that culture is an asset that represents something comparable to reputation and trustworthiness. If we think about corporations caring about some other dimension besides profitability, then some of the empirical findings about the impact of apologies on shareholder value can be reconciled. This other dimension may be business ethics (Ghosal, 2005), or reputation (Kreps, 1986), or a long-view/high discount factor. Though these last two should be incorporated into views on profitability, there could be market imperfections or short-sightedness that prevent it.

Lee et al. (2004) find that apologies made in corporate annual reports do have an impact on stock price. Specifically, they find that companies that took responsibility for their own poor past performance had a higher stock price one year later. Although there are significant statistical problems in their analysis, the paper raises the interesting possibility that public apologies provide information about some dimension of the company’s future profitability that is not immediately reflected in current shareholder value.

Aaker et al. (2004) consider the relationship between a firm and its customers by conducting an experiment where they set up photo developing companies with differing characteristics, and attempt to measure the reputational effects of an apology after an infraction (the photos were temporarily misplaced). The provided models provide insight into the differential effect apologies have on company reputation as a function of differing priors.

7 Future Research

There is considerable room for future research. Amongst which includes a generalization of the theory of dynamic signaling used in this model. Additionally, a more complete model of the applications of apologies to legal settings or corporate culture is in the works. The theory of apologies itself can be generalized to have a better notion of forgiveness and a better notion of the community of actors.

Formal empirical tests beyond the confirmation of pre-existing psychological studies is another avenue being pursued. The theory could be tested empirically by looking at court cases, political apologies or corporate scandals. Alternatively, data could be collected by surveying the corporate culture of existing firms. Finally, a controlled laboratory experiment is being considered.

References

Aaker, J. Fournier, S. & Brasel A. (2004) When good brands go bad. Journal of Consumer Behavior. Vol 31.

American Heritage Dictionary, Fourth Edition (2000)

Axelrod, R. (1984) The Evolution of Cooperation, New York: Basic Books.

Becker, G. (1976), *The Economic Approach to Human Behavior*, Chicago: University of Chicago Press.

Bengt Holmstrom, (1982). "Moral Hazard in Teams," Bell Journal of Economics, RAND, vol. 13(2), pages 324-340.

Bennett, M. and Dewberry, C. (1994) Ive Said Im Sorry, Havent I - A Study Of The Identity Implications And Constraints That Apologies Create For Their Recipients, *Current Psychology*, 13(1) p10-20.

Bennett, M. and Earwalker, D. (2001) Victim’s Responses to Apologies: The Effects of Offender Responsobility.

Bernheim, D. and Rangel, A. (2004) “Addiction and Cue-Triggered Decision Processes” forthcoming *American Economic Review*

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. Journal of Personality and Social Psychology, 55, 726-737.

Cohen, J. (2000) Apology and Organizations: Exploring an Example from Medical Practice, 27 Fordham Urb. L.J. 1447.

Cohen, J. (2000) Apology and Organizations: Exploring an Example from Medical Practice. 27 Fordham Urb. L.J. 1447.

Cohen, J. (2002) Legislating Apology: The Pros and Cons, 70 U. Cin. L. Rev. 819.

Cohen, J. (2002) Legislating Apology: The Pros and Cons. 70 U. Cin. L. Rev. 819.

Crawford, V. and Joel Sobel, (1982) “Strategic Information Transmission,” *Econometrica*, 50 (6), 1431–1451.

Crawford, Vincent P & Sobel, Joel, (1982). "Strategic Information Transmission," *Econometrica*, Econometric Society, vol. 50(6), pages 1431-51, November.

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole. (1999) "The Economics of Career Concerns, Part I: Comparing Information Structures." *Review of Economic Studies* 66: 183-98.

Ekman, P. (1969). Nonverbal leakage and clues to deception. Psychiatry [0033-2747] Ekman yr:1969 vol:32 iss:1 pg:88 -106

Fehr, E. and Schmidt, (1999), A Theory of Fairness, Competition and Cooperation, in Quarterly Journal of Economics, 817-868.

Frank, Robert (1989) *Passion Within Reason*, W.W. Norton and Company.

G. Bradley, Self-serving biases in the attribution process: A reexamination of the fact or fiction question. Journal of Personality and Social Psychology 36 (1978), pp. 56–71.

Gallagher TH, Waterman AD, Ebers AG, Fraser VJ, Levinson W. (2003) Patients' and physicians' attitudes regarding the disclosure of medical errors. JAMA. 2003 Feb 26;289(8):1001-7.

Ghoshal, S. (2005). Bad Managemnet Theories are Destroying Good Management Practices. *Academy of Management Learning and Education*. 4(1) p75-91.

Granovetter, Mark (1985). Economic Action and Social Structure: The Problem of Embeddedness, American Journal of Sociology, 91, 481-510.

Green, Edward J & Porter, Robert H, (1984). "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, Econometric Society, vol. 52(1), pages 87-100,

Heider, F. (1958) the psychology of interpersonal relations, New York: Wiley

Heider, F. (1958) the psychology of interpersonal relations, New York: Wiley

Hickson GB, Clayton EW, Githens PB, Sloan FA (1992) Factors that prompted families to file medical malpractice claims following perinatal injuries. JAMA. 1992; 267; 1359-1363

Huycke LI, Huycke MM. (1994) Characteristics of potential plaintiffs in malpractice litigation. Ann Intern Med. 1994;120:792-798.

Iyengar, S.S. & Lepper, M.R. (1999) Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology*, 76, 459-366.

Jones, E. E. and Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of the behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins and B. Weiner (eds.), *Attribution: Perceiving the causes of behavior* (pp. 79-94). Morristown, NJ: General Learning Press.

Jones, E. E. and Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of the behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins and B. Weiner (eds.), Attribution: Perceiving the causees of behavior (pp. 79-94). Morristown, NJ: General Learning Press.

Kahneman, D. & Tversky, A. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science* 185, pp1124-1131.

Kreps, D. (1986). "Corporate Culture and Economic Theory", 1986, in Alt and Shepsle, editors, Rational Perspectives on Political Science

Latif, E. (2001) Apologetic Justice: Evaluating Apologies Tailored Toward Legal Solutions, 81 B.U.L. Rev. 289.

Latif, E. (2001) Apologetic Justice: Evaluation Apologies tailored toward legal solutions. 81 B.U.L. Rev 289.

Lazear, E. (1989). “Pay Equality and Industrial Politics,” *Journal of Political Economy* 97:3 (June 1989): 561-80.

Lee, F. Peterson, C. & Tiedens, L. (2004) Mea Culpa: Predicting Stock Prices from Organizational Attributions. *Personality and Social Psychology Bulletin*. 30(12):1636-1649.

Lee, F., & Tiedens, L. (2001a). Who’s being served? “Self”-serving attributions and their implications for power. *Organizational Behavior and Human Decision Processes*, 84(2), 254-287.

Lee, F., & Tiedens, L. (2001b). Is it lonely at the top? Independence and interdependence of power-holders. In B. Staw and R. Sutton (Eds.), Research in Organizational Behavior, Vol. 23, p. 43-91.

M. Zuckerman, Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. Journal of Personality 47 (1979), pp. 245–287.

Markus, H.R. & Kitayama, S. (1991). Culture and the Self: Implicatoins for cognition, emotion and motivation. *Psychological Review*, 98, 224-253.

May ML, Stengal DB. (1990) Who sues their doctors? Law Society Rev. 1990; 24: 105-120.

McCullough, M., Rachal, K., Sandage, S., Worthington, E., Brown, S, Hight, T.. (1997) Interpersonal Forgiving in Close Relationships: II. Theoretical Elaboration and Measurement. *Journal of Personality and Social Psychology*. 75(6) p1586-1603.

Nisbett, R.E. (1993) Violence and U.S. Regional Culture. *American Psychologist*, 48, 441-449.

Novack DH, Detering bJ, Arnold R, Forrow L, Ladinsky M, Pezzullo JC. (1989) Physicians' attitudes toward using deception to resolve difficult ethical problems. JAMA 1989; 261:2980-2985.

Ohbuchi, K., Kameda, M. , Agarje, N. (1989) Apology as Aggression Control: Its Role in Mediating Appraisal of and Response to Harm. Journal of Personality and Social Psychology. 59(2), p219-227.

Pettigrew, T. F. (1979) the ultimate attribution error: Extending Allport's cognitive analysis of prejudice, Personality and Social Psychology Bulletin, 5, 461-476

Pfeffer, Jeffrey (2005). Why Do Bad Management Theories Persist? A Comment on Ghoshal , Academy of Management Learning & Education, 1537260X, Mar2005, Vol. 4, Issue 1

Pinkus, R. (2000) Learning to keep a cautious tongue: the reporting of mistakes in neurosurgery, 1890 to 1930. In Zoloth L, ed. Margin of Error: The EThics of mistakes in the practice of medicine. Hagerstown, Md. University Publishing Group; 2000.

Rehm, PH, Beatty, D. (1996) Legal Consequences of Apology. 1996 J. Disp. Resol, 115

Ross, L. (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (ed.), Advances in experimental social psychology (Volume 10, pp. 173-240), Orlando, FL: Academic Press

Ross, L. (1977) The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (ed.), Advances in experimental social psychology (Volume 10, pp. 173-240), Orlando, FL: Academic Press

Segal, Erich (1972) *Love Story* (movie).

Sen, Amartya (1977) "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," Philosophy and Public Affairs vol. 6, no. 4 (Summer 1977), pp. 317–344.

Spence, A.M. (1974), Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution, *Journal of Economic Theory*, 7(3):296-332.

Taft, L. (2000) Apology Subverted: The Commodification of Apology, 109 Yale L.J. 1135.

Taft, L. (2000) Apology Subverted: The Commodification of Apology. 109 Yale L. J. 1135.

Takaku, S., Weiner, B. and Ohbuchi, K. (2001) A Cross-Cultural Examination of the Effects of Apology and Perspective Taking on Forgiveness, *Journal of Language and Social Psychology*, 20(1,2) pp144-166.

Tannen, D. (1996) "I'm Sorry, I Won't Apologize" *New York Times*, July 21, 1996, Section 6, Page 34, Column 1, Magazine Desk.

Tavuchis, N. (1991), *Mea Culpa: A Sociology of Apology and Reconciliation*, Stanford, CA: Stanford University Press.

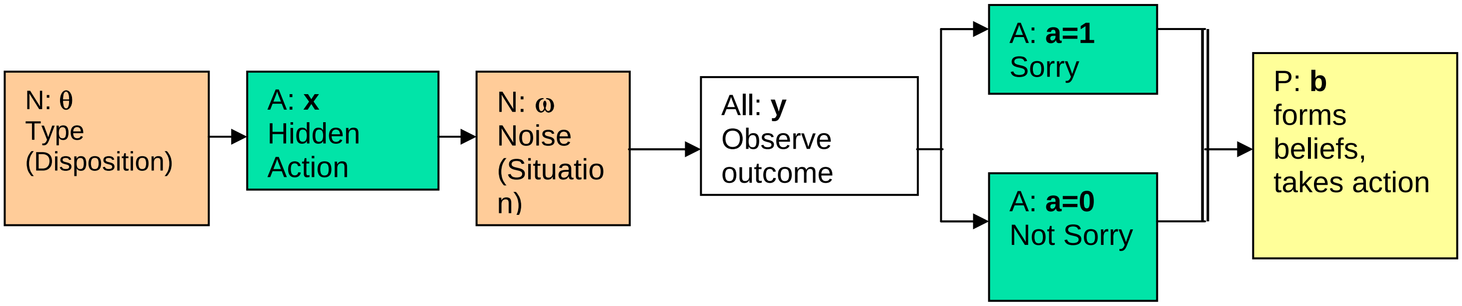
Tiedens, Larissa, (2001) "Anger and advancement versus sadness and subjugation: The effect of negative emotion expressions on social status conferral," *Journal of Personality and Social Psychology*, Jan 2001, 80 (1), 86–94.

Van den Steen, E. (2004) "Rational Overoptimism (and Other Biases)", *American Economic Review*. 94(4).

Weiner, B., Graham, S., Peter, O., Zmuidinas, M. (1991) Public Confession and Forgiveness, *Journal of Personality*, 59(2)

1 Figures

1.1 Figure 1



1.2 *Figure 2*

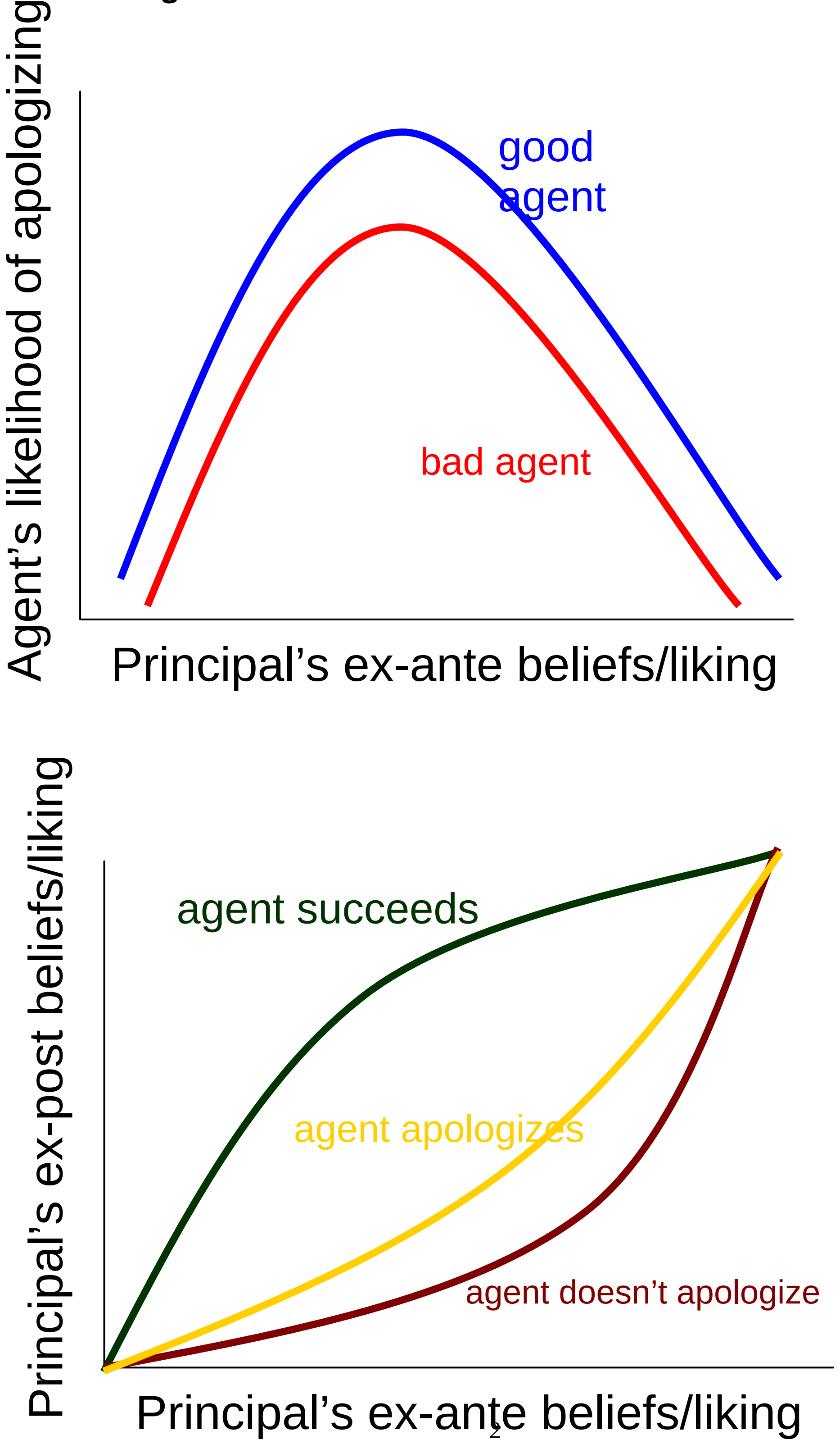
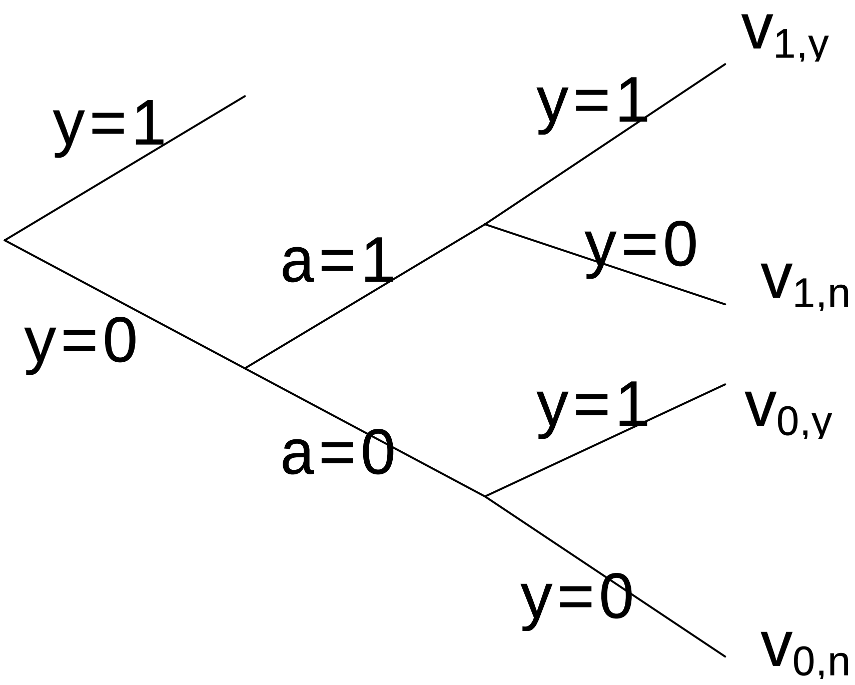

Agent's likelihood of apologizing
good agent
bad agent
Principal's ex-ante beliefs/liking
Principal's ex-post beliefs/liking
agent succeeds
agent apologizes
agent doesn't apologize
Principal's ex-ante beliefs/liking

Figure 3



$v_{1,y}$
y=1
y=1
a=1
y=0
$v_{1,n}$
y=0
y=1
$v_{0,y}$
a=0
y=0
$v_{0,n}$