1 Appendix B: Examples of Moral Hazard Games

In the first view, the agent chooses the action x before she learns the realization ω. In this view, x represents the agent's intentions, before the state of the world, ω, is realized. The state might represent new information about the situation or new thinking by the agent. The optimal x given ω would be different than the ex ante optimal x. This difference represents regret. An apology here signals that the outcome resulted from a bad situation rather than bad intentions.

In the second view, x is chosen after both θ and ω are realized. However, only θ is persistent, and thus for future interactions, the principal only cares about θ. In this view, ω represents a temporary mood as in Bernheim and Rangel (2004). The principal cares only about the agent's disposition, θ, but cannot easily tell given the confound of ω. An apology indicates that despite the bad outcome in the past, the principal can expect a good disposition in the future.

In the third view, ω represents the old type, while θ represents new thinking. However, the hidden action, x was chosen before the new thinking was realized. Again, the principal only cares about future interactions, thus an apology signals the change from old to new type.

The necessary common property for these production games is that the agent's type (θ,ω) is translated into an outcome that yields utility $u(\theta,\omega)$ for the agent and utility $y(\theta,\omega)$ for the principal, where $y(\theta,\omega)$ is increasing in θ via some action, x, by the agent. In most examples of interest, outcome is a function both of type and the agent's action: $\tilde{y}(x,\theta,\omega)$. However, in equilibrium the agent's action is uniquely determined by her type, $x^*(\theta,\omega)$; so long as the outcome function $\tilde{y}(x,\theta,\omega)$ has the necessary properties, we can focus on the reduced form:

$$y(\theta,\omega)=\tilde{y}(x^*(\theta,\omega),\theta,\omega)$$

The natural specification of utility is as a function of action, output, and type: $u(x,y,\theta,\omega)$. So long as $\tilde{y}(x,\theta,\omega)$ is invertible in x, it is possible to rewrite $u(x,y,\theta,\omega)$ as $\tilde{u}(y,\theta)$:

$$\tilde{u}(y,\theta)=E_{\omega}u(x(y,\theta,\omega),y,\theta,\omega)$$

Effectively, instead of choosing an action, the agent is choosing an expected output for the principal. The necessary assumption, then, is a simple restriction on the utility function, that $\tilde{u}(y,\theta)$ has increasing differences in y and θ. Either higher agent types value the principal's utility more, or it is easier for higher agent types to provide principals with higher utility. Thus, by Topkis' Theorem, a higher θ agent maximizing such utility yields higher y.

Recall, however, that the agent's choices are embedded in the larger apology game. Thus, in the larger game, it is necessary for the agent's full utility,

$$U_A(y\mid\theta)=E_{\omega}\left[\tilde{u}(y,\theta)-c(a,\theta,\omega)+v(b(a,y,p),\theta)\right]$$

to have increasing differences in y and θ. Since the continuation value is composed of the production utility and the cost, a simple sufficient condition is that the production utility is increasing in θ and that the cost function is independent of θ. Alternatively, one could assume that the agent has a sufficiently low discount rate for the future such that the supermodularity of the present is preserved.

This specification for the utility function and production technology is awkward, but it captures many common moral hazard problems. Some examples may help clarify.

The first example is the standard moral hazard with high and low productivity where x represents effort. The agent's cost of effort is increasing in x, but higher types have a lower marginal cost of effort.

$$u(x, y, \theta) = -\frac{x}{\theta}^2$$
$$y(x, \omega) = x + \omega$$

Assuming the noise term, ω has mean zero, this expression yields

$$\tilde{u}(y, \theta) = -\frac{y^2 + Var(\omega)}{\theta}$$

which can be differentiated to show that it satisfies increasing differences.

A second example is a political game, with a uni-dimensional policy space, and 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:

$$u(x, y, \theta) = -(x + \omega - \frac{1}{\theta})^2$$
$$y(x, \omega) = -(x + \omega)^2$$

A third example might have θ as an altruism parameter, and the task is some noisy gift giving game, where the agent's choice x is how much the agent gives to the principal, but the agent's choice is obscured by noise ω.

$$u(x, y, \theta) = \theta y - x$$
$$y(x, \omega) = x + \omega$$

The point again of ensuring supermodularity is to guarantee that higher types, θ, make choices that lead in expectation to higher utility for the principal, y. Having established this requirement, I return to the reduced form specification.

2 Appendix C –Cheap Talk Models of Apologies

2.1 Contracting with Commitment: Social Contracts "I'm sorry I'll never do it again"

One approach to modeling cheap apologies is to take a contracting approach where the principal can commit to future termination strategies, creating a mechanism for ensuring truthful revelation. In the other sections, Markov perfection limits the principal to choosing the most attractive agent. Here, I proceed with the same model from Section **Error! Reference source not found.**, except here there is no explicit cost of apologies, and I use the solution concept of a Markov Perfect Equilibrium that uses both the principal's beliefs and the agent's apology in the prior period as the state variable. I also allow the principal to commit ex ante to a mixed strategy of retaining the agent or not as a function of agent's apology and outcome. I then look for a separating equilibrium where good types always apologize and bad types never do.

Once again, agents produce an output and then choose whether to apologize or not. Principals now, instead of merely choosing between {continue, terminate}, now can choose $\delta_t(a_{t-1}, y_t)$, the probability of termination, as a function of the apology in the previous period, and the current period outcome.[1]

In a separating equilibrium, beliefs as a function of apologies would be

$$b(1, y) = 1 \quad (1)$$
$$b(0, y) = 0$$

[1] It would also be sensible to allow δ_t to depend on a_t, but I do not to simplify the analysis.

To implement such an equilibrium, the principal chooses $\delta(a, y)$ to arrive at a payoff function, $v(b, \theta)$, such that the following incentive compatibility constraints are satisfied:

$$v(1, \theta^G) \geq v(0, \theta^G)$$
$$v(0, \theta^B) \geq v(1, \theta^B) \tag{2}$$

One way to obtain such a payoff function in the first stage of a two-period game is to have 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; 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.

To simplify the problem, assume there are only two possible outputs for the principal so that $y \in \{\underline{y}, \bar{y}\}$.[2] Define the following:

$$s_G = \Pr[y(\theta^G, \omega) = \bar{y}] = F(\{\omega : y(\theta^G, \omega) = \bar{y}\})$$
$$s_B = \Pr[y(\theta^B, \omega) = \bar{y}] = F(\{\omega : y(\theta^B, \omega) = \bar{y}\}) \tag{3}$$

The utility of the agent is given by:

$$U(a, \theta) = u(\theta, \omega_1) + u(\theta, \omega_2) + \delta(a_1, y(\theta, \omega_2))v(\theta) \tag{4}$$

Then the agent's IC constraints so that only good types apologize are

$$E_\omega[\delta(1, y(\theta^G, \omega_2))v(\theta^G)] \geq E_\omega[\delta(0, y(\theta^G, \omega_2))v(\theta^G)]$$
$$E_\omega[\delta(0, y(\theta^B, \omega_2))v(\theta^B)] \geq E_\omega[\delta(1, y(\theta^B, \omega_2))v(\theta^B)] \tag{5}$$

which can be simplified using Equation (3) and rearranged to yield

$$\frac{s_G}{1 - s_G} \geq \frac{\delta(0, \underline{y}) - \delta(1, \underline{y})}{\delta(1, \bar{y}) - \delta(0, \bar{y})} \geq \frac{s_B}{1 - s_B} \tag{6}$$

Moral hazard concerns provides two more constraints—an agent in the second round must get higher payoffs for success than failure:

$$\delta(1, \bar{y}) > \delta(1, \underline{y})$$
$$\delta(0, \bar{y}) > \delta(0, \underline{y}) \tag{7}$$

Combining these constraints gives us the following ordering of $\delta(a, y)$:

$$\delta(1, \bar{y}) > \delta(0, \bar{y}) > \delta(0, \underline{y}) > \delta(1, \underline{y}) \tag{8}$$

Effectively, 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. An apology will lead the principal to believe the agent is a good type, but he will expect better performance from her in the future. In the unreduced form, this means higher effort for the agent after an apology.

Alternatively, the following ordering is also possible given the constraints:

$$\delta(0, \bar{y}) > \delta(1, \bar{y}) > \delta(1, \underline{y}) > \delta(0, \underline{y}) \tag{9}$$

I argue this ordering is unlikely by considering the principal's maximization problem. If the probability of the bad type succeeding is still relatively high, which is likely if the principal has difficulty differentiating between bad and good, then the principal would prefer the ordering given in Equation (8).

Given the above analysis, I now return to the case of Paul and Amy. Paul may tolerate one failure from Amy to show up on time, but given the repeated failure, he is forced to end the relationship.

This game could be extended to N-periods except that after the second period, there is complete separation making signaling uninteresting. To get around this, I relax the complete persistence of type. Let there be some probability of mutation:

[2] These results generalize easily to continuous θ. They are messier but similar for continuous y.

$$\Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^G] = \bar{p}$$
(10)
$$\Pr[\theta_{t+1} = \theta^B \mid \theta_t = \theta^B] = \underline{p}$$

Then even though there is full separation each period, the two period equilibrium can also serve as a Markov perfect equilibrium for an n-period game where the prior is reset to either $\bar{p}$ or $\underline{p}$ after each period.

In any case, the problem with the contract presented in this section is that it is not renegotiation proof. Once an agent has apologized, she has established herself as the good type. At that point, the principal would not want to end the relationship. To solve this problem I introduce the ability for the principal to offer different tasks.

2.2 Contracting without Commitment: Status "I'm sorry; I'm an idiot."

2.2.1 Introducing tasks

One reason why contracting is difficult for the principal in the previous case is the limitation of the principal's strategy space. In this section, I give apologies more sophisticated meaning by expanding the space of principal responses. Now, instead of "continue" or "terminate," I allow the principal to offer a menu of tasks. Let there be a set Z of tasks for each period, each task defined by an ordered triple $(\delta_z, \rho_z, \phi_z)$, where the discount rate δ_z reflects how long before that task comes up again and ρ_z is the correlation of the next task with the current task in the θ dimension and ϕ_z is the correlation in the ω dimension. Until now, I have assumed that θ is identical across periods, while ω was drawn independently, limiting the principal's choice set for the next period to

$$Z = \{(\delta_{cont} = 1, \rho_{cont} = 1, \phi_{cont} = 0), (\delta_{term} = 0, \rho_{term} = 0, \phi_{term} = 0)\} \quad (11)$$

In this section, I expand the set of tasks available to the principal to a larger set. As before, θ represents an internal dimension or disposition, while ω represents an external dimension or situation, but here I consider scenarios where the agent is expected to have some control over her situation. Whereas before, an agent could excuse poor performance to a bad situation, doing so now would admit to a lack of control, which is also bad for the agent. Formally, the change in the model is that now, payoffs to the agent are based not just on the principal's beliefs about θ, but also the principal's beliefs about ω as well.

Returning to the example of Paul and Amy, let the base task be "be on time." Let θ represent how much Amy cares about Paul, and let ω represent how able Amy is at showing up to events on time. Now, consider two other tasks that Paul might like fulfilled: "talk to at party," and "be on time for job interview." One would expect that success at "talk to at party" would be correlated with how much Amy cares about Paul, but not be correlated with how good Amy is at showing up on time for things. Conversely, "be on time for job interview" might depend very much on Amy's ability to show up on time for things, but not depend on Amy's liking of Paul. Thus one might imagine an equilibrium in which if Amy is late and apologizes, then Paul would be happy to talk to Amy if he sees her at a party, whereas if she did not apologize, Paul might be more likely to recommend Amy for a job opening that depended on her on time arrival for the interview.

This expansion of the set of tasks available allows the principal to offer a renegotiation proof menu that allows cheap apologies to carry meaning even if the principal is not able to commit to a contract.

2.3 Model Details

The setup of the game starts again with the base model, and once again, the cost of apology is set to zero. The main change is that now, both θ and ω are semi-persistent, and the degree of persistence—i.e. correlation, across periods—is determined by the principal's choice of tasks. The agent's choice of actions each period is the same as before, she produces an output for the principal, and then, upon realization of the output, decides to apologize or not. The principal offers a menu of two future tasks, one if the agent apologizes, and the other if she does not. The choice of task determines the distribution that governs the agent's type $(\theta_{t+1}, \omega_{t+1})$ in the next period.

The principal's payoffs are the same: maximize his output across periods.

$$U_P = \sum_t y(\theta_t, \omega_t) \tag{12}$$

The agent's per period payoff is also the same, except now, the agent's continuation value depends not just on the principal's beliefs about her internal type, b_θ, but also the principal's beliefs about her external type, b_ω:

$$U_A = u(\theta_t, \omega_t) + v(b_\theta, \theta_t, b_\omega, \omega_t) \tag{13}$$

In a two period game, the agent's second period payoff will depend on the task selected by the principal at the end of the first.

$$U_A = u(\theta_t, \omega_t) + \delta(z_{t+1}(a_t, y_t))u(\theta_{t+1}, \omega_{t+1}) \tag{14}$$

I add a number of simplifying assumptions. None are crucial, but they make analysis more comprehensible. Limit the external dimension to two values so that $\omega \in \{\omega^L, \omega^H\}$ one of which represents low ability, and the other representing high ability. I retain $\theta \in \{\theta^G, \theta^B\}$ so that there are good agents who care about the principal, and bad agents who do not care. Then, for any particular task in period *t*, the agent's type is given by (θ_t, ω_t) which can take one of four values.

Now, assume also that there are only two possible outputs for the principal, $y \in \{0,1\}$, a task can be a failure or a success. Assume also that an agent succeeds at a given task only if she is both of good disposition, and high ability: (θ^G, ω^H):

$$\begin{aligned} y(\theta^G, \omega^H) &= 1 \\ y(\theta^G, \omega^L) &= 0 \\ y(\theta^B, \omega^H) &= 0 \\ y(\theta^B, \omega^L) &= 0 \end{aligned} \tag{15}$$

Similarly, assume that agent's utility is increasing in type. Specifically, assume the agent's utility for consumption is given as:

$$\begin{aligned} u(\theta^G, \omega^H) &= 1 \\ u(\theta^G, \omega^L) &= 0 \\ u(\theta^B, \omega^H) &= 0 \\ u(\theta^B, \omega^L) &= 0 \end{aligned} \tag{16}$$

Note that the agent is also only happy when she is successful. In this simplified form, there is preference alignment between principal and agent, when it comes to the task at hand. This alignment is not necessary, but it makes the conflict of interest introduced by task assignment more apparent.[3]

Some notation will be helpful. Recall for a given task z to be assigned in the next period, ρ_z is the correlation of the new θ_{t+1} with the current θ_t, while ϕ_z is the correlation of ω_{t+1} with ω_t. The principal's prior that the agent is θ^G for a given task z is p_z and the prior that the agent is ω^H for a given task z is q_z. Define:

[3] As before, one could introduce moral hazard via a hidden action for the agent that would yield the same reduced form.

$$
\begin{aligned}
\overline{p}_z &= \Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^G, z] = p_z + \tfrac{\rho_z}{p}\sqrt{p(1-p)p_z(1-p_z)} \\
\underline{p}_z &= \Pr[\theta_{t+1} = \theta^G \mid \theta_t = \theta^B, z] = p_z - \tfrac{\rho_z}{1-p}\sqrt{p(1-p)p_z(1-p_z)} \\
\overline{q}_z &= \Pr[\omega_{t+1} = \omega^H \mid \omega_t = \omega^H, z] = q_z + \tfrac{\phi_z}{q}\sqrt{q(1-q)q_z(1-q_z)} \\
\underline{q}_z &= \Pr[\omega_{t+1} = \omega^H \mid \omega_t = \omega^L, z] = q_z - \tfrac{\phi_z}{1-q}\sqrt{q(1-q)q_z(1-q_z)}
\end{aligned}
\tag{17}
$$

I now look for a renegotiation-proof Markov Perfect Equilibrium in pure strategies again using beliefs and last period apologies as a state space. Given the stark production technology specified in Equation (15), if the principal observes a success, $y = 1$ then he knows for sure that the agent is (θ^G, ω^H) and seeks to assign a task as similar to the current task as soon as possible. That is, a task where δ, ρ, ϕ are all close to one. If Amy shows up on time, then Paul will ask her back to meet again the following week.

In the event of a failure, $y = 0$, the set of possible agent types narrows to $\{(\theta^G, \omega^L), (\theta^B, \omega^H), (\theta^B, \omega^L)\}$. Assuming that success is relatively common so that both p and q are relatively high, the third case would be rare. Consider a strategy by the principal that allows him to distinguish between the first two cases.

The principal offers a menu of two tasks after a failure. An agent who apologizes gets task z_1, and an agent who does not apologize gets task z_0. To get separation, the principal selects tasks so that (θ^G, ω^L) types apologize and (θ^B, ω^H) types do not. The incentive compatibility conditions are:

$$
\begin{aligned}
\delta(z_1)\overline{p}_{z1}\underline{q}_{z1} &\geq \delta(z_0)\overline{p}_{z0}\underline{q}_{z0} \\
\delta(z_0)\underline{p}_{z0}\overline{q}_{z0} &\geq \delta(z_1)\underline{p}_{z1}\overline{q}_{z1}
\end{aligned}
\tag{18}
$$

Ideally, the principal would offer one task perfectly correlated on the internal dimension, $(\delta_{z1} = 1, \rho_{z1} = 1, \phi_{z1} = 0)$, and the other task perfectly correlated on the external dimension, $(\delta_{z0} = 1, \rho_{z0} = 0, \phi_{z0} = 1)$, but the actual assignment depends upon task availability. The conflict arises because payoffs for the principal are undiscounted. I assume that the principal interacts with potentially many agents, and thus could have the task filled by another. Thus, the difference in correlations must be sufficient to overcome the difference in discount rates. For example, in the example with Paul and Amy, the "job interview" task may come quite infrequently, and would have a particularly low discount rate. If no appropriate task is available, then apologies would be uninformative. Often, agents apologize consequence free.

In the case of politics, as demonstrated by Lee and Tiedens (2001), an apology gains favor in the "liking" domain at the cost of the "respect" domain. A president who apologizes for a sexual indiscretion might be liked more, and thus given taks based on liking, such as "dating my daughter," but would not be given further tasks based on judgment, such as "running the country."

Returning to the aforementioned gender and cultural differences, it may be more difficult for men to apologize because they encounter more often (higher δ's) tasks based on status or competence. An evolutionary reason might be because women use measures of status and competence to choose their mates (Cole, Mailath and Postlewaite, 2001). In terms of culture, one might find that cultures that apologize more, such as in Japan, have production technologies based on group production, where preference alignment is more valuable, while in Western cultures, performance pay tied to individual competence is more common.

2.4 Partial Apologies: Empathy "I'm sorry your grandmother died"

In this section, I consider another possible dimension of type, but instead of control, I consider *empathy*, a measure reflected in the game's information structure.

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: e.g. "I am sorry to hear that your grandmother died." Alternatively, empathy can be interpreted as awareness by the agent of what the principal considers appropriate rules of conduct. This section presents a model of partial apologies: those apologies that do not come with an admission of guilt.

To capture this interaction, return to the base model where again, an apology will be a shift in attribution from preference alignment, θ, to environment, ω. Now, let type be three dimensional instead of two, given by the triple (θ,ω,τ) for each period. 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 in their information sets; non-empathic types do not observe the principal's payoff, y.[4]

Let there be a positive correlation, ψ, between θ and τ, either because the empathic types are more effective at producing given their better understanding of the principal, or for some external reason such as common upbringing. Let the prior on τ be defined as $q = \Pr[\tau = 1] > \frac{1}{2}$.

Again, there is no cost of apology so the agent receives utility only from production. Also, this model again restricts the principal's choice set to either stay with the current agent or switch to the outside option. The probability that the current agent is better than the outside option is again given by $\delta(b_\theta(a,y))$. In a two period game, the agent's utility is

$$U_A(a \mid \theta, \omega_1, \omega_2, \tau) = u(\theta, \omega_1) + \delta(b_\theta(a,y))u(\theta, \omega_2) \tag{19}$$

Assume that output, $y(\theta,\omega) \in \{0,1\}$, takes only two values, success and failure.

In such a game with cheap apologies, the agent's strategy is given by her apology decision $a \in \{0,1\}$. This decision is conditioned on y for empathic types, but non-empathic types have only one information set, and so their only pure strategy would be "always apology" or "never." Consider the following equilibrium: Empathic types always apologize in case of failure, and never apologize in case of success. Non-empathic types never apologize.

In this equilibrium, the principal's beliefs about the agent's empathy, τ, is

$$\begin{aligned} &\Pr[\tau = 1 \mid a = 0, y = 0] = 0 \\ &\Pr[\tau = 1 \mid a = 1, y = 0] = 1 \\ &\Pr[\tau = 1 \mid a = 0, y = 1] = q \\ &\Pr[\tau = 1 \mid a = 1, y = 1] = 0 \end{aligned} \tag{20}$$

An appropriate apology proves empathy, and conveys information about θ via its positive correlation with τ. Success is never accompanied by an apology and thus provides no information. If success is followed by an apology, this is off the equilibrium path, and I specify that this indicates a non-empathic type. An inappropriate apology automatically reveals the non-empathic types.

The principal's updated beliefs regarding the agent's θ, are

$$\begin{aligned} b_\theta(a=0,y=1) &= \frac{F(\Omega^G)p}{F(\Omega^G)p + F(\Omega^B)(1-p)} \\ b_\theta(a=1,y=0) &= \frac{F(\Omega \setminus \Omega^G)\Pr[\tau = 1 \mid \theta = \theta^G]p}{F(\Omega \setminus \Omega^G)\Pr[\tau = 1 \mid \theta = \theta^G]p + F(\Omega \setminus \Omega^B)\Pr[\tau = 1 \mid \theta = \theta^B](1-p)} \\ b_\theta(a=0,y=0) &= \frac{F(\Omega \setminus \Omega^G)\Pr[\tau = 0 \mid \theta = \theta^G]p}{F(\Omega \setminus \Omega^G)\Pr[\tau = 0 \mid \theta = \theta^G]p + F(\Omega \setminus \Omega^B)\Pr[\tau = 0 \mid \theta = \theta^B](1-p)} \end{aligned} \tag{21}$$

Using the positive correlation between τ and θ yields

$$\begin{aligned} &\Pr[\tau = 1 \mid \theta = \theta^G] > q > \Pr[\tau = 0 \mid \theta = \theta^G] \\ &\Pr[\tau = 0 \mid \theta = \theta^B] > q > \Pr[\tau = 1 \mid \theta = \theta^B] \end{aligned} \tag{22}$$

so we can get:

$$b_\theta(a = 1, y = 0) > b_\theta(a = 0, y = 0) \tag{23}$$

Empathic types follow this equilibrium, because doing so increases the principal's beliefs that the agent is a high type, *b(a,y)*, and hence increases payoffs. Deviations would decrease payoffs. Non-empathic types have no information about the state of the world and the payoffs to the principal. Since non-empathic agents cannot condition their apologies on the state of the world, they must choose to either always apologize or always not apologize. The equilibrium strategy is optimal for the non-empathic types so long as the expected benefit of not apologizing is greater than the expected benefit of always apologizing:

$$\begin{aligned} &F(\Omega^i)\delta(b_\theta(0,1)) + F(\Omega \setminus \Omega^i)\delta(b_\theta(0,0)) \\ &\quad > F(\Omega^i)\delta(b_\theta`(1,1)) + F(\Omega \setminus \Omega^i)\delta(b_\theta(1,0)) \end{aligned} \tag{24}$$

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:

$$F(\Omega_B^1)[v(b(0,1)) - v(b(1,1))] > F(\Omega \setminus \Omega_B^1)[v(b(1,0)) - v(b(0,0))] \tag{25}$$

Essentially, if in most situations, the agent is successful ($F(\Omega_B^1) >> F(\Omega \setminus \Omega_B^1)$) and an apology is typically unwarranted, then the non-empathic agent finds it optimal to never apologize.

It is useful to note that once the apology is made, in a simplified model without any further noise and type 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 a partial apology when it is expected, judgments can shift quickly against her.

Another interpretation of this model is in the situation where an apology is tendered before the principal is even aware of the mistake. An apology demonstrates awareness than a transgression occurred and that 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.

In the example of Paul and Amy, the first apology for her tardiness demonstrated to Paul that Amy knows enough to apologize for her mistakes, that she is aware of Paul's feelings, or that she acknowledges Paul's view that

[4] Empathy could be made continuous 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.

the tardiness is a mistake. However, once empathy has been established, repeated apologies no longer help. After the third time, Paul effectively concludes that Amy may be aware of his feelings but she is still of a bad type.

Incidentally, this empathy variant applies equally well for other perfunctory pleasantries such as "thank you" or "congratulations."

3 Appendix D – Experiment Instructions

Instructions Thank you for participating in this study of economic relationships in the presence of interaction.

Please be advised that there is no talking once the experiment has begun, except to ask questions. I will be available to answer questions especially during the practice game. Also, kindly turn off all cell phones.

Note: This is a study for an economics research project. It is the norm of the experimental economics profession for the experimenter to never deceive subjects. Please be assured that the game will proceed exactly as described here.

The Experiment There is a number written in the top right corner of this page. That is your ID number. It will be necessary so that proper payouts can be calculated.

You have been randomly assigned to one of the two experiment rooms. If you are assigned to room L5, you will be given the role of First Mover. If you are assigned to room L8, you will be given the role of Second Mover.

This experiment is designed to study two person relationships. You will play the same game five times, each time with a new partner. In each game, each First Mover will be paired with a randomly selected Second Mover to play a game that will last 10 periods. After each game is over, the Second Movers will switch seats and be paired with a new First Mover

The Game Each game will last for 10 periods. You will be interacting with the same partner throughout the 10 periods. You begin each game with 120 tokens. You can earn more tokens throughout the course of the game. At the beginning of each game, a new communication cost will be selected to be used for the duration of the game.

In each period, three things happen.

1) The First Mover (FM) is given 10 tokens to allocate. These tokens can be used for one of two purposes. Each token can either be allocated to the Second Mover (SM), or the tokens can be banked. The FM can allocate between 1-10 tokens to the SM. Tokens not allocated are banked.

2) The SM receives a number of tokens equal to the number the FM allocated **times three**. The SM can also do one of two things, keep the tokens, or allocate the tokens to a project that benefits only the FM. If the project is successful, the FM receives 20 tokens in addition to the ones he previously banked. If the project is a failure the FM receives 0 tokens in addition to the ones he previously banked. Each additional token that the SM allocates to the FM's project increases the probability of success by 5%.

The SM's earnings comes only from tokens he does not allocate to the FM's project. The FM's earnings comes both from tokens that he banked and from tokens earned from successful projects.

3) Both the FM and the SM observe the outcome of the project. The FM observes only whether the project was successful, but **does not observe** the SM's allocation. The SM now can choose whether to pay the communication cost and send a simple message that reads, "I am sorry."

The period thus ends and repeats to step 1.

In Summary, the payoffs for each period are

FM's tokens = 10 – Number Entrusted + Project Earnings

SM's tokens = Number Banked – Communication Costs

Payment At the end of each game, the profits will be displayed. Each SM will switch seats and be paired with a new FM, and a new game will begin with profits reset to 120 tokens. After all the games are complete, one of the games will randomly (by six sided die) be selected for payment. You are only paid the profits that you earn for the randomly selected game. However, it is equally likely for any particular game to be selected, so try to earn as much as possible in each game.

The exchange rate for tokens to dollars: 120 tokens is worth $10. You will be paid rounded up to the nearest dollar.

If you have any questions please ask them at this time.

There will be a practice round before we begin.