Working Paper in Economics No. 689

Would I lie to you? Strategic deception in the face of uncertain penalties

Buly A Cardak, Ananta Neelim, Joe Vecci, Kevin Wu

Department of Economics, January 2017

ISSN 1403-2473 (Print)
ISSN 1403-2465 (Online)

UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Would I lie to you? Strategic deception in the face of uncertain penalties[*]

Buly A Cardak[†], Ananta Neelim[‡], Joe Vecci[§], Kevin Wu[**]

January, 2017

Abstract

Using an experiment we investigate the effect of different centralised punishment mechanisms on deception and beliefs about deception in a principal-agent interaction that resembles many everyday expert advisor - client relationships. Agents have private information to transmit to Principals who must decide whether to follow Agent advice. Across our treatments, Agents face a range of expected penalties for deceptive behaviour with varying severity and monitoring probability. The Bayesian Nash equilibrium of the principal-agent interaction predicts penalties to have no effect on Agent behaviour. We find the magnitude of penalties to have important deterrent effects on deceptive Agent behaviour while Agents do not respond to changes in monitoring probabilities. Principal following behaviour increases in response to high penalties. However, it is unaffected by equivalent increases in monitoring. To help us understand the mechanism through which penalties deter deceptive behaviour, we test whether framing activates norms, providing an additional deterrence effect. We find norms are only activated by large penalties, providing a possible explanation for the impact of penalties on deceptive behaviour.

***Keywords*:** Punishment, Deception, Principal Agent, Norm Induced Behaviour
***JEL codes*:** K42, L51, C91

---

[*] We thank Jordi Brandts, Utteeyo Dasgupta, Lisa Farell, Iain Fraser, Lana Friesen, Robert Hoffman, Fahad Khalil, Gari Walkowitz and participants at the Australia and New Zealand Workshop in Experimental Economics, Association of Economics and Development Studies on Bangladesh Workshop and seminar participants at RMIT for helpful comments and suggestions. Funding provided by University of Nottingham and RMIT University.
[†] Buly Cardak (b.cardak@latrobe.edu.au), Department of Economics and Finance, La Trobe Business School, La Trobe University, Bundoora, Victoria 3086, Australia.
[‡] Ananta Neelim (ananta.neelim@rmit.edu.au) School of Economics, Finance and Marketing, RMIT University, Melbourne, Victoria 3000, Australia. [corresponding author]
[§] Joe Vecci (joseph.vecci@economics.gu.se) Department of Economics, University of Gothenburg, Sweden.
[**] Kevin Wu (kevin.wu@monash.edu.au) Department of Economics, Monash University, Australia.

1. Introduction

The increasing complexity of many tasks and the specialisation required in completing these tasks leads people to often engage experts for advice and assistance. For example, individuals routinely engage experts for financial advice on saving for retirement; experts provide assistance when applying for home and other loans and in the preparation of tax returns; experts are typically engaged when repairing our homes, home appliances and vehicles; and when we are ill we seek medical treatment from medical experts.[1] Given the information asymmetry in such relationships, a common challenge is to minimise the chances of deceptive conduct by expert advisors and the accompanying costs to society. In this paper, we use an experiment to study the effects of uncertain punishment on deceptive behaviour in an asymmetric expert advisor-client relationships. We focus on the trade-off between punishment severity and detection probability, comparing explanations of deception and following based on Becker's deterrence hypothesis as well as models of behavioural norms.

Deceptive behaviour in expert advisor-client relationships has been an increasingly important policy issue due to its economic consequences. In recent research on medical over servicing, Schwartz *et al.* (2014) find between 25% and 42% of US Medicare patients are provided some care of little or no benefit with a total cost of between \$1.9 billion and \$8.5 billion in 2009. This involves agents (health service providers) behaving opportunistically when providing advice and services to uninformed principals (patients).[2]

Another recent example involves the packaging of risky home loans into residential mortgage backed securities (RMBS) in the lead up to the global financial crisis (GFC). Many of these loans did not comply with regulatory guidelines and were not appropriate for securitization. However, the resulting RMBS were misrepresented by banks (expert advisors) as high quality to investors, contributing to the GFC of 2008. Two of the major banks involved, JPMorgan and Bank of America, have agreed to record \$13 billion and \$16.65 Billion settlements respectively. These bank's admissions include "misleading investors about securities

[1] Even experts seek the advice of those more specialised than themselves, for example investment decisions of professional portfolio managers depend on the decisions and advice of ratings agencies and general medical practitioners seek advice from and refer complex patients on to specialists.

[2] It is however complicated by the role of insurance and the fact that patients often do not pay directly for medical treatment. Where health care is provided through full or partial insurance, the Principal may be thought of as the insurer. The patient's interests may also align with those of the insurer in that they do not wish to undergo un-necessary treatments.

containing toxic mortgages" and "and failing to disclose key facts about the quality of securitized loans to investors, including misrepresentations to Fannie Mae, Freddie Mac and the Federal Housing Administration"; for more details see Department of Justice (2013, 2014). Another important factor leading to the GFC is the deceptive behaviour of lenders when signing up home loan borrowers by not fully evaluating or potentially misleading borrowers about their capacity to meet repayments over the life of their loans. In both cases, clients sought advice and services from experts and were often deceived, despite the existence of regulations (including penalties), costing society many tens of billions of dollars. The policy response has been the introduction of new laws and regulations governing the behaviour of financial institutions in these types of interactions through the Dodd-Frank act, 2010.[3]

Given the economic significance and diversity of applications, much work has been undertaken in the economics literature to design and study the efficacy of regulatory mechanisms, including penalties; see Pollinsky and Shavell (2007) for a review of this literature. A key insight from theoretical models is that with uncertain monitoring, greater expected penalties reduce the incidence of crime. Assuming agents are on average risk averse, the deterrent effect of increased penalty severity has been shown to be stronger than equivalent increases in monitoring (Becker, 1968). Recent experimental literature corroborates the first part of these findings, see for example Harbaugh *et al.* (2013) and Khadjavi (2015), with Friesen (2012) providing evidence of the latter; these studies all focus on non-strategic non-interactive contexts.

This literature has yet to examine the impact of uncertain penalties on deceptive behaviour in the strategic expert advisor-client relationships we study here. In this strategic context, Agent (or advisor) payoffs are not only dependent on their actions, but also on the actions of the principal (or client). While penalties decrease the expected monetary returns to opportunistic advisor behaviour, it also serves to reassure the client that an advisor will behave honestly, increasing their susceptibility to opportunistic behaviour. Aware of this latter effect, any deterrence due to penalties will be moderated in equilibrium by a more trusting clientele.[4] This can perhaps explain why, despite the existence of regulations and penalties in these expert advisor-client relationships, and a diverse range of approaches to understanding the

[3] The act simplified disclosure requirements and strengthened penalties for non-disclosure.
[4] These equilibrium effects are highlighted in Gneezy and Rustichini (2004).

implications of penalties, we continue to see experts behave opportunistically and in violation of regulations with accompanying penalties.

In this paper, we use an experiment to study how expert advisors and clients respond to penalties in a stylised principal agent model with information asymmetry. In this setting, expert advisors are the Agents who have private information about the state of the world and are required to signal this information to clients, who are the Principals. By sending deceptive signals Agents can potentially improve their own payoffs in the bad state. Principals are aware of these incentives and must decide whether to follow the signals provided. The experiment builds on the Sender-Receiver deception game of Gneezy (2005) but differs in three important ways: (i) unlike Gneezy (2005), we examine the impact of penalties on deceptive behaviour; (ii) Principals have full payoff information and as a consequence, the message sent by Agents is not cheap talk but is informative, albeit uncertain; and (iii) our choice of payoffs ensure that Agents do not have the incentive to engage in sophisticated deception, which Sutter (2009) has highlighted as one of the motivations for Sender behaviour in Gneezy (2005).

In this experimental setup, we analyse the effect of introducing a centralised punishment mechanism and systematically varying the severity and probability of enforcement of penalties on both Principal and Agent behaviour and beliefs. In all punishment conditions, the penalty for deceptive behaviour is directed at the Agent. Given this design, we characterise the Bayesian Nash equilibria (BNE) for treatments with and without penalties. The model predicts that the introduction of penalties: (i) has no effect on Agent behaviour, explained by a crowding out of the deterrent effect of penalties by increased Agent beliefs of Principal following; and (ii) increased Principal following as expected penalties increase.

Contrary to these predictions, we find in our experiment that penalties deter Agent deception conditional on the penalty being of large magnitude. Treatments with large penalties have the strongest deterrence leading to a 50% reduction in deception compared to the condition without penalties. An equivalent small certain penalty has no effect on Agent deception. We also find increasing the probability of enforcement does not have any additional deterrence effect, even in the presence of large penalties. On the other hand, Principal behaviour is partially consistent with the BNE predictions. Larger penalties increase following rates irrespective of the probability of punishment. However, Principals do not respond to expected penalties as predicted; a small certain penalty did not affect Principal behaviour relative to the control treatment while a large penalty with the same expected value induced greater following.

In a strategic setting beliefs are important and may explain behaviour, particularly Agent responses to penalties. Our experimental design collects information on the beliefs of Principals and Agents. Contrary to expectations, we find with small expected penalties Agent beliefs about Principal following rates are not impacted by our punishment mechanisms. If Agents do not update their beliefs about Principal following, the introduction of penalties may have a deterrent effect on deceptive behaviour. This explanation breaks down for the case of high expected penalties where Agent beliefs about following rates are higher but we observe lower deception rates than in the control.

An alternative explanation for behaviour not account by the BNE model is the impact of norms on behaviour. As argued by Elster (2009) among others, norms are important in explaining behaviour as actions that violate norms are psychologically costly. Individuals trade-off these costs with material benefits when making choices involving deceptive behaviour; Gibson et al. (2013). Penalties not only affect material benefits of an action, but also influence the psychological costs associated with that action. Penalties, in principle, can signal to individuals that an action is undesirable (Benabou and Tirole, 2006) by activating or increasing the saliency of norms (Cialdini *et al.*, 1990), (Krupka and Weber (2009, 2013). Conversely, penalties may serve as a price and if set very low can reduce the psychological costs associated with the undesirable behaviour (Gneezy *et al.*, 2011). Despite the importance of norms on penalties, and the role of penalties in deceptive behaviour, the literature is silent on how norms influence deception with varying penalties. To test this formally, we implement two additional treatments where we vary the material payoffs of deceptive behaviour without introducing the additional psychological costs associated with a penalty. We provide novel evidence that deceptive behaviour is reduced by the introduction of psychological costs only in the presence of large but not small penalties. We then show that these psychological costs explain up to two-thirds of the total effect of penalties on deception. This implies that framing of a penalty has additional deterrence effects that may be unaccounted for in the theoretical and policy literature.

Our work adds to the literature in a number of important ways. First, we bring together the literature analysing deception in experiments in strategic contexts (Gneezy, 2005, Sutter, 2009) with the experimental literature testing the deterrence effects of uncertain penalties on stealing and cheating; Nagin and Pogersky (2003), Friesen (2012), Harbaugh et al. (2013) and Khadjavi (2015).[5] Unlike these separate literatures, participants in our experiments interact strategically *and* face uncertain penalties. As suggested by earlier examples, strategic interactions are common in practice and differ from the non-strategic settings more commonly studied in the literature. In strategic contexts, subjects will be influenced by beliefs about the uncertain behaviour of their interacting partner. We design an experiment to measure deception in this setting while accounting for beliefs. Second, we add to the literature on crime and punishment (Becker, 1968) by building our knowledge of how deterrence mechanisms influence behaviour. While others have shown the relative effectiveness of penalty severity over equivalent increases in monitoring in reducing socially undesirable behaviour (Friesen, 2012, Anderson and Stafford, 2003), we test whether norms can explain changes in behaviour upon the introduction of a penalty. That is, if penalties reduce deception through an increase in the salience of norms. We thereby add to the literature on the effect of social and ethical norms on behaviour (Krupka and Weber, 2009, 2013).

2. Theory and experimental design

We use a principal agent signalling game to examine the effects of severity of sanctions and the probability of enforcement on behaviour. We simulate an environment in which Player 1 (the Agent) observes the state of nature while Player 2 (the Principal) does not. The Agent then either reveals the true state of nature or sends a deceptive signal to the Principal. The Principal must decide whether to follow the signal provided by the Agent, aware that it could be deceptive and the range of potential payoffs. We consider four treatments with all treatments comprising these key characteristics. The key differences between treatments are the possibility of centrally administered punishment of the Agent for deceptive signalling, explained below.

Our experimental design builds on the experimental mechanism of Gneezy (2005) where the Receiver (Principals in our setting) can choose to follow the Sender's (Agents in our setting) signal. However, our design differs from Gneezy (2005) as the Principal has full payoff

[5] Other studies have investigated the effects of external punishment in reducing corruption (Abbink et al. 2002), collusion (Block and Gerety, 1995) and roadway speeding (DeAngelo and Charness, 2012). In a non-regulatory setting, i.e. public goods games with peer punishment, the introduction of such institutions improves cooperation (Fehr and Gachter, 2000, Leibbrandt and Lopez, 2011). The efficacy of these institutions is strongly influenced by how they are selected (Tyrian and Feld, 2006). Peeters et al. (2013) also show the importance of selection of punishment institutions on behaviour in Sender – Receiver deception games. These studies do not investigate an expert advisor – client asymmetric environment that we study here.

information and as a consequence, the message sent by Agents is not cheap talk but rather contains some information, albeit uncertain. Further, our choice of payoffs ensure that Agents do not have the incentive to engage in sophisticated deception.

2.1 Interaction without punishment

First, the state of nature (*N*) is randomly determined as *Good (G)* with 25% probability and *Bad (B)* otherwise.[6] The Agent observes *N* and provides a signal about the state to the Principal. This signal may or may not be the true state of nature. Upon receiving a signal that the state is *Good*, the Principal assigns a probability of q to the likelihood of the state of nature being *Good* and (1-q) to the likelihood it is *Bad*. Analogously, after receiving a signal of *a Bad* state, the Principal assigns a probability of p to the likelihood of the state of nature being *Bad* and (1-p) to the likelihood it is *Good*. The Principal then chooses to follow (*F*) or not follow (*NF*) the Agent's signal. The Principal's optimal strategy is to follow when the Agent sends a truthful signal and conversely, it is optimal to not follow when the Agent misreports the state of nature. The game tree with payoffs is shown in Figure 1, where Z is the penalty for sending a deceptive signal and α is the probability of the penalty being enforced; these features are discussed further below.

Payoffs were selected to ensure the following criteria are observed:

(i) The payoff to both the Principal and Agent are highest when the state of nature is *Good*, the Agent truthfully reports this state and the Principal follows.

(ii) The Principal should always be at least as well off, if not better off, following a signal if it is truthful than following the same signal when it is deceptive. This implies principals must be wary of the truthfulness of Agent claims in signals; i.e. being deceived hurts the Principal.[7]

(iii) Given the state, Agent payoffs are lower if they deceive and are not followed than if they tell the truth and are followed. The idea behind such payoff differences is the Principal discovers the signal was deceptive and the Agent justifiably loses credibility.

(iv) Agents sending truthful signals are never better off when Principals do not follow such truthful signals. This ensures that Agents have no incentive to engage in sophisticated

[6] We use the neutral frame, state *A* and *B* in the instructions. Strictly speaking, state *A* in the game tree in Figure 1 is always a good state only for the Principal; i.e. state *A* is always weakly preferred by the Principal.

[7] In our game, see Figure 1, following a *Bad* signal offers the same Principal payoff whether it is true or false, as following such a signal is treated as a baseline or defensive strategy; e.g. in a financial advisor – investor setting, choosing to not invest when the state is *Good* or *Bad* preserves the investor's (Principal's) capital.

deception (Sutter, 2009) where they report truthfully, with the expectation that Principals will not follow.

An extensive set of possible payoffs would satisfy these criteria. Those shown in Figure 1, with $Z = 0$, provide a simple set of satisfactory payoffs.

Assuming participants are sophisticated and risk neutral we can identify multiple sequentially rational Perfect Bayesian Equilibria. While there is no separating equilibrium for this game, we can identify a pooling equilibrium at (*B B*, *NF F*; 0.25, 0.75).[8] As this equilibrium implies the Principal places no value on the signal received, i.e. cheap talk, we discount it from further consideration.

[Figure 1 around here]

Instead, we consider mixed strategy equilibria.[9] The Agent's weakly dominant strategy in state *G* is to signal *G*. We thus focus on equilibria where the Agent plays this weakly dominant strategy in state *G* but randomises between a signal of *G* or *B* in state *B*. The Principal will thus know with certainty the state is *B* when a signal of *B* is observed, setting $p = 1$. For the Agent at 1*B* to randomise between a signal of *G* or *B*, the Principal at 2*G* must randomise between actions *F* and *NF* to the point of indifference between these actions. The expected value of following, *F*, which is $(70 \times q)$, must equal the expected value of not following, *NF*, which is 30, based on payoffs in Figure 1. The Principal is therefore expected to form the belief $q = 3/7$. We use Bayes' theorem to compute the probability the Agent signals *G* at node 1*B*, denoted as *x*. It is shown in Appendix A that $q = 3/7$ only if $x = 4/9$. Finally, the Agent must be indifferent between signalling *G* or *B* in order to employ such a mixed strategy. Allowing *y* to denote the probability the Principal chooses *F* at 2*G*, it is also shown in Appendix A that the Agent is indifferent between the two signals at 1*B* when $y = 1/5$.

We can characterise a mixed strategy equilibrium with (i) the Agent signalling *G* with certainty at 1*G* and with probability 4/9 at 1*B*; and (ii) the Principal choosing *F* with probability 1/5 at 2*G* and always choosing *F* at 2*B*. The Principal's beliefs are $q = 3/7$ and $p = 1$. We refer to this version of the experiment, where there is no punishment mechanism, or $Z = 0$, as the control treatment. We now turn to characterising various punishment treatments.

[8] Equilibria are expressed in terms of the strategy at nodes (1*A* 1*B*, 2*A* 2*B*) and beliefs (*q*, *p*); see Figure 1.
[9] We use the terms mixed strategy Nash Equilibrium, Bayesian mixed strategy Nash Equilibrium and Bayesian Nash Equilibrium interchangeably.

2.2 Introducing a punishment mechanism

We build on the experiment described in Section 2.1 by introducing a punishment mechanism for deception by the Agent. In the high penalty high enforcement (HPHE) treatment, a penalty of Z=50 is imposed with probability $\alpha = 0.50$, should the Agent send an untrue signal of the state of nature and the principal follows. In the high penalty low enforcement (HPLE) treatment, the same penalty of Z=50 is imposed with probability $\alpha = 0.10$. In the low penalty high enforcement treatment (LPHE), a smaller fine of Z = 5 is imposed with certainty, $\alpha = 1$.

The game tree for the HPHE and HPLE treatments are shown in Figure 2 for our control and punishment treatments are shown in Appendix B, Tables A1-A4. There are no separating or pooling equilibria for this game. As the Agent's dominant strategy at node 1*G* is to signal *G*, we continue to focus on mixed strategy equilibria where the Agent plays this dominant strategy at 1*G*. The Principal must again be indifferent between the *F* and *NF* actions at 2*G* requiring that $q = 3/7$ thus, the Principal's payoffs are unaffected by the penalty. Again, this belief is only consistent if the probability of the Agent choosing *G* at 1*B* is $x = 4/9$.

[Figure 2 around here]

Thus, the only difference between the HPHE and HPLE equilibria and the control treatment mixed strategy equilibrium is, *y*, the probability the Principal chooses *F* at 2*G*. Assuming risk neutral Agents and given the expected penalty is $P = \alpha Z$, we find $y = \frac{10}{50-\alpha Z}$, which is derived in Appendix A. For the HPHE treatment, $Z = 50$ and $\alpha = 0.50$, so we have $y = 2/5$, while in the HPLE treatment, $Z = 50$ and $\alpha = 0.10$, so $y = 2/9$.

Setting Z = 5 in Figure 1 provides the LPHE game tree with payoffs. Maintaining the assumption of Agent risk neutrality, the LPHE mixed strategy equilibrium is identical to that of the HPLE equilibrium as the expected penalty is identical. The equilibrium is characterised as (i) the Agent always signalling *G* at 1*G* and signalling *G* with probability 4/9 at 1*B*; and (ii) the Principal choosing *F* with probability 2/9 at 2*G* and always choosing *F* at 2*B*. The Principal's beliefs are $q = 3/7$ and $p = 1$.

The counterintuitive result that the introduction of penalties on Agents changes Principal behaviour but has no impact on Agent behaviour is a common feature of Bayesian mixed strategy Nash equilibria. It is due to the fact that Agent's equilibrium behaviour is determined solely by the Principal's payoffs, which in our case do not change with the introduction of penalties. At an intuitive level, we can think of the penalty having two opposing effects in this strategic setting. The first being the conventionally anticipated deterrent effect on deceptive behaviour by the agent, while the second is the reassurance provided by the penalty to the Principal that the Agent is more likely to behave honestly, thereby inducing the Agent to behave dishonestly. These opposing effects perfectly crowd out each other in our Bayesian mixed strategy Nash equilibrium, leading the penalty to have no effect on deceptive behaviour. These types of unexpected outcomes of punishments on behaviour in game theoretic settings are highlighted in Gneezy and Rustichini (2004) and in Tsebelis (1989) who concludes in a strategic setting that a penalty has no impact on crime.

2.3. Hypotheses

It is important to note that we do not discuss Agent behaviour when the true state of the world is *G*, since there is no incentive for Agents to be deceptive and send signal *B*. Further, since sending a signal *B* when the state of the world is *G* is a weakly dominated strategy for Agents, Principals should always follow a *B* signal, for this reason we also do not discuss Principal behaviour when they receive a *B* signal. However, for completeness, we discuss these more trivial cases in Appendix C.

Following the mixed strategy Nash equilibrium discussed in Section 2.2 we offer the following hypotheses.

Hypothesis A1: At node 1*B*, a large fraction (4/9) of Agents will send deceptive signals about the state of the world in all treatments.

Hypothesis A2: The introduction of a punishment mechanism will have no impact on the number of Agents sending deceptive signals at node 1*B*, with no variation across the three punishment treatments.

The mixed strategy Nash equilibrium predicts that higher expected penalties for deceptive behaviour should increase the rate of following among Principals as the probability of following is $y = \frac{10}{50-\alpha Z}$, i.e. the following probability rises with the expected penalty αZ until $\alpha Z = 40$; the principal should follow with certainty for expected penalties greater than 40, however all expected penalties in our experiments are less than 40.[10] This presents us with two hypotheses regarding Principal behaviour and beliefs when observing a signal *G*.

Hypothesis P1: Compared to the control treatment, following at node 2*G* will be higher in the punishment treatments, being higher and equal in the LPHE and HPLE treatments and highest in the HPHE treatment;

Since Principal following probabilities are based on their beliefs about Agent truthfulness:

Hypothesis P2: Principal beliefs about receiving a deceptive signal when the state is *G* will be highest with the control treatment. Beliefs will be lower and equal in the LPHE and HPLE treatments and lowest in the HPHE treatment.

Given hypotheses *P1* and *P2*, we have the following hypothesis regarding Agent beliefs about the following behaviour of Principals.

Hypothesis A3: Agent beliefs about the following rate of Principals receiving signal *G* will increase as expected penalties from deceptive behaviour increase at node 1*B*.

2.4 Risk preference Task

Given that Principals are uncertain of the truthfulness of Agent signals and two of our treatments involve uncertain punishments, risk attitudes may play an important part in the behaviour of Principals and Agents. In order to elicit risk preferences we implement a risk game introduced by Eckel and Grossman (2002) as the final task. Participants had the opportunity to select one payoff option from a number of possibilities; see Table A7 in Appendix B. All payoff options had a high and a low payoff outcome, each with equal probability. At the end of the experiment a coin was tossed to decide the payoff. Participants are attracted to higher numbered game options if they are more risk tolerant. The elicited risk attitudes are used to test the robustness of player behaviour to these otherwise unobserved risk attitudes.

[10] In the foundational theoretical literature, Becker (1968) shows that risk preferences can influence how individuals are deterred by an expected fine with differing severity and probability; see Friesen (2013) for a discussion in an experimental setting. For risk-neutral individuals, the severity and probability trade-off does not matter. For risk-averse individuals, increases in severity will have a larger effect and for risk-loving agents, increases in probability will have a larger effect. We control for risk preferences as part of our robustness checks in Sections 3.1 and 3.2.

2.5. Experimental procedures

The experiment was conducted at Monash University's Laboratory for Experimental Economics (MonLEE) using the experimental program z-Tree (Fischbacher, 2007). Participants were recruited using ORSEE (Greiner, 2015). In total we ran 8 sessions with 160 participants, of which half were randomly assigned as Agents and half as Principals. All participants played three rounds, where each round was a separate treatment. This partially within subject design was utilised to reduce the influence of individual level heterogeneity on behaviour. Each participant, in the first round played in the control condition. In the second round the participants either played the HPLE treatment (n = 84) or the HPHE treatment (n = 76) while in the final round all participants played the LPHE treatment. After completing the third round all subjects participated in the risk experiment. The points that subjects earn during the experiment are converted to Australian dollars at the end of the experiment with an exchange rate of 5 points to $1 (Australian dollar or AUD). Along with a $5(AUD) turn up fee, participants were paid for one of the three rounds, which was chosen randomly in addition to the risk experiment. Each session lasted an hour with the average participant earning $21(AUD).

All subjects received written instructions (attached in Appendix D)[11] that were read aloud to establish common knowledge. Instructions for the subsequent tasks were only provided once the previous task was completed. Understanding of the rules was assured by a set of control questions that subjects were required to complete prior to making decisions. Answers to these questions were verified and the decision-making did not commence until all subjects indicated that they understood the instructions.

All responses were solicited using the strategy method. More specifically, all Agents were asked if the actual state of the world was *G*, what state of the world would you inform Player 2? And on the same screen were asked the same question but if the state of the world was *B*. Similarly, Principals were asked for both signals *G* and *B*-- if Player 1 signalled state *G* (*B*), would you follow or not follow? We implemented a stranger matching design, so that subjects were randomly rematched every round. All participants understood that they were to be randomly matched at the start of each round. In order to avoid experiences from past rounds

[11] We provide the instructions for the control, HPLE and LPHE version of the experiment. Other treatments are available on request.

influencing behaviour, there was no feedback provided to subjects between rounds, reducing confounding impacts of beliefs being updated between rounds and driving behaviour.

In addition to actual decisions of subjects, we also collected information on beliefs about their partner's choice. In order to capture beliefs we used incentivised survey questions after subjects made their decisions in each round. Specifically, Agents were asked to predict the percentage of Principals that would follow (*F*) if they received signal *G* and if they received signal *B*. Similarly, Principals were asked to predict the percentage of the Agents that misreported the state of the world when the true state was *G* and when it was *B*; i.e. Agents sent signal *G* (*B)* when the state was *B* (*G*). Participants were paid 10 experimental dollars if their prediction was within 10% of the actual behaviour of participants in their respective sessions. We also collected information on basic demographics and socio-economic status and choice of studies at university with a post-experiment survey.

2.6 Norms based experiments

To further understand the impact of penalties on behaviour, we consider theories of behavioural norms as an alternative explanation. While the introduction of penalties may not change the economic incentives of Agents, it can change the framing of a decision task and activate or increase the focus of norms that can affect behaviour; see for example (Krupka and Weber, 2009, 2013) and (Cialdini et al., 1990). In this case, norms could lead to lower levels of deception by: (i) signalling to Agents that deceptive behaviour is undesirable; and (ii) increasing the psychological costs of deception. Conversely, the introduction of penalties may increase deception because a punishment mechanism might serve to value the norm, reducing intrinsic motivation for adherence. This will occur if the size of the penalty is smaller than the psychological cost of deviating from the norm, thus lowering the value and psychological cost of deception (Gneezy et al. 2011). This is similar to the finding of Gneezy and Rustichini (2000) who show that penalties for late pick-up in a child care centre raised the incidence of late pick-ups rather than reducing it.

We hypothesise norms may influence *Hypothesis A2***.** The punishment mechanisms in these treatments may activate (or increase the salience of) norms that reduce deceptive signalling. The simplest possibility is that all penalties (irrespective of size) activate (or increase the salience of) norms in the same way, reducing deception relative to the control treatment, with no variation in deceptive behaviour due to differences in penalty magnitudes and probabilities of enforcement across punishment treatments.

However, the type of punishment mechanism may provide more nuanced information to Agents, with the salience of norms depending on the size or enforcement of the penalty. For example, expected penalties may affect salience monotonically, i.e. the higher the expected penalty the higher the salience of the norm. If this is the case, deception will be greatest in the control treatment, followed by the LPHE and HPLE treatments (due to identical expected penalties), with the lowest levels of deception in the HPHE treatment. Along similar lines, if the probability of the enforcement of a penalty determines norm salience, we expect deception to be highest in the control treatment and to be progressively lower in the HPLE, HPHE and LPHE treatments respectively.

To test the impact of norms, we collect data from a second experiment. This experiment is similar to the first but contains two new treatments where we alter the payoffs of our original control treatment. We implemented 4 sessions resulting in 66 participants. In the first additional treatment, which we label Min-5, we modify the original control treatment by reducing Agent payoffs from deception and being followed by 5 experimental dollars. The expected payoffs from this treatment are identical to LPHE and HPLE treatments, The key difference is that like the control, subjects are not informed that deceiving is accompanied by a penalty (see Appendix B Table A5). We can thus compare this treatment to LPHE and HPLE treatments to identify if using a penalty frame changes behaviour and to identify the impact of the size of a penalty on norms made salient through the frame. The second treatment, Min-25 is like Min-5 except we reduce Agents payoff from deceiving and being followed by 25 experimental dollars (see Appendix B Table A6). Expected payoffs are equivalent to the HPHE treatment with the exception that we do not use the term penalty. We compare this treatment to HPHE to identify the influence of norms activated through the penalty frame.

Like our first experiment, all participants played three different treatments. Subjects participated in Min-5 and Min-25 followed by LPHE in the third round. The rationale for including LPHE in the final round is to check the robustness of the effect of our experiment 1 LPHE treatment on behaviour. In the first experiment, the LPHE treatment was always played in the 3rd round and was preceded by a treatment where high penalties were introduced. If norms were made salient by these high penalty treatments, then it might have spilled over to the LPHE treatment. In experiment 2, as LPHE treatment is not preceded by a treatment where the penalty is made salient, this spill over should not exist. Finally, all other protocols for the second experiment was the same as Experiment 1.

3. Results

In this section, we begin by examining the signal transmitted by the Agent and Agent's beliefs about the behaviour of Principals. Following this we investigate the response of Principals to signals received and their beliefs about the proportion of agents that misreport.

3.1 The effects of punishment mechanisms on Agent behaviour.

The behaviour of Agents across all treatments is presented in Table 1 and the difference estimates across treatments are presented in Panel A, Table 2.

When the state of the world is *B*, the mixed strategy Nash equilibrium predicts that 4/9 of signals will be deceptive in the control treatment. In this treatment we find that Agent deception is much higher at 59% (p = 0.01; Wilcoxon Signed-rank). Deception decreases across both high penalty treatments. Agents send deceptive signals 24% of the time in the HPHE, and 29% of the time in the HPLE treatment. In both cases, deception rates are statistically different from the control treatment (p < 0.01; Mann-Whitney).[12] There is no statistically significant difference in deception between the control and LPHE treatments (p = 0.27). It follows that the HPLE and HPHE treatments have significantly lower rates of deception than the LPHE treatment (p < 0.05).

[Table 1 and Table 2 around here]

Given the expected penalties for the LPHE and HPLE treatments are identical, these results imply that raising the severity of the penalty has a larger deterrent effect on deceptive behaviour than an equivalent increase in the probability of enforcement. This is further supported by the fact that the rates of deception are not statistically different between HPHE and HPLE treatments (p = 0.8), implying that in the presence of high penalties, increasing enforcement probability, from 0.1 to 0.5, has little impact on deceptive behaviour. These findings reject both *Hypothesis A1* and *A2* and are summarised in the following results.

[12] Unless stated otherwise, we report p-values for two-tailed Mann-Whitney Wilcoxon tests. In Appendix D we provide p-values for two tailed t-tests as a robustness check.

Result A1: *When the state of the world is Bad, with no punishment mechanism in place, a majority (more than the predicted 4/9) of Agents send deceptive signals about the state of the world. This deception is not constant across treatments.*

Result A2: *The punishment treatments reduce the rate of deception conditional on the penalty being of large magnitude. A higher probability of punishment in the presence of large penalties has no additional deterrent effect.*

Agent beliefs about Principal following rates are shown also documented in Table 1 and differences in beliefs across treatments are shown in Panel A, Table 2. On average, Agents believe that 52% of Principals will follow a signal *G* in the control treatment. Under the HPHE treatment, Agent beliefs increase to 60%, which is significantly higher than the control (p = 0.06). There are no statistical differences between Agent beliefs in the control, HPLE and LPHE treatments ($p > 0.33$ for all pairwise tests). A Kruskal-Wallis test of equality of means of beliefs across the treatments (experiment 1) is also statistically insignificant ($p = 0.25$). We thus do not find support for *Hypothesis A3*.

Result A3: *Agent beliefs of Principal following rates of a G signal are unaffected by our punishment treatments when the expected penalty is low.*

Agent beliefs are quite close to the actual following behaviour of Principals in the control treatment and LPHE treatments, where the following rate is 49% ($p = 0.7$, Wilcoxon Signed-rank) and 56% ($p = 0.21$; Wilcoxon Signed-rank) respectively; Panel A, Table 2. In the presence of large penalties, Agents systematically underestimate Principal following rates. In the HPHE treatment the following rate is 71%, which is significantly greater than the 60% predicted by Agents ($p < 0.01$; Wilcoxon Signed-rank). Similarly, the following rate for the HPLE treatment is 69%, which also exceeds Agents' beliefs of 55% ($p < 0.01$; Wilcoxon Signed-rank). These findings, which relate Agent beliefs to Principal following rates when sending a signal *G* are summarised in the following result.

Result A4: *When facing no punishment or punishments of small magnitude, Agents accurately predict the behaviour of Principals facing a signal of G. In the presence of large penalties, Agents underestimate the rate of Principal following of a signal G.*

Behaviour displayed in the experiment may be driven by an individual's risk preferences. In order to isolate the treatment effect and to confirm the robustness of our previous results we undertake an econometric analysis, controlling for risk preferences elicited from the Eckel and

Grossman (2002) risk task, along with other individual characteristics. In columns (1) and (2) of Table 3, we present OLS estimates for a model of Agent behaviour in states *G* and *B* respectively. The explanatory variables include dummies for treatments and a dummy if the subject is risk averse.[13] We also control for individual heterogeneity. These estimates show that after controlling for risk and individual characteristics our results are robust and consistent with those found in our non-parametric analysis.

[Table 3 around here]

3.2 Principal responses to punishment mechanisms

Principal behaviour is reported in Table 1 and tests for differences in behaviour across treatments are reported in Panel A, Table 2. In the case of signal *G*, the following rate was 49% in the control. This is significantly higher than the mixed strategy Nash equilibrium, which predicts 20% following ($p < 0.01$, Wilcoxon Signed-rank). In the HPHE and HPLE treatments, following rates increase further to 71% and 69% respectively, both of which are statistically significantly higher than in the control ($p < 0.05$). The following rate in the LPHE treatment is 7 percentage points higher than in the control but this difference is not statistically significant ($p > 0.60$). Our results suggest that the introduction of penalties only reassures principals if they are large, leading them to increase following behaviour. The marginal impact of increasing monitoring in the presence of severe penalties appears negligible when comparing the HPHE and HPLE treatments. These findings allow us to reject *Hypothesis P1* and are summarised as follows.

Result P1: *When Principals receive signal G, following rates depend on the severity of the penalty. Following rates are higher when penalties are high and do not vary with the probability of punishment. Low but certain penalties do not induce higher following rates.*

Turning to Principal's beliefs about deceptive signals when the actual state of the world is *B*. We find that in the control treatment, Principals believe that 60% of signals sent by Agents when the state of the world is *B* are deceptive; that is, Principals believe Agents send signal *G* when the true state is *B*. In both the HPHE and HPLE treatments, Principals believe Agent deception decreases to less than 40%, which was statistically different from both the control and LPHE treatments ($p < 0.01$; Mann-Whitney, two tailed). While the average belief about

[13] Following Eckel and Grossman (2002), a subject is risk averse if they select an option less than 5; see Appendix B, Table A7. Since individuals play the control, HPHE or HPLE, and LPHE treatments, we are analysing the impact of risk preferences within each of the treatments and not across.

misreporting is lower in the LPHE (53.1%) compared to the control treatment, the difference is not statistically significant ($p > 0.12$; Mann-Whitney, two-tailed). There is also no statistical difference in Principal beliefs across HPHE and HPLE treatments ($p > 0.62$; Mann-Whitney, two-tailed). We thus reject *Hypothesis P2*, summarizing our results on Principal beliefs in the following result.

Result P2: *Principal beliefs about deception when the state is B depend on the severity of the punishment. A higher penalty decreases Principal's beliefs that Agents will send a deceptive signal, however, a higher probability of punishment has little effect on beliefs.*

Principal beliefs about Agent actions are quite accurate in the absence of large penalties. In the control treatment and LPHE, the actual rate of deception was 59% and 50% while the corresponding beliefs were 60% ($p > 0.4$, Wilcoxon Signed-rank, two tailed) and 53.1% ($p > 0.2$, Wilcoxon Signed-rank, two tailed) respectively. However, when the penalties imposed were large, Principals systematically over-estimated the probability of receiving deceptive signals by over 10 percent ($p < 0.02$, Wilcoxon Signed-Rank, two tailed, for both HPHE and HPLE treatments). These findings compare Principal beliefs to Agent deception rates, and are summarised in the following result.

Result P3: *When Agents face no punishment or punishment of small magnitudes, Principals accurately predict the rate of Agent deception. In the presence of large penalties, Principals overestimate the rate of Agent deception.*

As with Agent behaviour, we test the robustness of Principal behaviour to risk preferences and individual characteristics by estimating OLS regressions of Principal following behaviour. Results are presented for signals *G* and *B* respectively in columns (3) and (4) of Table 3. Our main results about Principal following behaviour are robust to the inclusion of individual level differences.

3.3 Discussion

In this subsection we summarise our main findings and relate them to the predictions from the Bayesian Nash equilibrium we have characterised. When the true state of the world is *Bad*, we find the magnitude of penalties matters more than the magnitude of expected penalties for Agent signalling (*Result A2*). Agents reduce their deceptive behaviour in the presence of large potential penalties, irrespective of the probability of enforcement. This result is inconsistent with the predictions of constant rates of deception across all treatments. Key to explaining this inconsistency is understanding the trade-offs that underlie the Bayesian Nash equilibrium. Intuitively, the drop in expected monetary benefits from deception should be counter balanced by changes in Agent beliefs about Principal following. A lack of evidence of the latter suggests that the drop in deception can be explained by changes in the expected penalty size. This is the prominent mechanism outlined in the traditional deterrence model.

We find some support for this. The HPHE treatment, which has the highest expected penalty of 25 experimental dollars (ED), elicits significantly less deceptive behaviour than the control and LPHE treatments with expected penalties of 0 ED and 5 ED respectively. However, comparing the HPLE treatment with expected penalty of 5 ED to the HPHE and LPHE treatments shows different deception rates with identical expected penalties (HPLE vs LPHE) and identical deception rates with different expected penalties (HPHE vs HPLE). As a consequence we conclude that deterrence models that focus on expected penalties cannot fully explain the effect of penalties in our Principal-Agent setting. We offer an alternative explanation in the next section.

Turning to Principals, results in Panel A, Table 2 show Principals faced with a potentially deceptive *Good* signal respond to large penalties. A significant difference in following behaviour is observed between the cases where fines are high (HPHE and HPLE) and the control and LPHE treatments. According to the mixed strategy Nash equilibrium, following should increase with expected penalty and we only find partial evidence for this – Principal following increases between the control and HPHE and HPLE. However, we also find that following is the same in (i) the HPHE and HPLE treatments; and (ii) the LPHE and control treatments. However, it is different across LPHE and HPLE, implying the model's predictions are not in line with observed behaviour. In addition, we find that Principals believe that Agents will send fewer deceptive signals in the high penalty, HPHE and HPLE, treatments. This pattern of beliefs is consistent with the actual actions taken by Agents in that only high penalties are believed to be effective and is consistent with the norms explanation of behaviour offered below.

3.4 Alternative explanations

In this subsection we investigate an alternative norms based explanation outlined in Section 2.6. Norms provide a set of rules that govern individual behaviour that can at times conflict with incentives based on pecuniary interests. This is particularly true in the case of deceptive behaviour where pecuniary incentives may encourage deception, while norms often encourage honest behaviour due to the high social and moral cost of lying (Gibson *et al*, 2013). The impact of norms on behaviour is pertinent when decision tasks are framed in a way to induce moral behaviour by increasing the salience of these norms.

We combine data from our norms based treatments with our data from experiment 1. Table 1 and Panel B, Table 2 present results for the norm treatments Min-5 and Min-25, corresponding to LPHE/HPLE and HPHE expected payoffs respectively. Several important results are evident. First, when the penalties are large, framing punishment as a penalty appears to have an additional deterrence effect. We find that the HPLE treatment has a 20 percentage point lower rate of deception compared to Min-5 despite having the same expected payoffs (p = 0.08). We also find a 10 percentage point difference between HPHE and Min-25, although it is not statistically significant ($p > 0.4$). If we were to decompose the total effect of penalties in the high penalty treatments between: i) the effect of deterrence due to pay-off decrease and ii) norm salience, we find that 2/3 and ¼ of the total effect in the HPLE and HPHE treatments respectively, can be explained by an increase in the salience of norms (Figure 3).[14]

[Figure 3 around here]

Second, in the Min-5 treatment, the Agent deception rate is identical to that in the low penalty treatment (LPHE) ($p > 0.99$). This implies that when the penalty is small, making norms salient has little additional deterrence effect. Note that the fine size in the LPHE treatment is 12.5% of the benefit of deception. Gneezy and Rustichini (2000) highlight the pricing effect of fines: in this context this implies that the guilt from deception is priced and hence reduces the intrinsic motivation to send truthful signals. These results imply that the activation of norm saliency among Agents is not uniform across penalties but rather depends on the size of the penalty.

Similar to the behaviour of Agents, Principals are also influenced by the saliency of norms, but again only when penalties are large. We find that following rate is 30 percentage points higher in HPLE compared to Min-5 (p = 0.01) and 24 percentage points higher in HPHE compared to

[14] The total effect for HPLE (HPHE) is calculated by comparing the means between control and HPLE (HPHE) treatments. The effect of the payoff is calculated by comparing the means between control and Min-5 (Min-25). The difference between the total effect and the effect of the payoff is the effect based on norm saliency.

Min-25 (p = 0.06). We find that 2/3 of the total effect in Principal behaviour in the HPLE treatment is driven by norm saliency. For HPHE, the entire deterrence effect is driven by norms. When the penalty is smaller, there is no additional effect of introducing a punishment mechanism: we find that the following rate is not statistically different between LPHE and Min-5 (p > 0.69). Principal behaviour is consistent with their beliefs. On average, Principals believe there will be lower rates of deception in the high penalty treatments: in HPHE relative to Min 25 (p = 0.00) and HPLE relative to Min-5 (p = 0.00). When penalties are small, beliefs about deception are not different between LPHE and Min-5 (p = 0.11). These results imply that the use of the term penalty, particularly when they penalties are large, increases the saliency of norms among Principals and induces higher following rates and beliefs that Agents will be honest.

Since subjects in experiment 1 participate in multiple treatments prior to the LPHE treatment, norms made salient in earlier treatments may influence LPHE behaviour. The design of experiment 2, where penalty frames were not used prior to LPHE, allows us to test this impact. Comparing behaviour in LPHE experiment 1 and LPHE experiment 2 we find that for both Principals and Agents, the behaviour across the two experiments are similar (p > 0.88 for Agents, p > 0.29 for Principals, see Panel B, Table 2). This suggests that framing punishment as a penalty has a negligible spill over effects across treatments.

To summarise, our results from experiment 2 suggest the increased psychological cost of deception induced by norm saliency plays an important role in the deterrence of deceptive behaviour and its perception. Clearly framing punishments as a penalty increases honesty but only when the penalty is appropriately large.

4. Conclusion

This paper examines deceptive behaviour in a stylised expert advisor-client relationship. Such relationships are common and include financial advisor-investor, doctor-patient or mechanic-car owner. The inherent information asymmetry in these relationships provides scope for deceptive behaviour on the part of agents and as a consequence, regulatory schemes exist to deter and punish such behaviour. Our findings suggest that large penalties, even with low levels of monitoring are more likely than small certain penalties to be effective at deterring opportunistic behaviour and perceptions among clients of such behaviour. This deterrence is not only driven by the potential losses associated with large fines when they are realised, but is also determined by increases in the saliency of norms due to framing punishment as a penalty. We find that a penalty frame produces an additional deterrence effect, but only when penalties are large. When penalties are smaller, this additional effect is not observed.

Our results suggest that policy makers can utilise penalties to reduce deceptive behaviour. In our experiment we have demonstrated that large fines, in contrast to small fines, will achieve this. However, if large fines are accompanied with strong wording to emphasize (frame) the social and moral costs of deceptive actions, this deterrence effect should be stronger. While our experiment assumes that the criminal justice system is perfect in its enforcement, this might not be the case. In an imperfect system with reasonable doubt tests, Andreoni (1991) provide a theoretical model which show that large absolute fines can lead to increases in criminal behaviour if jurors are averse to convict. It would be worthwhile to examine whether behaviour is in line with the theoretical prediction in Andreoni, (1991) in light of the strong deterrence effect presented in our paper. Finally, while large penalties deter deception, we do not test whether the ensuing lower rates of deception will persist once these fines are removed/reduced through the lowering of monitoring. This is an issue for future research.

References

Abbink, K., Irlenbusch, B., and Renner, E. (2002). An experimental bribery game. *Journal of Law, economics, and organization*, *18*(2), 428-454.

Anderson, L.R. and S.L. Stafford, (2003), "Punishment in a regulatory setting: Experimental evidence from a VCM", *Journal of Regulatory Economics* 24(1), 91-110.

Andreoni, J. (1991): Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime? *RAND Journal of Economics*, 22(3), 385-395

Becker, G., (1968), "Crime and Punishment: An Economic Approach", *Journal of Political Economy* 76(2), 169-217.

Bénabou, R., and J. Tirole (2006), "Incentives and pro-social behaviour." *American Economic Review* 96(5), 1652-1678.

Block, M.K. and V.E. Gerety, (1995), "Some experimental evidence on differences between student and prisoner reactions to monetary penalties and risk", *The Journal of Legal Studies* 24(1), 123-138.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, *58*(6), 1015.

DeAngelo, G., and Charness, G. (2012). Deterrence, expected cost, uncertainty and voting: Experimental evidence. *Journal of Risk and Uncertainty*, *44*(1), 73-100.

Department of Justice, (2013), "Bank of America to Pay $16.65 Billion in Historic Justice Department Settlement for Financial Fraud Leading up to and During the Financial Crisis", Office of Public Affairs, 21 August 2014. https://www.justice.gov/opa/pr/bank-america-pay-1665-billion-historic-justice-department-settlement-financial-fraud-leading , accessed 5 September 2016.

Department of Justice, (2014), "Justice Department, Federal and State Partners Secure Record $13 Billion Global Settlement with JPMorgan for Misleading Investors About Securities Containing Toxic Mortgages", Office of Public Affairs, 19 November 2013. http://www.justice.gov/opa/pr/justice-department-federal-and-state-partners-secure-record-13-billion-global-settlement , accessed 7 December 2015.

Eckel, C.C., Grossman, P.J., (2002), "Sex differences and statistical stereotyping in attitudes toward financial risk", *Evolution and Human Behaviour* 23(4), 281–295.

Elster, J (2009), Social norms and the explanation of behaviour. Hedström P, Bearman P, eds. The Oxford Handbook of Analytical Sociology (Oxford University Press, Oxford, UK), 195–217.

Fehr, E. and S. Gachter, (2000), "Cooperation and Punishment in Public Goods Experiments", *American Economic Review* 90(4), 980-994.

Fischbacher, U., (2007), “z-Tree: Zurich toolbox for ready-made economic experiments”, *Experimental Economics* 10(2), 171-178.

Friesen, L., (2012), “Certainty of Punishment versus Severity of Punishment: An Experimental Investigation”, *Southern Economic Journal* 79(2), 399-421.

Gibson, R., Tanner, C. and A.F. Wagner, (2013), “Preferences for truthfulness: Heterogeneity among and within individuals”, *American Economic Review* 103(1), 532-548.

Greiner, B., (2015), “Subject pool recruitment procedures: organizing experiments with ORSEE”, *Journal of the Economic Science Association* 1(1), 114-125.

Gneezy, U., (2005), “Deception: The role of consequences”, *American Economic Review* 95(1), 384-394.

Gneezy, U., and A. Rustichini, (2000), “A fine is a price”, *Journal of Legal Studies* XXIX, 1-17.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behaviour. *The Journal of Economic Perspectives*, *25*(4), 191-209.

Gneezy, U., and A. Rustichini, (2004), “Incentives, punishment and behaviour”, In: Camerer, C.F., Lowenstein, G., Rabin, M. (Eds.), *Advances in Behavioural Economics,* Princeton University Press: Princeton and Oxford, 572-589.

Harbaugh, W. T., Mocan, N., and Visser, M. S. (2013). Theft and deterrence. *Journal of Labor Research*, *34*(4), 389-407.

Khadjavi, M. (2015). On the interaction of deterrence and emotions. *Journal of Law, Economics, and Organization*, *31*(2), 287-319.

Krupka, E., and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behaviour. *Journal of Economic psychology*, *30*(3), 307-320.

Krupka, E. L., and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, *11*(3), 495-524.

Leibbrandt, A., and López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behaviour and Organization*, *84*(3), 753-766.

Nagin, D. S., and Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology*, *41*(1), 167-194.

Peeters, R., Vorsatz, M. and M. Walzl, (2013), “Truth, trust and sanctions: On institutional selection in sender-receiver games”, *Scandinavian Journal of Economics* 115(2), 508-548.

Polinsky, A. M., and S. Shavell, (2007), “The theory of public enforcement of law”, In: Polinsky, A. M., Shavell, S. (Eds.), *Handbook of Law and Economics*, Elsevier-North Holland, Vol. 1, 403-454.

Schwartz, A. L., Landon, B. E., Elshaug, A. G., Chernew, M. E., and McWilliams, J. M. (2014). Measuring low-value care in Medicare. *Journal of the American Medical Association: Internal medicine, 174*(7), 1067-1076.

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal*, *119*(534), 47-60.

Tsebellis, G., (1989), “The abuse of probability in political analysis: The Robinson Crusoe fallacy”, *The American Political Science Review* 83(1), 77-91.

Tryan, J.-R. and L.P. Feld, (2006), “Achieving Compliance when Legal Sanctions are Non-deterrent”, *Scandinavian Journal of Economics* 108(1), 135-156.

Table 1: Mean Behaviour and Beliefs of Agents and Principals across treatments in Experiment 1 and 2.

	Agent		Principal	
Description	Deception at Bad State	Belief about Following of Good Signal	Follow of Good Signal	Belief about Deception when state is Bad
Control	58.8	51.8	48.8	51.8
HPHE	23.7	59.6	71.1	59.6
HPLE	28.6	55.4	69.0	55.4
LPHE_1	50.0	53.3	43.8	53.3
Min_5	48.5	61.3	42.4	61.3
Min_25	33.0	59.2	48.4	59.2
LPHE_2	48.5	55.9	33.0	55.9

	Agent		Principal	
	Deception at Good State	Belief about following of Bad Signal	Follow of Bad Signal	Belief about Misreport when state is Good
Control	6.3	62.5	78.8	33.7
HPHE	0.0	68.0	92.1	18.9
HPLE	2.4	65.7	71.4	26.6
LPHE_1	5.0	68.3	71.3	22.3
Min_5	3.0	64.9	84.8	22.5
Min_25	0.0	63.5	75.7	19.1
LPHE_2	6.1	65.0	84.8	15.2

Notes:LPHE_1 and LPHE_2 correspond to results from the LPHE treatment in Experiment 1 and 2 respectively.

Table 2: Treatment differences for behaviour and beliefs of Agents when in the *Bad* state and Principals upon receiving *Good* signal

Treatment Differences	Agent				Principal			
	Deception at Bad State		Belief about Following of Good Signal		Follow of Good Signal		Belief about Deception when state is Bad	
	Difference (1)	*p-value* (2)	Difference (3)	*p-value* (4)	Difference (5)	*p-value* (6)	Difference (7)	*p-value* (8)
Panel A: Experiment 1								
Control - HPHE	35.1	0.00***	-7.8	0.06*	-22.3	0.02**	-7.8	0.00***
Control - HPLE	30.2	0.00***	-3.5	0.33	-20.3	0.03**	-3.5	0.00***
Control - LPHE_1	8.8	0.27	-1.5	0.46	5.0	0.53	-1.5	0.11
HPHE- HPLE	-4.9	0.62	4.2	0.32	2.0	0.84	4.2	0.62
HPHE -LPHE_1	-26.3	0.00**	6.3	0.13	27.3	0.00***	6.3	0.00***
HPLE - LPHE_1	-21.4	0.02**	2.0	0.77	25.3	0.00***	2.0	0.01**
Panel B: Experiment 2								
Control - Min_5	10.3	0.32	-9.5	0.03**	-6.4	0.33	-3.7	0.26
Control - Min_25	25.8	0.01**	-7.4	0.11	-0.4	0.87	5.4	0.63
LPHE_1 - Min_5	1.5	0.88	-8.0	0.07*	-1.4	0.62	-10.4	0.11
HPLE- Min_5	-19.9	0.08*	-6.0	0.2	-26.6	0.03**	-24.1	0.00**
HPHE - Min_25	-9.3	0.37	0.4	0.96	-22.7	0.05*	-17.2	0.00**
LPHE_1 - LPHE_2	1.5	0.88	-2.6	0.07**	-10.8	0.29	1.6	0.87
Control - LPHE_2	10.3	0.96	-4.1	0.29	15.8	0.13	8.3	0.18

Notes: All p-values based on Mann-Whitney-Wilcoxon test; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 3: Models of Agent and Principal behaviour in both the *Good* (*G*) and *Bad* (*B*) states.

	(1) Truth, State *G*	(2) Truth, State *B*	(3) Follow Signal *G*	(4) Follow Signal *B*
Panel A: Coefficient Estimates				
HPHE	0.046	0.346***	0.225**	0.134**
	(0.033)	(0.089)	(0.091)	(0.066)
HPLE	0.054	0.306***	0.201**	-0.073
	(0.040)	(0.082)	(0.088)	(0.077)
LPHE	0.013	0.088	-0.050	-0.075
	(0.039)	(0.074)	(0.064)	(0.058)
Risk Averse Dummy	-0.051	0.116	0.036	-0.088
	(0.044)	(0.086)	(0.136)	(0.097)
Postgrad Dummy	-0.060	0.134	-0.151	0.156
	(0.038)	(0.092)	(0.129)	(0.097)
Female Dummy	0.053	0.122	0.155	-0.041
	(0.038)	(0.079)	(0.095)	(0.078)
Trust in Stranger Dummy	-0.036	0.005	-0.077	-0.000
	(0.032)	(0.080)	(0.097)	(0.080)
Father Completed University Dummy	-0.019	0.064	0.064	0.149
	(0.031)	(0.072)	(0.129)	(0.090)
Mother Completed University Dummy	-0.045	-0.032	0.070	-0.207**
	(0.037)	(0.066)	(0.116)	(0.091)
Low Income Dummy	-0.085**	-0.248**	-0.226*	0.085
	(0.039)	(0.095)	(0.126)	(0.107)
Constant	1.147***	0.218	0.579***	0.668***
	(0.087)	(0.164)	(0.174)	(0.168)
Panel B: Difference Estimates				
Control – HPHE	-0.05	-0.34***	-0.225**	-0.13**
Control – HPLE	-0.05	-0.31***	-0.20**	0.07
Control – LPHE	-0.01	-0.08	0.05	0.08
LPHE – HPHE	-0.03	-0.26***	-0.275***	-0.20***
HPHE – HPLE	-0.008	0.04	0.02	0.20**
LPHE – HPLE	-0.04	-0.22***	-0.25***	-0.001
Observations	240	240	240	240
R^2	0.087	0.215	0.106	0.11

Notes: Estimates from OLS regressions with standard errors, clustered at the individual level, in parentheses.

Dependent Variable: (1) Agent sending true message if state is *G*; (2) Agent sending true message if state is *B*; (3) Principal following signal *G*; (4) Principal following signal *B*.

Panel A presents coefficients of OLS regressions. All regressions included session dummies. Panel B uses post-estimation linear combination tests of hypotheses using *t*-tests.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

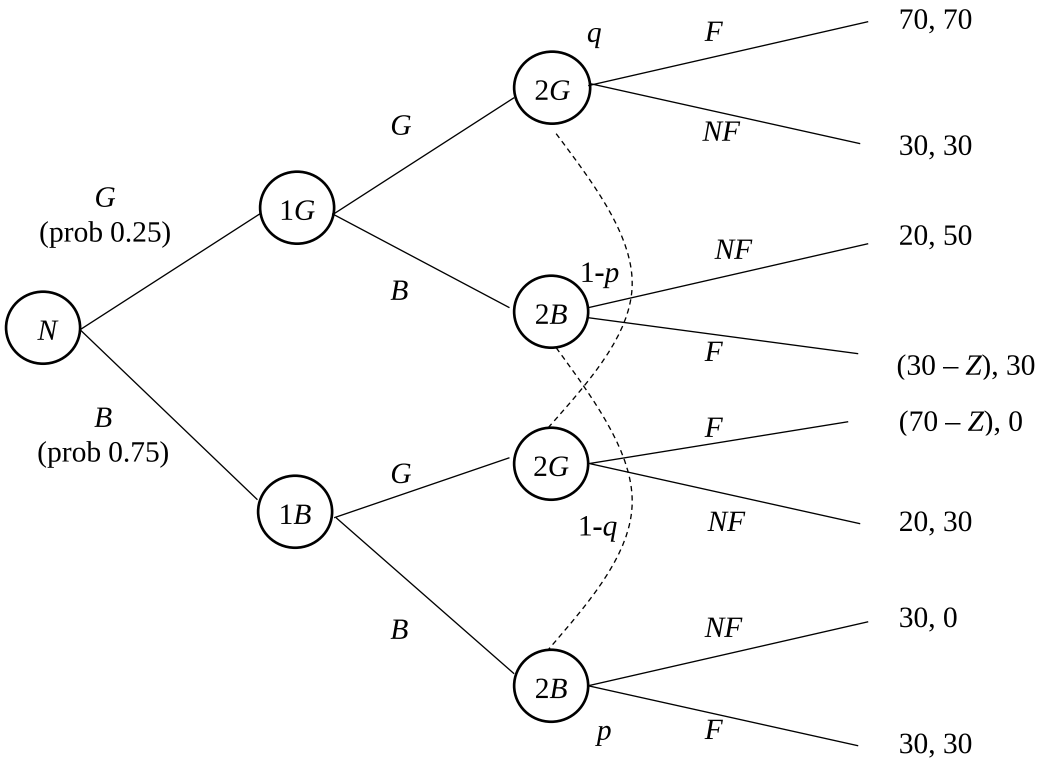

Figure 1: Game tree for the (i) control treatment with no punishment (Z = 0); and (ii) low penalty, high enforcement (LPHE) treatment which has a penalty Z = 5 with certainty. In this Figure *B (G)* refers to the *Good* (*Bad*) state of the world, *F (NF)* occurs when a Principal decides to follow (not follow) a signal, Z refers to the penalty, p is the likelihood the state of nature is bad and q is the likelihood the state of nature is good.

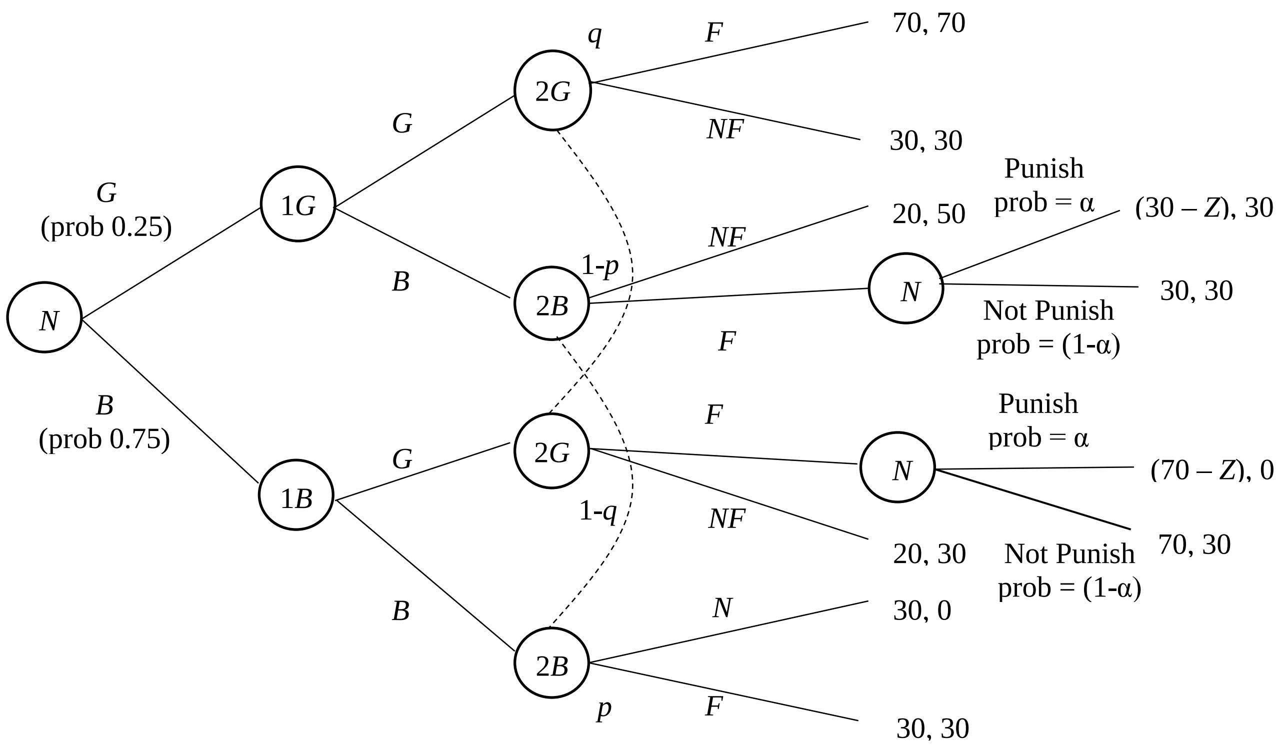

Figure 2: Game tree for uncertain punishment treatments. The high penalty, high enforcement (HPHE) treatment assumes a penalty $Z = 50$ and an enforcement probability $\alpha = 0.5$. The high penalty, low enforcement (HPLE) treatment assumes the same penalty with an enforcement probability $\alpha = 0.1$. In this Figure *B (G)* refers to the *Good* (*Bad*) state of the world, *F (NF)* occurs when a Principal decides to follow (not follow) a signal, *Z* refers to the penalty, *p* is the likelihood the state of nature is bad and *q* is the likelihood the state of nature is good.

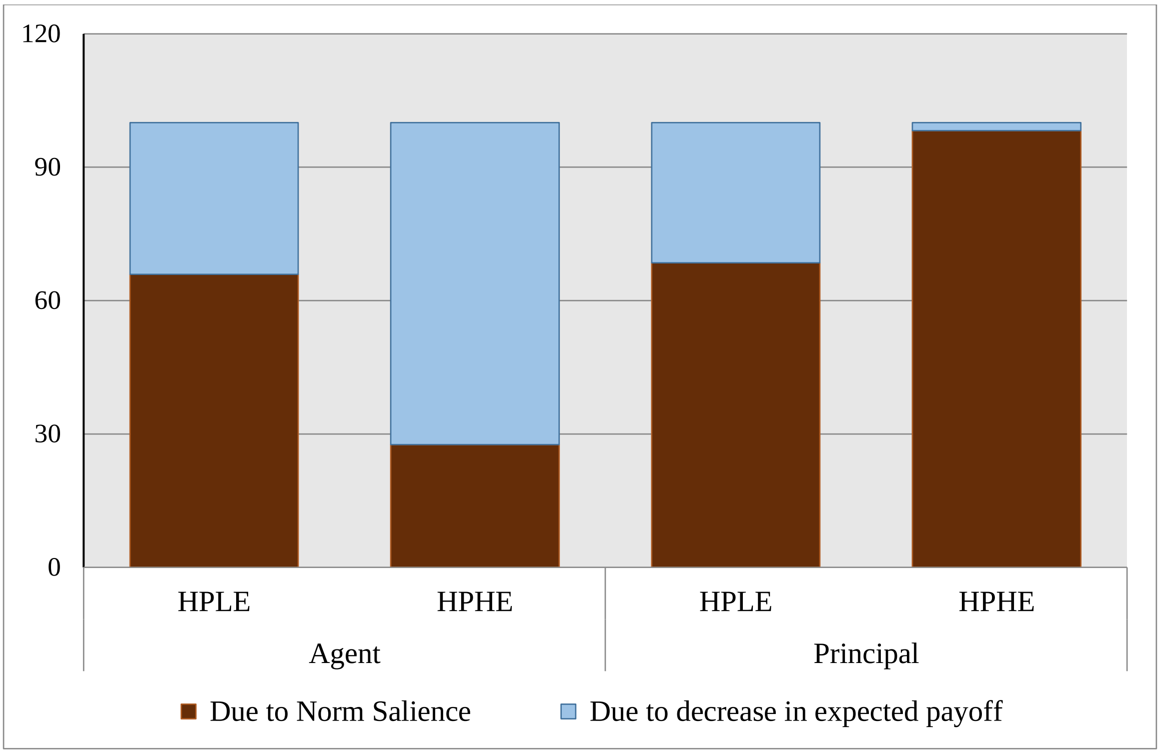

Figure 3: Decomposition of the total deterrence effect due to norm salience and decreases in expected payoff for Agents and Principals.

Appendix A: Computations for Bayesian Nash equilibrium

1. Mixed Strategy Equilibrium for Control Treatment

As the Agent's weakly dominant strategy is to choose *G* at 1*G* and the Principal knows that a signal of *G* implies $P(\text{State } G|\text{signal } G) = 3/7$, we obtain the probability the Agent sends a signal *G* at node 1*B*, denoted as *x*, by applying Bayes' theorem as:

$$P(\text{State } G|\text{Signal } G) = \frac{P(\text{Signal } G|\text{State } G)P(\text{State } G)}{P(\text{Signal } G|\text{State } G)P(\text{State } G) + P(\text{Signal } G|\text{State } B)P(\text{State } B)}$$

$$\Rightarrow \frac{3}{7} = \frac{\frac{1}{4}}{\frac{1}{4} + x\frac{3}{4}}$$

$$\Rightarrow x = \frac{4}{3}\left(\frac{7}{12} - \frac{3}{12}\right) = \frac{4}{9}$$

The Agent's expected payoff for signalling *G* at 1*B* is $70y + 20(1 - y)$. Signalling *B* ensures a payoff of 30. Thus we have:

$$70y + 20(1 - y) = 30$$

$$\Rightarrow y = \frac{1}{5}$$

2. Mixed Strategy Equilibirum when Expected value of punishment is αZ

An Agent's expected payoff for signalling *G* at 1*B* is now $(70 - \alpha Z)y + 20(1 - y)$, while signalling *B* still ensures payoff of 30.

$$(70 - \alpha Z)y + 20(1 - y) = 30$$

$$\Rightarrow y = \frac{10}{50 - \alpha Z}$$

Appendix B: Payoff Tables

Table A1: Payoff from the Control Treatment

State of the World	P1 Signal	P2 Choice	Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70
		NF	30	30
	B	NF	20	50
		F	30	30
B (Probability: 0.75)	G	F	70	0
		NF	20	30
	B	NF	30	0
		F	30	30

Table A2: Payoff from High Penalty Low Enforcement Treatment (HPLE)

State of the World	P1 Chooses to Message	P2 Chooses to pick	Payoff if not punished		Payoff if punished	
			Payoff P1	Payoff P2	Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70	70	70
		NF	30	30	30	30
	B	NF	20	50	20	50
		F	30	30	-20	30
B (Probability: 0.75)	G	F	70	0	20	0
		NF	20	30	20	30
	B	NF	30	0	30	0
		F	30	30	30	30

Note: 10% probability of being punished

Table A3: Payoff from High Penalty and High Enforcement Treatment (HPHE)

State of the World	P1 Chooses to Message	P2 Chooses to pick	Payoff if not punished		Payoff if punished	
			Payoff P1	Payoff P2	Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70	70	70
		NF	30	30	30	30
	B	NF	20	50	20	50
		F	30	30	-20	30
B (Probability: 0.75)	G	F	70	0	20	0
		NF	20	30	20	30
	B	NF	30	0	30	0
		F	30	30	30	30

Note: 50% probability of being punished

Table A4: Payoff from Low Penalty High Enforcement Treatment (LPHE)

State of the World	P1 Chooses to Message	P2 Chooses to pick	Payoff	
			Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70
		NF	30	30
	B	NF	20	50
		F	25	30
B (Probability: 0.75)	G	F	65	0
		NF	20	30
	B	NF	30	0
		F	30	30

Note: 100% probability of being punished

Table A5: Payoff from Min-5 Treatment

State of the World	P1 Signal	P2 Choice	Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70
		NF	30	30
	B	NF	20	50
		F	25	30
B (Probability: 0.75)	G	F	65	0
		NF	20	30
	B	NF	30	0
		F	30	30

Table A6: Payoff from Min-25 Treatment

State of the World	P1 Signal	P2 Choice	Payoff P1	Payoff P2
G (Probability: 0.25)	G	F	70	70
		NF	30	30
	B	NF	20	50
		F	5	30
B (Probability: 0.75)	G	F	45	0
		NF	20	30
	B	NF	30	0
		F	30	30

Table A7: Payoff Possibilities from the Risk Game; payoffs listed in experimental dollars.

Option No.	Heads	Tails
1	ED 28	ED 28
2	ED 24	ED 36
3	ED 20	ED 44
4	ED 16	ED 52
5	ED 12	ED 60
6	ED 2	ED 70

Appendix C: Discussion of Agent behaviour when the state is G and Principal behaviour when they receive a signal *B*

Results for Agent behaviour when the state of the world is *G* are presented in Table 1 and Table A8. As expected we find that deception is very low in state *G*, with very little difference across treatments. Overall, 230 out of 240 signals sent when the state of the world was *G* were true signals. This result is expected as deception is a weakly dominated strategy in this state.

Turning next to Agent beliefs about the behaviour of Principals when they receive a B signal we find that Agent's on average predicted that only 66% of Principals would follow a signal *B*. There is very little impact of our treatments on Agent beliefs that a Principal will follow a Signal *B* (p = 0.29, Kruskal-Wallis, two sided). Given the incentives, we would expect to see higher following rates among Principals and similarly high beliefs among Agents that Principals will follow such a signal. In the control, HPHE and HPLE treatments, beliefs about Principal following are always below the actual Principal's following rate ($p \leq 0.02$, Wilcoxon signed rank). However, beliefs about following rates are not statistically different from the actual following rate in the LPHE treatment ($p = 0.13$, Wilcoxon signed rank).

Given that sending a signal *B* when the state of the world is *G* is a weakly dominated strategy for Agents and Principals are aware of payoffs, Principals should always follow a *B* signal. We find that, on average, only 77% of Principals in all treatments follow signal *B*. Following rates are highest in the HPHE treatment at 92% followed by 79% in the control and 71% in the LPHE or HPLE treatments, with following in the HPHE treatment statistically significantly higher than the HPLE and LPHE treatments ($p = 0.02$ and $p < 0.01$ respectively). However, overall these rates are not statistically significantly different across treatments ($p = 0.28$, Kruskal-Wallis Test, two tailed).

Principal's beliefs about Agent truthfulness when the true state of the world is *G* is presented in Panel A of Table A8. On average, Principals believe that 31% of Agents send a deceptive signal, that is, they send signal *B* when the state of the world is *G*. This is far greater than the average actual deception rate of 4.2% ($p < 0.01$, Wilcoxon Signed-rank). In the control treatment, Principals believe that 34% of Agent signals will be deceptive when the state of the world is *G*. This is significantly higher than the expected rate of deception of 19% in the HPHE treatment ($p = 0.01$) and 22% in the LPHE treatment ($p = 0.007$). The highest actual following rate for a signal *G* is 71% for the HPHE treatment, which corresponds to the most optimistic Principal beliefs about truthful reporting in the HPHE treatment of 81%. This provides support for the fact that Principal beliefs about deception will be greatest in the control treatment but is not consistent with the manner in which beliefs are expected to vary with punishment treatments.

Table A8: Treatment Differences for behaviour and beliefs of Agents when in the Good state and Principals upon receiving Bad signal

Treatment Differences	Deception at Good State		Belief about following of Bad Signal		Follow of Bad Signal		Belief about Misreport when state is Good	
	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Panel A: Experiment 1								
Control - HPHE	6.3	0.12	-5.5	0.15	-13.4	0.07*	14.8	0.01**
Control - HPLE	3.9	0.35	-3.2	0.44	7.3	0.37	7.2	0.14
Control - LPHE_1	1.3	0.73	-5.8	0.09*	7.5	0.27	11.4	0.00***
HPHE- HPLE	-2.4	0.34	2.3	0.41	20.7	0.01**	-7.6	0.49
HPHE -LPHE_1	-5.0	0.16	-0.3	0.85	20.9	0.01**	-3.3	0.97
HPLE - LPHE_1	-2.6	0.49	-2.6	0.42	0.2	0.98	4.3	0.45
Panel B: Experiment 2								
Control - Min_5	3.2	0.14	-2.38	0.76	-6.1	0.46	11.20	0.04**
Control - Min_25	6.3	0.58	-0.98	0.59	3.1	0.73	14.68	0.00***
LPHE_1 - Min_5	2.0	0.65	3.41	0.33	-13.6	0.13	-0.25	0.93
HPLE- Min_5	-0.7	0.86	0.81	0.83	-13.4	0.17	4.03	0.61
HPHE - Min_25	0.0	1.00	4.53	0.45	16.4	0.06*	-0.11	0.64
LPHE_1 - LPHE_2	-1.1	0.81	3.31	0.55	-13.6	0.12	7.08	0.11
Control - LPHE_2	0.1	0.97	-2.48	0.50	-6.1	0.46	18.53	0.00***

Notes: All *p*-values based on Mann-Whitney-Wilcoxon test

Appendix D: Alternative tests of robustness

Table A9: Robustness of all treatment differences across Agents and Principals.

	Panel A: Agent					
	Action			Beliefs		
	Difference	P (MWW)	P (*t*-test)	Difference	P (MWW)	P (*t*-test)
		State Good			*Signal Bad*	
Control - HPHE	6.3	0.12	0.12	-5.5	0.15	0.21
Control - HPLE	3.9	0.35	0.35	-3.2	0.44	0.41
Control - LPHE	1.3	0.73	0.73	-5.8	0.09*	0.07*
HPHE - HPLE	-2.4	0.34	0.34	2.3	0.41	0.61
HPHE - LPHE	-5	0.16	0.16	-0.3	0.85	0.94
HPLE - LPHE	-2.6	0.49	0.49	-2.6	0.42	0.46
Control - Min_5	3.2	0.14	0.49	-2.8	0.76	0.57
Control - Min_25	6.3	0.58	0.14	-1.0	0.59	0.83
LPHE- Min_5	2	0.65	0.65	3.4	0.33	0.38
HPLE -Min_5	-0.7	0.86	0.86	0.8	0.83	0.84
HPHE - Min_25	0	1.00	1.00	4.5	0.45	0.42
		State Bad			*Signal Good*	
Control - HPHE	35.1	0.00***	0.00***	-7.8	0.06*	0.08*
Control - HPLE	30.2	0.00***	0.00***	-3.5	0.33	0.39
Control - LPHE	8.8	0.27	0.27	-1.5	0.46	0.66
HPHE - HPLE	-4.9	0.62	0.63	4.2	0.32	0.41
HPHE - LPHE	-26.3	0.00**	0.00**	6.3	0.13	0.15
HPLE - LPHE	-21.4	0.02**	0.02**	2	0.77	0.61
Control - Min_5	10.3	0.32	0.32	-9.5	0.03**	0.03**
Control - Min_25	25.8	0.01**	0.01**	-7.4	0.11	0.12
LPHE- Min_5	1.5	0.88	0.88	-8	0.07*	0.07*
HPLE -Min_5	-19.9	0.08*	0.07*	-6	0.2	0.23
HPHE - Min_25	-9.3	0.37	0.37	0.4	0.96	0.95
	Panel B: Principal					
		Signal Good			*State Bad*	
Control - HPHE	-22.3	0.02**	0.02**	-7.8	0.00***	0.00***
Control - HPLE	-20.3	0.03**	0.03**	-3.5	0.00***	0.00***
Control - LPHE	5	0.53	0.53	-1.5	0.11	0.11
HPHE - HPLE	2	0.84	0.85	4.2	0.62	0.71
HPHE - LPHE	27.3	0.00***	0.00***	6.3	0.00***	0.00***
HPLE - LPHE	25.3	0.00***	0.00***	2	0.01**	0.01**
Control - Min_5	-6.4	0.33	0.54	-3.7	0.63	0.46
Control - Min_25	-0.4	0.87	0.98	5.4	0.26	0.29
LPHE- Min_5	-1.4	0.62	0.89	-10.4	0.11	0.06*
HPLE -Min_5	-26.6	0.03**	0.02**	-24.1	0.00**	0.00**
HPHE - Min_25	-22.7	0.05*	0.05*	-17.2	0.00**	0.00**
		Signal Bad			*State Good*	
Control - HPHE	-13.4	0.07*	0.07*	14.8	0.01**	0.00***
Control - HPLE	7.3	0.36	0.37	7.2	0.14	0.17
Control - LPHE	7.5	0.27	0.28	11.4	0.00***	0.00***
HPHE - HPLE	20.7	0.01**	0.02**	-7.6	0.49	0.13
HPHE - LPHE	20.9	0.01**	0.01**	-3.3	0.97	0.42
HPLE - LPHE	0.2	0.98	0.98	4.3	0.45	0.36
Control - Min_5	-6.1	0.46	0.46	11.2	0.04**	0.05**
Control - Min_25	3.1	0.73	0.73	14.68	0.00***	0.00***
LPHE- Min_5	-13.6	0.13	0.13	-0.25	0.93	0.95
HPLE -Min_5	-13.4	0.17	0.17	4.03	0.61	0.49

HPHE - Min_25	16.4	0.06*	0.06*	-0.11	0.64	0.98

Note: P (MWW) implies *p*-value from two tailed Mann-Whitney-Wilcoxon test, while P (*t*-test) is *p*-value from two tailed *t*-test