Why Blame?

Mehmet Gurdal, Joshua B. Miller, Aldo Rustichini

No 1022

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF
WARWICK

Why Blame?

Mehmet Gurdal[a], Joshua B. Miller[b], Aldo Rustichini[c] *

August 31, 2013

Abstract

We provide experimental evidence that subjects blame others based on events they are not responsible for. In our experiment an agent chooses between a lottery and a safe asset; payment from the chosen option goes to a principal who then decides how much to allocate between the agent and a third party. We observe widespread blame: regardless of their choice, agents are blamed by principals for the outcome of the lottery, an event they are not responsible for. We provide an explanation of this apparently irrational behavior with a delegated-expertise principal-agent model, the subjects' salient perturbation of the environment.

JEL Classification Numbers: C92; D63; C79.
Keywords: Experiments; Rationality; Fairness
Journal of Political Economy, forthcoming

*a: Department of Economics, Boğaziçi University, Istanbul, Turkey, b: Department of Decision Sciences and IGIER, Università Bocconi, c: Department of Economics, University of Minnesota. The authors are grateful to Alexander Vostroknutov, Adam Sanjurjo, Connan Snider and Julia Thorton-Snider for comments on the draft and to Jan Werner, Jarek Grygolec, participants in the Mathematical Economics workshop at the University of Minnesota for helpful comments, participants at Washington University's Graduate Student Conference in St. Louis, and seminar participants at the Federal Reserve Bank of Boston, Einaudi Institute and Bocconi University. We also thank the editor and two anonymous referees for insightful comments that guided us to a complete revision of an earlier draft. Corresponding author: Joshua Miller. Address: Department Of Decision Sciences, Università Bocconi, Via Roentgen 1, Milano, 20136, Italy, telephone: +39 02 5836.3411, e-mail: *joshua.miller@unibocconi.it*.

We assign responsibility to a man, not in order to say that as he was he might have acted differently, but in order to make him different. F.A. Hayek

1 Introduction

We define blame as the channeling of negative feelings produced by an undesirable event towards someone associated with that event. Quite often this behavior is misguided, i.e. it is directed at those who are not responsible for the event. This (implicit) assignment of responsibility when it is unjustified can have important economic and political consequences. For example, there is evidence that U.S. presidents are blamed by voters for poor economic performance that cannot be credibly attributed to their decisions (Converse 1964). Similarly, CEOs who are under-performing the industry average are blamed by shareholders and voted out more readily during industry downturns than they are during booms(Jenter and Kanaan 2011).[1]

Blame can serve a functional purpose. In a repeated relationship between parties, blame can be viewed as a way of expressing voice in alternative to exit (Hirschman 1970). Voicing dissatisfaction may be one way of signaling to a partner a discrepancy between current expectations and realizations, and of warning about future changes in behavior if the discrepancy were to persist. Unjustified blame is not easily explained in this way, because it might well be counterproductive in repeated relationships. The essential problem for a functional explanation of blame as voice, however, is that it occurs even when the relationship is not repeated.

Unjustified blame, particularly in exchanges that only occur once, can therefore be viewed as irrational and unfair. To formally determine what is rational and fair, one needs a benchmark, which can be constructed by appealing to two philosophical principles: the control principle and the merit principle. The control principle asserts that individuals should be considered responsible only for events that are under their control (Nelkin 2004) while the merit principle states that responsibility is a necessary condition for blame and praise (Kleinig 1971).[2] Taken together, these principles seem to provide support for the claim of the irrationality of unjustified blame. This position has a long tradition in ethics, made explicit for example in the initial paragraphs of Kant's

[1] Depending on one's point of reference, one could argue that U.S. presidents and CEOs are instead unjustly praised in the state of the world where economic and industry factors are favorable. In practice, when a difference in behavior is observed, without knowing an individual's point of reference, one cannot know if the responsibility being assigned is blame, or less praise.

[2] One may also establish the unfairness of unjustified blame by defining fairness via the equal treatment of equal individuals (Moulin 2003). Accepting that uncontrollable events should not break an existing equality relation between individuals necessitates one to view blame as unfair if based on an event someone cannot control.

Foundations: "A good will is good not because of what it performs or effects, not by its aptness for the attainment of some proposed end, but simply by virtue of the volition" (Kant 1784).

In much observed behavior and in the experiments presented here, however, people regularly violate these principles. This violation follows a precise pattern: it appears to be based on a counterfactual evaluation, similar to the one used in envy and regret. When obtaining an outcome determined at least in part by the action of others, people compare the outcome obtained with the one they could have obtained if the other person had behaved differently. The affective response, and the vocal complaints, are proportional to this difference. Why then do people blame others? In particular, why do they blame them when they are not responsible and what explains the use of a counterfactual evaluation? Is this evidence of irrationality?

Rational Blame

Blaming someone based on an event one knows they have no control over is not immediately explicable in terms of two of the most successful economic approaches to actor-responder type relationships: contract theory and reciprocity theory. Among the main general predictions of the contract theoretic approach is the *informativeness principle* from the principal-agent model. The principle predicts that a principal's behavior, described by the optimal contract, will incorporate only signals that provide information about an agent's payoff-relevant actions (Bolton and Dewatripont 2004; Hölmstrom 1979). Economic theories of reciprocity and the available experimental evidence they organize additionally suggest that a principal's behavior—reciprocal actions in this case—will incorporate information about the "kindness" of the agent's actions and the intentions that produced them, as well as account for the distributional consequences (Bolton and Ockenfels 2000; Charness 2004; Charness and Levine 2007; Charness and Rabin 2002; Falk, Fehr, and Fischbacher 2008; Falk and Fischbacher 2006; Fehr and Gächter 2000; Fehr and Schmidt 1999; Rabin 1993). Both theories seem unable to provide an explanation of blame when it is unjustified. Events that an agent cannot control or influence in any way cannot be informative about an agent's payoff-relevant actions, hence the principal-agent model appears silent for these environments. Similarly, these uncontrollable events are not informative on the "kindness" of those actions, or the intentions which produced them; and don't always affect the distributional consequences. Therefore theories of reciprocity do not seem to be directly relevant for explaining unjustified blame.

This apparent inability of contract theory or reciprocity theory to explain why people are blamed for events they cannot control could be due to the confounding factors that are always

present in uncontrolled natural settings, rather than any incompleteness in the theories themselves. A particularly obvious problem in natural settings is that wealth is often correlated with outcomes, so behavior that appears to arise from blame might instead be driven by budget constraints or distributional preferences. The uncontrolled structure of uncertainty presents another difficulty. Legitimate ambiguity with regard to cause and effect may be a sufficient reason to believe that an event may be under the control of someone when it is not. In addition, there always exist some payoff-relevant actions or kindness-relevant actions which are known to be unobservable and may influence events. These potential simple confounds should be kept in mind when analyzing data and when designing experiments: one must first control and exclude them. We will return to this point when we present our experimental design. But these considerations leave the general questions still open, should unjustified blame remain in a controlled setting: why is there blame when it is unjustified? Where do we look for a theoretical explanation?

A potential insight into blame can, in fact, be provided from the basic structure of principal-agent theory, with an expanded conception of its applicability. In situations where an agent's actions can increase the probability of a good outcome, it is rational (and is part of the optimal contract) to punish the agent for a poor outcome, even if it is commonly known, in equilibrium, that the agent must have chosen the good action (work and not shirk) and hence the poor outcome is known not to be the agent's responsibility. So a basic lesson of the principal-agent model is that as long as the agent's action influences the probability of a good outcome and is under his control, then if nature yields a poor outcome, the principal must punish the agent even though the principal knows in equilibrium that the agent chose the desired action; that is, a feature of the optimal contract is precisely unjustified blame. Though blame is not justified on what can be observed or inferred, it nevertheless becomes justifiable on rational grounds because it induces, ex-ante, the appropriate incentives. Therefore, by guaranteeing the accountability of decision makers in society, blame can be rationally supported as a part of a normative morality.

The idea that the foundational norms for a contractual relationship between society (the principal) and individual (the agent) can rationally feature unjustified blame (and praise) was advanced by)Hayek (2011 (1960).[3] The hypothesis we suggest is that blame is the emotional expression of

[3] "Our problem is generally not whether certain mental factors were operative on the occasion of a particular action but how certain considerations might be made as effective as possible in guiding action. This requires that the individual be praised or blamed, whether or not the expectation of this would in fact have made any difference to the action. Of the effect in the particular instance we may never be sure, but we believe that, in general, the knowledge that he will be held responsible will influence a person's conduct in a desirable direction. In this sense the assigning of responsibility does not involve the assertion of a fact. It is rather of the nature of a convention intended to make people observe certain rules." *The Constitution of Liberty*, Chapter V, Responsibility and Freedom.

this feature of the optimal contract. In environments where payments and rewards are administered according to a contract (may be a formal one, or may be informal, supported for instance by social norms) this feature of the optimal contract in a principal-agent relationship provides a simple explanation of blame. Does this explanation extend to other environments, where the principal-agent model does not directly apply? We will argue it does; but to make this connection clear it is necessary first to describe our experimental setup.

Our study

The present study consists of a main experiment, a proposed model to explain the results, and two additional experiments which vary the incentives and allow for learning. The main experiment, which we term the Allocate treatment, is carefully designed to identify blame as the causal mechanism determining observed behavior. The design can be briefly described as follows (the complete description is in Section 2.1): In each round of the experiment two subjects were paired for the first and only time. One subject, the "agent", decided between a risky alternative, consisting of a lottery, and a safe alternative, consisting of a certain payment, on behalf of another subject, the "principal", who then observed the decision.[4] Next, the outcome of the lottery was determined by the public toss of a die so the principal could observe the outcome of the lottery no matter what the agent chose for him. Then the principal observed his payoff from the agent's choice and engaged in an allocation task, deciding how much money to assign to the agent and a random third party, to be selected among the other agents in the session. The total amount assigned could be between \$0 and \$15 and was from an outside account that did not belong to the principal. Finally, the two payments were communicated to the agent and any unassigned amounts were left to the experimenter.

Several features of the design are worth noting. The determination of the outcome of the lottery through the public toss of a die made it completely clear and salient that the agent had no influence on the outcome of the lottery. The fact that the payment of the principal came from an account owned by the experimenter assured that the principal had no direct incentive— including risk sharing motivations—involved in their payment decision. The fact that the principal could pay the agent or a third party assured that he was deciding how to allocate a dollar between two otherwise identical individuals from his point of view, with the only exception that one and only one of the two is associated with the choice and the outcome. If the principals were allowed to

[4] In the experiment itself the agent was called the "option chooser" and the principal was called the "rule chooser"; while the lottery was called the "risky option".

choose between allocating only to their agent or returning the money to the experimenter we may have been unable to control for two motivations: (1) principals could have efficiency motives, or construe a subject vs. experimenter environment and simply assign $15 regardless of the event, and more importantly, (2) principals could be motivated to assign less to all other subjects, regardless of their association with the uncontrollable event, which in the case of only one subject would be observationally equivalent to blame.

In the experiment, principals assign money to agents based on the outcome of the lottery, an event the agents cannot control. Regardless of the choice of the agent, principals assign less money to the agent if their payoff from the chosen alternative is lower than the payoff from the unchosen alternative. This behavior persists with the introduction of a direct cost and an extension from ten decision periods to twenty-five. The design assures that when the principal assigns a lower payment to the agent this is directed at the agent. Thus we may conclude the principal is holding the agent individually accountable and we may label this behavior blame.

Blame in the literature

Our results are related to earlier work on social judgement in the social psychology literature. Findings from that literature indicate that when evaluating decisions of others, people will over-emphasize information revealed after a decision is made. For example, a behavioral phenomenon termed *outcome bias* (Baron and Hershey 1988), occurs when individuals ("judges") who are asked to evaluate the choice of a "decision maker" take into account outcome information that is irrelevant for the particular judgment they are making regarding anothers decision. The experimental demonstration of outcome bias typically involves subjects reading a hypothetical vignette describing a decision maker choosing between two actions, where one action leads to a deterministic outcome and the other action is followed by two possible random outcomes with probability independent of the action. The subjects judge the hypothetical decision maker after reading the vignette. Outcome bias occurs when the judgement is more favorable following a good outcome than a bad outcome, in spite of the fact that the outcome is known to be due to chance and in no way determined by the actions of the decision maker. Outcome bias has been observed in studies involving decision quality evaluation (Baron and Hershey 1988; Mowen and Stone 1992; Tan and Lipe 1997), responsibility attributions (Walster 1966), ethicality judgments (Gino, Moore, and Bazerman 2008), judgements of culpability (Gino, Moore, and Bazerman 2008; Mazzocco, Alicke, and Davis 2004), and punishment recommendations (Mazzocco, Alicke, and Davis 2004). Outcome bias in judgment may

be moderated somewhat by perceived controllability (Tan and Lipe 1997), but its existence in the domain of vignette-based experimental studies, is significant, substantial and robust to changes in the response measure.

The presumption behind the term outcome bias seems to be that when an outcome does not reveal new information about an agent's decision process, then that outcome should not be considered in the evaluation of the agent, which should instead only depend on the action; that is the judgment should follow the control principle. This presumption may not be justified: in the optimal contract of the principal-agent problem the principal's payment depends on the outcome even if at the equilibrium of the game the principal knows that the agent chose the good action. The reason for behavior that appears to be biased by the outcome is that a differential payment is necessary ex ante to induce the agent to choose the right action, as Hayek's insight suggests.

While outcome bias is clear and established in the social psychology literature, it has not been demonstrated that this bias in *judgement* translates into a "bias" in decision making, in particular that it would lead to unjustified blame in the context of real subject interactions with real monetary consequences which affect both parties. In the experimental economics literature there is evidence against unjustified blame for uncontrollable events, at least in situations where first movers can transparently reveal their costly sacrifice, or first movers have no control over both the events and their own actions which precede the events (Charness and Levine 2007; Charness and Rabin 2002; Falk, Fehr, and Fischbacher 2008). In our experiment, the ability to signal costly sacrifice was minimal (only via good decisions) and agents had full control over the actions they made, which was independent of the uncontrollable event. This created an experimental environment which shared more features in common with studies in the decision evaluation literature and thus, presuming the payment choice of principals was mediated by the same outcome-sensitive emotional or cognitive process, led to the presence of the unjustified blame that we observed.

The findings in the social judgement literature and ours pose a common puzzle: why does unjustified blame exist in such settings? As we mentioned, the behavior in the experiment may not be immediately explicable in terms of contract theory, because there is no incentive for the subjects in the role of principal to put effort into evaluating their own preference (between the risky and safe alternative) before their agent chooses as they cannot write a contract to influence the decision of the agent.

Salient Perturbations

We propose an explanation of the data that is based on the principal-agent model for environments where the action of the agent is not observable. In outlining a theory of salient perturbations, proposed by Roger Myerson (Myerson 1991), he conjectures that an individual responds to a strategic but unfamiliar situation by interpreting it as an instance of a more familiar situation, close enough (salient) to the present one to be taken as guidance for rational behavior. For example, a theory of salient perturbations can provide an explanation of ambiguity aversion, understood as the specific aversion (additional to simple risk aversion) that an individual may feel when confronted with uncertainty that is difficult to quantify. The *reason* for such aversion is that if an individual is offered a bet with payment contingent on such an uncertain event, he may interpret the situation as an instance of the more familiar situation where the person who proposes the bet has some control over the outcome, or private information relating to it, and is trying to profit from this position. Thus the apparent irrationality of ambiguity aversion in the specific situation is what would be the rational response in a related more familiar environment.

In our experiment, the salient perturbation is the standard hidden-action principal-agent situation, where agent *may* influence the outcome or may acquire information on its likelihood, and the equilibrium choice for the principal implies a reward to the agent for good outcomes, and a punishment for bad ones. The principal's lack of incentive to engage effort and attention to evaluate the agent's decision problem or monitor the agents choice creates a situation similar to the one where the action of the agent has an effect on the outcome and cannot be observed. Upon seeing the outcome, subjects respond as if they are in a hidden information environment and assign less to the agent when they are worse off (counterfactually). We contend that the behavior in the experiment corresponds to unjustified blame in natural settings, which can appear as a lower bonus, a lower wage, or exit.

The remainder of the paper is as follows: in Section 2.1 we present the design of the experiment, in Section 2.2 we present the theory and predictions, in Section 2.3 we present the results, in Section 2.4 we present the design and results from additional treatments investigating the robustness of the results to incentives and learning, and in Section 3 we conclude.

2 Experiments

A total of 554 subjects were recruited from first year principals of economics courses and compensated between \$8 and \$70, depending on their decision and that of their partner in a period that was randomly selected (an average payment of \$25).[5] When subjects arrived at the laboratory they were randomly assigned to carrels and then given an overview of what to expect, including how they were paid based on a randomly selected decision. Next, they were presented with common instructions after which they were assigned to one of two fixed roles for the remainder of the session. Before the task began subjects completed two practice rounds to familiarize themselves with the interface, and then given a questionnaire to test their understanding that involved customized feedback, conditional on their responses. In Appendix Section A a full description of the procedures is presented.

2.1 Design

Risky versus Safe Alternative

The experiment involved 10 decisions between a risky (R) and safe (S) alternative. The risky alternative consisted of a lottery which yielded a high payoff \$$h$ with probability $P(h)$ and low payoff \$$\ell$ otherwise, where $\ell = 0$ in every period. The safe alternative yielded a certain payment \$$c$. The values for h, $P(h)$, and c varied across periods, with $h > c > \ell = 0$ and $P(h) \in \{.25, .5, .75\}$. For each period these parameters were fixed and identical across all subjects, all experimental treatments and all sessions (see Table 1).

The outcome of the lottery was determined by the roll of a single 4-sided physical die which was observed by all subjects. They were informed that the mapping between die numbers and payoffs was randomly permuted across subjects, as this assured that the realized outcomes varied between subjects. For example, as can been seen in Table 1, in period 1 all subjects had the choice between \$9 and (\$30, .25), but any given four subjects could see a different representation of the lottery: for the first subject a die face of 1 could yield \$30 with the other die faces yielding \$0, for the second subject a die face of 2 could yield \$30 with the other die faces yielding \$0 each, etc.

The lotteries were designed so that nearly risk-neutral principals would want the safe alternative in Periods $1, 6, 7, 8, 10$ and the risky alternative in Periods $3, 4, 9$ and would be nearly indifferent between safe and risky for Period $2, 5$, as indicated in the final two columns of Table 1.

[5] A total of 304 subjects participated in the main experiment, with an additional 250 subjects participating in the two additional treatments with costly transfers.

Table 1: *The Binary decision problem for each period (all treatments)*

Period	Risky(R)*	Safe(S)	EV[R-S]	Risk Neutral Choice
1	(\$30, .25)	9	−1.5	Safe
2	(\$20, .50)	10	0	Indifferent
3	(\$20, .25)	4	1	Risky
4	(\$10, .75)	7	0.5	Risky
5	(\$30, .50)	15	0	Indifferent
6	(\$20, .25)	8	−3	Safe
7	(\$10, .25)	5	−2.5	Safe
8	(\$30, .25)	12	−4.5	Safe
9	(\$20, .50)	6	4	Risky
10	(\$10, .25)	3	−0.5	Safe

* The low outcome for the lottery was \$0 for all decisions.

Design for each Period

At the beginning of each of the 10 periods, each principal was randomly and anonymously matched with an agent they hadn't faced before.[6] In the first stage of a period the agent chose without cost between the risky and safe alternative; the principal was idle in this stage, but was shown the lottery and certain amount the agent was choosing between. In the second stage the principal observed the decision the agent made for him. After several seconds the die was thrown so all subjects could view the die and infer the outcome of their lottery. If the agent chose the safe alternative, the principal received the payoff \$$c$. If the agent chose the risky alternative then the principal won \$$h$ if the lottery outcome was high and \$$\ell$ = \$0 if the lottery outcome was low. In the third stage of the period the outcome of the lottery was circled on the screen of all participants and the amount won was presented on the screen of the principal, who was then prompted to assign a payment to the agent and to a third party, to be selected randomly in the set of all other agents (and not principals).

The payment to the agent and the third party was between \$0 and \$15 and together had to total to \$15 or less. Any unassigned money was not kept by the principal. Thus, the principal could divide a total of \$15 between the agent, the third party and the experimenter. Next, the agent viewed both payments determined by the principal. The third party benefitting from the transfer was randomly selected, and found out about their payment at the end of the experimental session. Finally, both agent and principal rated how they felt about their partner's decision on a 1-10 scale from very bad to very good. Then the next period began.

[6] Recall the subjects were randomly assigned to these fixed roles immediately after the instructions were presented.

The design has several important features that deserve further elaboration. First, the fact that the experimenter and not the principal kept any unassigned money assures us that any motivation that drives the real allocation decision of the principal cannot be individual financial gain. This means that variation in the principal's payment to the agent cannot be explained by the principal's ability to pay, the marginal utility of assigning payment or risk sharing between principal and agent. The lack of financial incentives tied directly to the principal's choice could have induced them to behave in the same way in every instance or without any pattern altogether, but they clearly did not: the economic consequences of the principal's decisions for the agent were real, and principals took this fact into account.[7] Second, having the principal pay both their agent and a random other agent had a crucial purpose which goes beyond simply eliminating the subject versus experimenter effects or efficiency motives. As we are observing payments and not directed judgment, it was essential to have a design feature which allowed us to identify that the principal's payment decision to the agent was directed specifically at the agent. If the principal was allowed only to adjust how much their agent was paid, then observing the principal pay less to the agent when the principal was relatively worse off could be driven by confounding factors: (1) irritation or disappointment, where principals have a propensity to share their irritation with all others not just the agent, or (2) perception, if a principal's payoff is \$0, then assigning \$1 to anyone may seem as a lot more than if the principal's payoff is \$30.[8] Thus, with this third-party payment feature, when we see a difference between treatments in the payment to the agent, we know that it is the principal holding the agent responsible and we can properly call it blame.

2.2 Theory and Predictions

We propose that a hidden-action principal-agent model may explain the occurrence of blame in the experiment and provide precise predictions. While the choice of the agent was observable and under the agent's control, the fact that the principal assigned payment after the agent had chosen and the outcome was revealed meant that there was no incentive for the principals to evaluate which decision was best for themselves. Upon seeing the outcome of the lottery the situation might be perceived by the principals as a hidden-action environment, where the hidden action is not the agent's choice, but the effort employed to evaluate which alternative is best for the principal. We claim that this is the most natural environment a subject may have in mind when making his

[7] In Section 2.4 we investigate the role of incentives tied directly to the principal's decision.

[8] These concerns proved to be warranted. While we did identify that the differences in the principal's payment to the agent was directed at the agent, the total amount assigned was lower when the payoff was relatively low for principals.

decisions, the salient perturbation of the current environment. In most environments others do have an influence over outcomes, and typically the effort, care and attention they put in evaluation and execution is not observable.

The basis for our predictions is provided by a principal-agent model where an agent can first exert hidden effort to evaluate which alternative is better for the principal and then makes an observable choice. With the counterfactual outcome produced by the unchosen alternative observable, as in our design, this *delegated expertise* model makes more stringent predictions than the standard hidden-action principal-agent model. In Appendix B we consider the optimal contract in the following set-up: an agent has to choose between a risky asset R (with two outcomes, one high h and one low ℓ) and a safe asset S for a principal; where the payoff of the chosen asset will be paid to the principal. He can devote hidden effort in variable degrees to acquire information on the potential return of the asset. The probability of a high outcome h is equal to $P(h)$ prior to devoting effort. The principal pays the agent according to what he can observe, namely the choice of the asset and the outcome of the asset. Denoting the payment that principals assign to the agent in the four possible outcomes defined by the agent's choice and the outcome of the lottery as $w(R,h), w(R,\ell), w(S,h)$ and $w(S,\ell)$, one can easily show (see Appendix B) that the optimal contract for an agent (with an outside option equal to zero) is such that:

$$w(S,h) = w(R,\ell) = 0; w(S,\ell) > 0, w(R,h) > 0 \tag{1}$$

Taking Equation 1 as a prediction of the model for our experiment, two clear implications are that $w(S,\ell) > w(S,h)$ and $w(R,h) > w(R,\ell)$ (even if we allow for a nonzero outside option) i.e. we predict unjustified blame in states (S,h) and (R,ℓ) (and equivalently unjustified praise in states (S,ℓ) and (R,h)). Clearly what matters are the inequalities between payments: without knowing a principal's point of reference we cannot determine, for example, if a positive wage $w(S,\ell)$ by itself signifies praise, or less blame.

Additionally, in the optimal contract, the absolute size of the two positive payments in the counterfactually better state of the world—$w(S,\ell)$ and $w(R,h)$—depend on the parameters. Their ratio is equal to the odds that the lottery doesn't pay off, i.e. $w(R,h)/w(S,\ell) = \frac{1-P(h)}{P(h)}$.[9] The intuition behind this comes from the principal wanting the agent to put in effort and choose the alternative they discover to be more promising; the more unlikely it is for the lottery to yield

[9] See Equation 18 of Appendix Section B.

a high payoff ex-ante, the more the agent needs to be rewarded in the state when they choose it and it pays off, (R, h), than when they don't choose it and it doesn't pay off, (S, ℓ). While this intuition is sensible in the delegated-expertise setting, it is not in the experimental setting. Lowering the odds that the lottery yields a high outcome while holding the payoff parameters h, ℓ, and c constant makes the lottery less attractive in an ex-ante sense, and thus makes choosing the lottery a worse choice. This means if principals follow what the comparative statics result predicts, in these cases they are rewarding poor decisions. Though this appears to be a rather unlikely prediction, a plausible underlying psychological mechanism does exist: the feeling of relief from avoiding a probable loss.

Deciding how to operationalize the comparative statics prediction presents a challenge. A within-subjects test of the comparative statics relationship cannot be performed because in a given period subjects are randomized into state (R, h) or (S, ℓ), but they cannot appear in both states. This means we must formulate our prediction using between-subjects data. A reduced-form prediction, which could provide a suggestive test for the comparative statics relationship, is that there should be a negative correlation between the probability the lottery yields a high outcome, $P(h)$, and the ratio of the average payments to the agent in each state, $w(R, h)/w(S, \ell)$.[10] In the results section below we test this prediction, along with the main predictions of the counterfactual evaluation.

2.3 Results

Summary Comparisons

In any given period of the experiment, our design induced a 2×2 between-subjects randomization where subjects in the role of principal were assigned to one of two decisions, indirectly, by being matched with an agent that chooses the risky (R) or the safe (S) alternative, and one of two outcomes, where the die roll determined the high (h) or low (ℓ) outcome of the lottery.

The rational ethical benchmarks, (merit and the control principle) together predict that payment may depend on the choice of the risky or safe alternative, for which the agent is responsible, and not the outcome:

$$w(R, h) = w(R, \ell) \quad \& \quad w(S, h) = w(S, \ell) \tag{2}$$

[10] While straightforward, this measure is delicate; the number of observations in each state may not be equal. For example, the random assignment to state (R, h) depends on the (endogenous) probability that the randomly matched agent chooses the lottery and the probability the lottery yields the high outcome, if both probabilities are low, there may be few observations in state (R, h) and many observations in state (S, ℓ). In the results section we address this issue with a weighted regression.

Instead, if subjects follow the counterfactual evaluation and blame their agents when their payoffs are worse than if the agent had chosen otherwise—as predicted by the optimal contract of the delegated expertise model—then we should see the following pattern of payments assigned to the agent:

$$w(R,h) > w(R,\ell) \quad \& \quad w(S,\ell) > w(S,h) \tag{3}$$

We shall see that the evidence is clearly supporting the second alternative, described by equation 3.

As the parameters of the risky and safe alternatives–payoffs and probabilities–vary between periods, we analyze the periods individually before we pool them together and control for these potential confounds in a regression model.

We present in Figure 1a, for the case when the agents chose the risky alternative, both the principals' average payment to the agent and their average rating of their feeling towards the agent's decision, for each lottery outcome. Focusing on the payment to the agent, it is clear that in each period the principals pay more to their agents if the lottery yields the high outcome than if it yields the low outcome. For the purposes of statistical tests we focus on the five periods with at least 10 observations for each lottery outcome, Periods $2, 3, 4, 5$ and 9. (see Table 9 in Appendix D.5 for other periods).[11] In each of these periods, the difference in the principal's payment to the agent is significant ($p < 0.01$ Mann-Whitney U-test). It is also apparent that being relatively better off ex-post is not the only factor that determines the average size of the payments. In periods 4 and 9, periods where the overwhelming majority of principals prefer the lottery—as evidenced in a separate preference elicitation task conducted at the end of the session—principals pay more for the low payoff of $\$\ell = \$0$ than in other periods.[12] It appears that the principal's preference for the chosen alternative influences the payment they assign to the agent, suggesting that the perceived quality of the agent's decision mitigates blame, which is evidence of a form of intentions-based reciprocity. In regressions discussed further below, and presented in Table 2, this perceived quality effect is shown to be significant, though not substantial relative to the effect of the outcome.

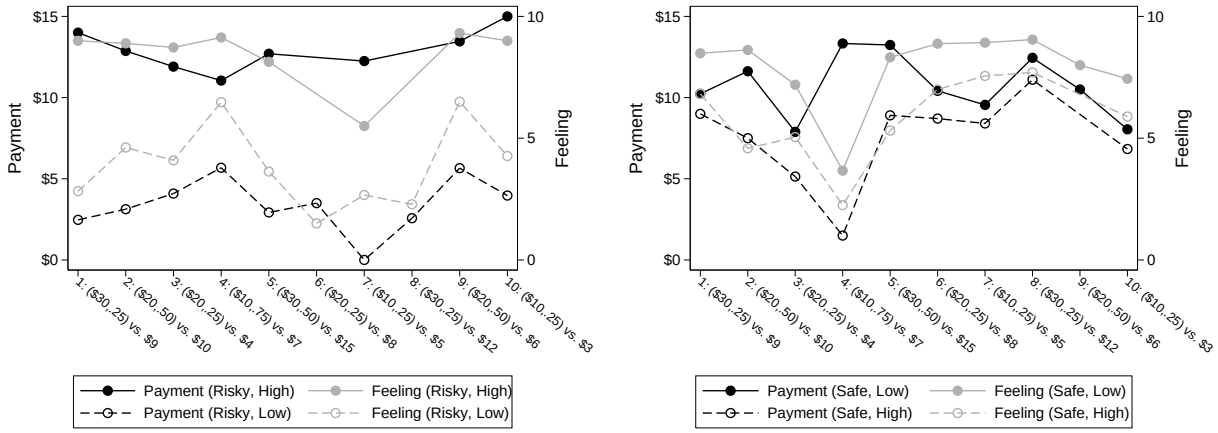
While the significant treatment effect of the lottery outcome is unambiguously causal in the data

[11] That only half the periods had more than 10 observations in each condition is to be expected; the experiment did not elicit contingent responses from the principals and the probability of being randomly assigned to a particular event could be quite low. For example, if a high payoff on the lottery had a 25% chance of occurring and the lottery was unfavorable, then the probability of both the lottery yielding a high outcome and being assigned to an agent who chose the lottery was quite low, in fact in periods 6 and 8, with 228 observations, this did not happen once.

[12] The preference elicitation task, where both principals and agents faced the same 10 risky and safe alternatives and chose for themselves, is described in Appendix Section A; the actual choices are reported in Figure 6 of Appendix Section C.

Figure 1: *Principals' average allocation to their agents and feeling towards their agents' choice (Allocate treatment)*

(a) *Risky Alternative Chosen* **(b)** *Safe Alternative Chosen*



where the risky alternative is chosen, this alone cannot identify blame as the causal mechanism. When agents choose the risky alternative, two factors vary between the treatments: the outcome of the lottery, and the wealth of the principal. The difference in wealth can drive differences in payments via consequence-based reciprocity, arbitrary numeric anchoring or, if the third party is not seen as an equal, a distributional preference to equalize risk or payoffs with the agent. In Figure 1a, if we look at the pattern exhibited in the principals' subjective ratings of how they feel about their agent's choice we can see that it follows a similar pattern to that of the payment. This, along with the fact the principals routinely allocate substantially more than \$0 to the agent when they receive \$0, suggests that distributional preferences are an unlikely mediator leading to the payment differences. We cannot rule out consequence-based reciprocity or numeric anchoring and therefore identifying the influence of blame requires us to look at the states where the agent chooses the safe alternative and then pool all the data into a regression model to take advantage of between period variations in wealth.

When the agent chooses the safe alternative for the principal, the outcome of the lottery does not influence the payoff of the principal: the only factor that varies is whether the principal is better off or worse off than if the agent had chosen otherwise. This control means that the data where the safe alternative is chosen provide us with the perfect setting to test if people will blame others for events they know they cannot control. We present in Figure 1a, for the case when the agents chose the safe alternative, both the principals' average payment to the agent and their average rating of their feeling towards the agent's decision, for each lottery outcome. In each period, principals

allocate less to their agent if the lottery yields the high outcome than if it yields the low outcome, consistent with a counterfactual evaluation, as the principals are relatively worse off than if the agent had chosen otherwise. These differences are significant ($p < 0.05$, Mann-Whitney U-test) for 6 out of 7 periods that have at least 10 observations in each treatment, periods $1, 3, 5, 6, 8$ and 10.[13] As payments to the agent follow the counterfactual comparison when the safe alternative is chosen, we should also observe that the payments and the principals' ratings of their feeling towards agent's choice move together. In Figure 1b we do in fact observe the principals' average ratings match the pattern of average payments rather closely. The pattern in the ratings is consistent with earlier work on outcome bias in social judgement and evaluation (Baron and Hershey 1988). Our results suggest that this feeling produced by the outcome is channeled into the principals payment decision, which is in turn directed towards the agent associated with the outcome, i.e. the principals engage in unjustified blame.

In order to investigate the influence of the outcome on the principals' payment decisions relative to other factors related to the parameters of the decision problem–probability and payoffs–we perform a regression with statistical controls.

Regression analysis

Pooling observations together and performing a regression analysis of the principal's payment allows us to use data from all periods and take advantage of the between period variability of the parameters for safe and risky alternatives. We can isolate the blame effect on the principal's payment in the risky choice data using the predictor variables determined by random assignment and quantify the association between the principal's payment and important factors such as the quality of the agent's choice as well the principal's utility from the payoff they receive. We analyze principal behavior from all four decision and lottery outcome conditions in a single regression model. The regression model below indicates that blame is significant, substantial, and in the predicted directions for both the risky and safe decisions.

We employ a panel data regression model, as we have repeated observations from subjects. Let $w_{i,t}$ be the payment assigned by the ith principal to the agent in period t. We estimate the following model:

$$w_{i,t} = \boldsymbol{\beta}' \boldsymbol{x}_{i,t} + \eta_i + \varepsilon_{i,t} \tag{4}$$

where $\boldsymbol{x}_{i,t}$ is a column vector consisting of 1 and the independent variables, η_i represents the latent

[13] See Table 9 in Appendix D.5 for a complete listing of all periods

individual characteristics of subject i, and $\varepsilon_{i,t}$ is the error term with the appropriate gaussian distribution assumptions. Unless stated otherwise we assume that the latent effects η are random, with mean zero, and uncorrelated with $\boldsymbol{x}$ and ε, i.e. a random-effects model.[14]

We use seven independent variables: *risky, safe, high, low, utility, E(utility premium)* and *period*. The variable *risky* is a dummy variable equal to one if the risky alternative is chosen, *safe=1-risky*. The variable *high* is a dummy variable equal to one if the die face indicates that the lottery yields a high payoff, while *low=1-high*. The variable *E(utility premium)* is a control equal to the difference between the principal's expected utility of the chosen alternative and that of the alternative not chosen. It is a measure of the principal's subjective attitude towards the quality of the chosen alternative, and it is intended to capture any between period variation in a possible effect driven by the principal's recognition of the agent's effort, or intentions-based reciprocity. The control variable *utility* is an ex-post measure of the principal's well-being, their utility from the payoff, which might be influenced by disappointment and relief, distributional preferences, or consequences-based reciprocity. Both *E(utility premium)* and *utility* are computed using the risk aversion parameter of the principal's CRRA utility function, which is estimated at the group-level for all principals in the final preference elicitation task.[15] The variable *utility* takes values in $[0, 28.26]$ while the variable $E(utility\ premium)$ takes values in $[-4.5, 4.5]$. The variable *period* records which period the decision was made in from 1 to 10.

The estimated coefficients for the regression model of the principal's payment to the agent under three alternative specifications are listed in Table 2. The first column presents the coefficients for the specification with only the experimentally manipulated variables. The columns two through four introduce the statistical controls, while the fifth column checks for a period effect.

Counterfactual evaluation, not merit principle

The estimations in Table 2 corroborate the findings of Figure 1: the payment depends on the outcome of the lottery, and not just on the choice of the agent. The main interactions to test for the randomized treatment effects are $safe \times low$, $risky \times low$ and $risky \times high$, all relative to

[14]There is every reason to assume that latent individual effects are independent of the explanatory variables in our estimated models as the individuals have no control over explanatory variables. The Hausman test justifies the use of a random effects model as the test statistic $H \sim \chi^2(7)$, is $H = 1.65$ ($p = 0.98$) failing to reject the null hypothesis of consistency of the random effects model.

[15]The preference elicitation was always the final task of the experimental session, described in the procedures section of Appendix A. The coefficient of relative risk aversion was estimated to be $r = 0.031$ using a binomial discrete choice model detailed in Appendix C. The estimation in Table 2 is not sensitive to replacing the principal's estimated CRRA utility function with a risk neutral one, or alternative ways of measuring the quality of the agent's decision (see Online Appendix Section D.4).

Table 2: *Estimated effects on the principal's payment to the agent (Allocate treatment)*

	(1)	(2)	(3)	(4)	(5)
constant	8.319***	6.266***	7.455***	5.908***	5.935***
	(0.344)	(0.436)	(0.335)	(0.419)	(0.460)
safe × low	1.848***	2.012***	1.466***	1.664***	1.665***
	(0.318)	(0.289)	(0.308)	(0.284)	(0.284)
risky × low	−4.100***	−2.044***	−3.509***	−1.915***	−1.915***
	(0.466)	(0.497)	(0.449)	(0.480)	(0.481)
risky × high	3.921***	1.829***	4.279***	2.499***	2.505***
	(0.391)	(0.409)	(0.374)	(0.407)	(0.411)
utility		0.251***		0.207***	0.206***
		(0.030)		(0.031)	(0.031)
E[utility premium]			0.581***	0.485***	0.489***
			(0.060)	(0.064)	(0.075)
period					−0.005
					(0.038)
R^2 overall	0.349	0.387	0.384	0.409	0.409

Robust standard errors in parentheses

1140 observations, 114 principals

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the baseline where the safe alternative is chosen and the lottery yields the high outcome. In each specification the counterfactual evaluation is significant. We focus on the specification in column five, which includes the controls. As can be seen with the highly significant positive coefficient of \$2.51 on $risky \times high$ ($p < 0.01$) and negative coefficient if −\$1.92 on $risky \times low$ the treatment effect of the lottery outcome has a large effect on principal's response when the agent chooses the risky alternative, with the difference \$4.42 indicating the size of the effect of having a favorable vs. unfavorable counterfactual comparison.[16] When the agent chooses the safe alternative, the lottery outcome also has a substantial and highly significant ($p < 0.01$) impact of \$1.67 on the principal's payment to the agent, despite the fact that the outcome of the lottery does not affect the payoffs of the principal.

Blame is not blind

Figure 1 shows period by period variability beyond the main effect of choice and outcome that we have just identified. An examination of the coefficients of the other variables, in particular $E[utility\ premium]$ shows the source of this variability: principals take into account the quality of choice made by the agent, making the payment increasing in the difference between

[16] A Wald test confirms that this difference is significant ($p < 0.01$).

their expected utility from the chosen alternative and the unchosen one. The 0.489 coefficient on $E[utility\ premium]$ is significant, and since principals are nearly risk-neutral this indicates that principals transfer approximately half a dollar for every dollar gain in expected value, when controlling for other factors. On the other hand, the significant coefficient on *utility* indicates principal's are significantly influenced by their terminal payoff when determining the amount to assign to the agent.[17]

Blame is not disappointment

A possible explanation of the deviation of payments from the values predicted by the control principle is that principals are disappointed by the outcome, and express their negative affective response transferring it onto others, regardless of their involvement in the choice. In our experiment, such an indiscriminate negative affective response should be expressed by lowering the payment of all parties involved, agent and third party. In Figure 2, we can see this is not the case: rather than lowering the third party's payment when they are in a counterfactually bad state, the principal uses them as a buffer, transferring more to the third party.

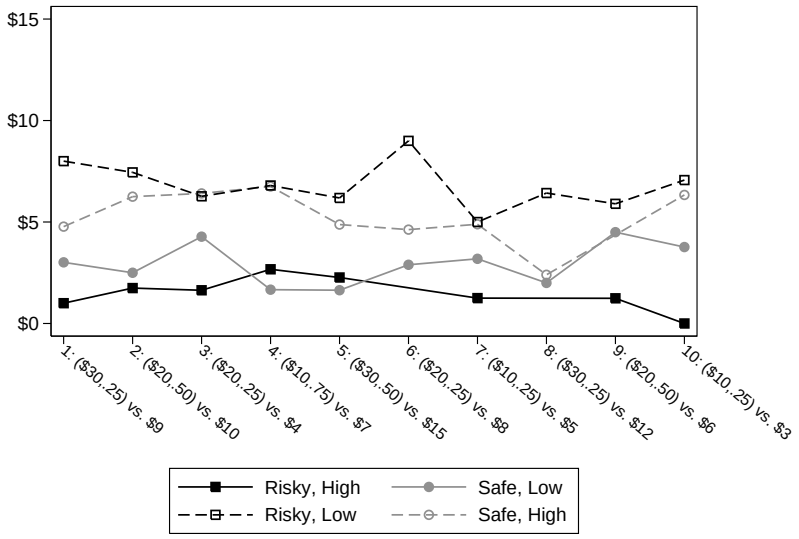


Figure 2: *Principals' average allocation to the third party (Allocate treatment)*

This conclusion is confirmed by the regression relating to the principal's payment to the third party in Table 3 below, which parallels the one for the payment to the agents in Table 2. We restrict our discussion to the specification with all control variables, column five of each table. The

[17] These regression results are robust to very different ways of measuring the desirability/quality of the chosen alternative. Alternate specifications of which are discussed in Appendix D.4

coefficients for a counterfactually good choice are all negative for the third party; in state (S, ℓ) the coefficients in each regression are of almost equal size, and opposite sign to those for the agent; in state (R, h) the \$4.42 increase in the transfer to the agent relative to state (R, ℓ) is balanced by a somewhat smaller \$3.10 reduction to the third party. In summary, principals use the transfer to the third party as a buffer (they prefer, everything else being equal, to transfer money from the experimenter to other subjects) and the payment to the agent as a way of expressing blame.

Table 3: *Estimated effects on the principal's payment to the third party (Allocate treatment)*

	(1)	(2)	(3)	(4)	(5)
constant	4.900***	5.941***	5.424***	6.168***	6.208***
	(0.317)	(0.402)	(0.338)	(0.404)	(0.429)
safe × low	−1.959***	−2.045***	−1.733***	−1.830***	−1.830***
	(0.272)	(0.264)	(0.253)	(0.247)	(0.247)
risky × low	1.916***	0.876*	1.560***	0.795	0.794
	(0.492)	(0.520)	(0.480)	(0.510)	(0.510)
risky × high	−2.944***	−1.888***	−3.163***	−2.310***	−2.301***
	(0.351)	(0.355)	(0.352)	(0.379)	(0.388)
utility		−0.127***		−0.099***	−0.100***
		(0.027)		(0.028)	(0.029)
E[utility premium]			−0.351***	−0.305***	−0.298***
			(0.064)	(0.068)	(0.081)
period					−0.008
					(0.042)
R^2 overall	0.176	0.186	0.191	0.197	0.197

Robust standard errors in parentheses
1140 observations, 114 principals
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Comparative Statics

An ancillary prediction of the delegated expertise model is that the ratio of the average payments to the agent between the two states counterfactually better for the principal, $w(R, h)/w(S, \ell)$, should be negatively correlated with the probability the lottery yields a high outcome, $P(h)$. In each period we computed the average payments for the group of principals in state (R, h) and divided it by the average for the subjects in state (S, ℓ). Figure 3 shows that data match this predicted pattern for the 8 periods where there were a sufficient number of observations to compute the ratio. We conducted a regression of the ratio of averages on the probability of a high lottery outcome to see if they were inversely related. As we noted in Section 2.2, we did not expect the number of

observations to be equal in each state, and they were not.[18] In order to address this issue in the regression, we weighted the ratio of averages in each period with an analytic weight proportional to the number of observations in the state with the fewest number of observations. Using this test we found that the relationship is significantly negative ($p < 0.01$).

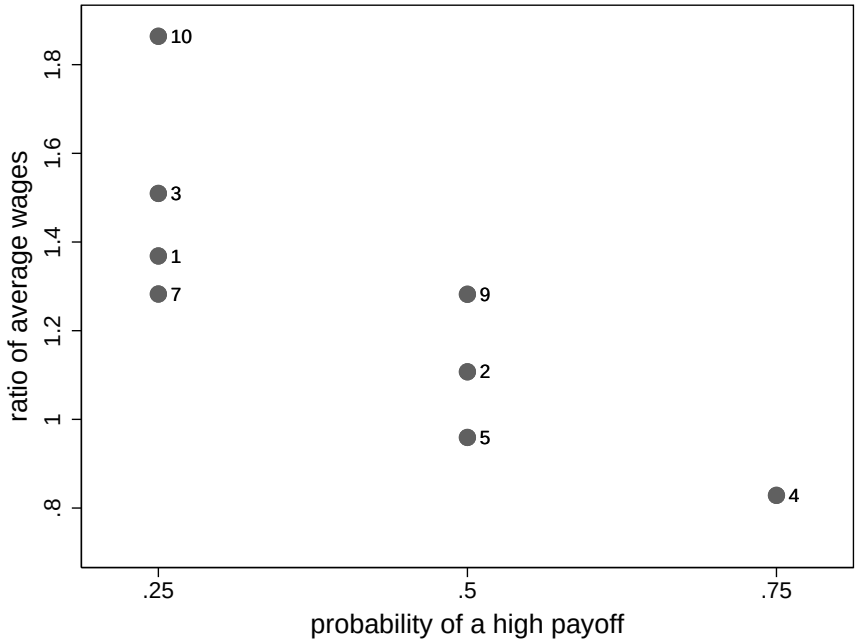


Figure 3: *A plot of the ratio of the average payments in each counterfactually favorable state, $w(R,h)/w(S,\ell)$, against the probability of the lottery yielding a high outcome, $P(h)$, for each period where data is available.*

2.4 Robustness to incentives and learning

While the experimental evidence is strong that the principals reward agents in a way consistent with unjustified blame and the prediction of the delegated expertise principal-agent model, there might remain some ambiguity with regards to what exactly has been identified. The design feature which allows the principal to assign payment to the agent without incurring any cost has advantages in terms of experimental control, but may present some difficulties for inference. Without direct personal financial incentives, the subjects in the role of principal might just be minimizing their decision cost by correlating their forced-responses to whatever feature of their environment is most salient, principally, how they feel. Therefore, while the consequence of the principal's decision is economically meaningful to the agent, the principal may be indifferent to this, treating the payment

[18] In fact, Periods 6 and 8 did not have any observations of R, h so we could not compute the average $w(R,h)$.

task as simply a measurement scale for their emotions, rather than a decision with respect to how much they want to assign to the agent. In order to investigate this possibility, we performed additional experiments which made the payments costly to principals and extended the number of decision periods from 10 to 25 in order to see if allowing more time to learn moderated the effect.

Design

The two additional experimental conditions termed here *Pay* and *Pay*×3 treatments have two modifications of the Allocate treatment: they make the principal's payment decision costly, and have a larger number of periods. The new treatments, with a total of 250 subjects, were conducted in separate sessions at the University of Minnesota with the same experimenter, laboratory, and subject recruitment methods as in the Allocate treatment.[19] Both conditions had 25 periods rather than 10 periods, and each of the first 10 periods involved the same risky and safe alternatives in the same order as the Allocate treatment, while the risky and safe alternatives in the final 15 periods were new (see Table 8 for a the complete list of alternatives). As in the sessions involving the Allocate treatment, a 10-period preference elicitation task was conducted as the final task of the session using the same lottery and safe alternatives as the first 10 periods of the experiment.[20] Exactly one of the 35 periods in the session was chosen at random to count for payment. In the Pay and Pay×3 treatments the design was exactly as outlined in Section 2.1 for the Allocate treatment, the principal engaged in an allocation task where they could assign payments to the agent and a third party (up to \$15 total), the only difference being that each dollar assigned had a direct cost to the principal. In the Pay treatment the cost of each \$1 assigned was \$1, while in the Pay×3 treatment the cost was \$0.33.[21] Understanding the costs of assigning a given payment in the Pay×3 treatment involved some calculation, so an on-screen cost calculator was provided to the principals.[22] In both conditions, after the on-screen video instructions were delivered, subjects were assigned to their fixed roles, given a chance to practice for one period, and then tested on their understanding of how their choice, the choice of their partner, and the roll of the 4-sided die determined joint payoffs.

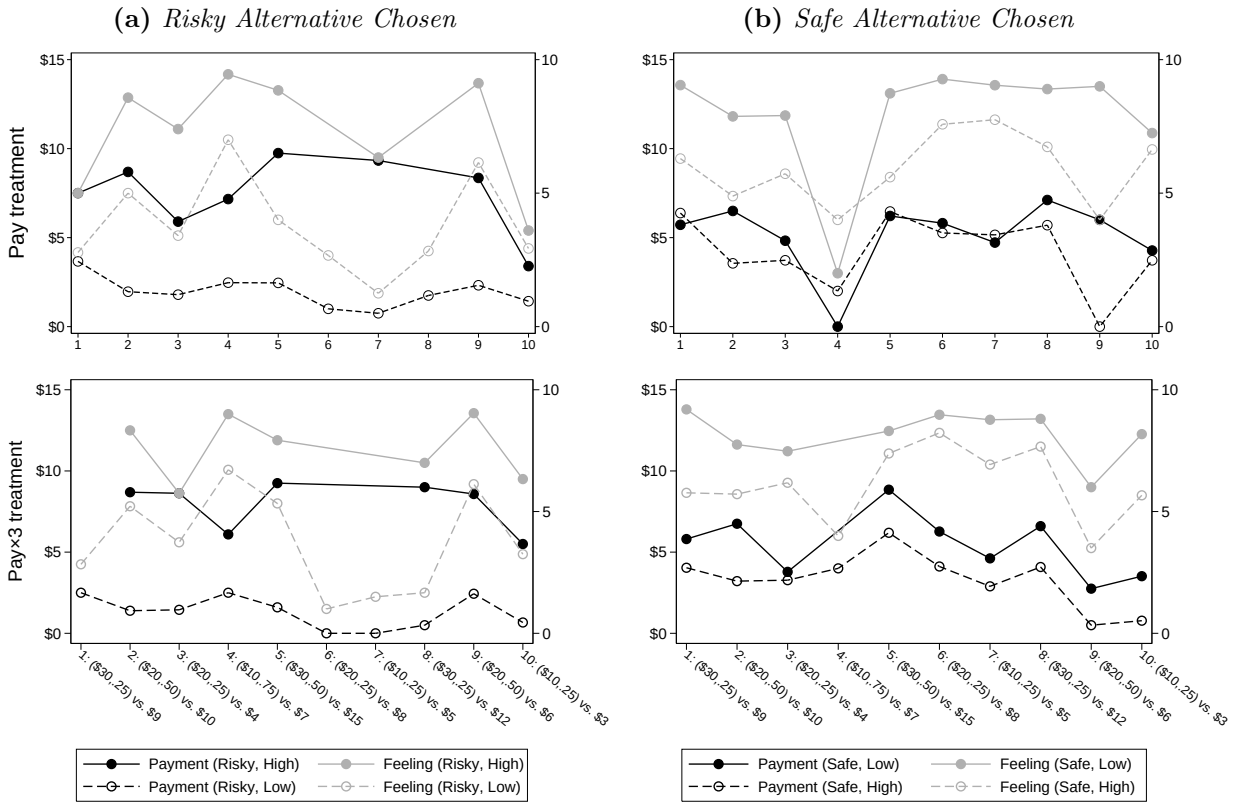
[19] The Pay treatment had 140 subjects and the Pay×3 treatment had 110 subjects. See Appendix Section A for a complete description of the new sessions.

[20] The preferences of principals did not differ significantly (See Appendix Section C.1).

[21] In the Pay treatment, in each period the principal was endowed with a \$15 balance in addition to the payoff received from the agent's choice and kept any amount that was not assigned to the agent or third party. For the Pay×3 treatment, the principal was endowed with a \$5 balance, keeping any unassigned amount as well.

[22] See online Appendix Section D.3 for a screen shot.

Figure 4: *Principals' average payment to their agents and feeling towards their agents' choice. The left vertical axis is the amount the principal transfers to the agent, and the right vertical axis is the principal's feeling towards the agents' choice.*



Robustness to Incentives

In the Pay and Pay×3 treatments the allocation task was made costly for subjects in the role of principal, who now faced a direct cost in expressing blame or praise, instead of the opportunity cost of allocating free (for him) funds to the agent rather than to the third party.

While the addition of a direct cost to the allocation task should not influence a principal's subjective rating of their agent's choice, it must reduce the amount they transfer—as long as subjects in the role of principal prefer more money to less. This narrowing of the response range combined with the cost of adjusting the payment itself means that any potential difference in the payments due to the counterfactual comparison will be smaller. This makes the Pay and Pay×3 treatments less powerful tests for detecting the effect of the counterfactual comparison on behavior. Nevertheless, as the regression further below indicates, the effect remains significant.

For the purpose of comparability, we first restrict the analysis to the first 10 (out of 25) periods in the Pay and Pay×3 treatments so that the sequence of decisions are identical to the Allocate treatment. As can be seen in Figure 4a, when the agent chooses the risky alternative the outcome of

the lottery affects the payment principals assign to the agent and their feeling towards the agent's choice in a way similar to the Allocate treatment. As expected, the subjects find the allocation task costly and the magnitude of the transfer is reduced, while the subjective reports do not differ markedly.[23] The pattern in subjects' subjective reports towards their agent's choice matches that of their payment, which suggests a common connection to the counterfactual comparison.

The addition of a cost to the allocation task introduces a potential confound between blame and the income level: if the lottery doesn't pay off after the agent has chosen it, subjects may simply be less willing to allocate payments to the agent or third party as it involves sacrificing the only payoffs they have. This confound was not present when the risky alternative was chosen in the Allocate treatment. To address this issue we focus on the principals that were randomly assigned to an agent who chose the safe alternative: in this case the income level is controlled for between subjects and, as discussed in Section 2.3, we can also control for the influence of distributional preferences, consequence-based reciprocity and arbitrary numeric anchoring. As can be seen in Figure 4b, when considering the counterfactual choice of the lottery, in both treatments principals consistently pay more to their agents and feel better about the choice when the lottery doesn't yield the high payoff.[24]

Since it is unlikely that both an undesirable alternative is chosen and an improbable lottery outcome occurs, the Pay and Pay×3 treatments have, as in the Allocate treatment, many periods which do not have a sufficient number of observations for a within-period between-subjects statistical test. In addition, with a higher standard error, lower sample size, and the aforementioned narrowed range of response, these designs are even less powerful for detecting a significant effect within a given period.[25] As a consequence, few single periods demonstrated a significant effect related to the experimentally manipulated variables. Despite this fact, as we have seen in Figure 4b, principals pay more to the agent for the counterfactually better outcome in most periods of the Pay treatment and in *all* periods of the Pay×3 treatment. The consistency of this relationship across periods suggests the effect of the lottery outcome is causal and that this can be identified if we pool the data from all 10 periods together and perform the same random effects regression as done for the Allocate treatment, i.e. controlling for the factors that depend on the decision-specific

[23] Surprisingly, the actual amount assigned when the agent chose the lottery and it paid off was at a comparable level between the Pay and Pay×3 treatments, even though any given payment was triple the cost in the Pay treatment.

[24] In Period 4 of the Pay treatment they appear to feel worse and assign less when better off in the counterfactual sense. As can be seen in Table 11 of Online Appendix Section D.5, there is only one observation in each condition in Period 4.

[25] See Tables 11 and 13 of Appendix Section D.5 for summary statistics.

parameters of each period—the potential payoffs and their probabilities.

In Table 4 and 5 we report the regressions from the Pay and Pay×3 treatments, where as before $safe \times low$ is the reference category for the experimentally manipulated variables, $safe \times high$, $risky \times low$, and $risky \times high$. For all specifications the estimation results are qualitatively similar to those reported in Table 2 for the Allocate treatment, but with two differences that can be highlighted by focusing the specification from each table that controls for the decision period (column five). The first difference with respect to the Allocate treatment is that subjects appear to assign less over time, approximately \$0.10 less in each period for both the Pay and Pay×3 treatment, indicating that subjects not only find the allocation task costly, but increasingly so. The second difference, as anticipated from the narrowed range of response, is that the (significant) effect from the counterfactual comparison is smaller in size. When the safe alternative is chosen, controlling for period-specific factors, the effect of the counterfactual comparison on the amount principals assign to the agent is highly significant ($p < 0.01$). The effect of being counterfactually better off when the safe alternative is chosen leads to an extra \$0.79 and \$1.07 being assigned to the agent in the Pay and Pay×3 treatments respectively, compared to the extra \$1.67 estimated from the identical specification in the Allocate treatment. As can be seen from the size of the difference in the coefficients on $risky \times high$ and $risky \times low$, the effect of winning the lottery, controlling for its payoff and quality, is substantially smaller in both the Pay and Pay×3 treatment, approximately \$2.51 and \$1.72 respectively, compared to \$4.42 in the Allocate treatment.[26]

As in the Allocate treatment the coefficient on the variable $E(utility\ premium)$, the proxy measure for the quality of the decision, is positive and significant overall in both treatments, albeit half the size.[27] This shows principals continue to account for the quality of the agents choice when assigning the payment. With the significant positive coefficient on the variable *utility*, the payment also continues to be sensitive to the utility of payoff and of comparable magnitudes.

Also, as in the Allocate treatment, there is a possibility that subjects in the role of principal are simply disappointed with the outcome and transfer this negative affect onto both the agent and the third party. This possibility was ruled out in the Allocate treatment by the fact that the third party was treated as a buffer, i.e. subjects assigned *more* money to the third party when they were worse off (counterfactually). In the Pay and Pay×3 treatments the payment assigned to the third party is small and displays no particular dependence on the counterfactual comparison,

[26] This should not be surprising, as subjects are using their own money in the allocation task.

[27] A specification which measures the quality of the decision according to the log-odds of the rate at which it was chosen in the population of principals can be found in Tables 19 and 20 of Online Appendix Section D.

Table 4: *Estimated effects on the principal's payment to the agent (Pay treatment, first 10 periods)*

	(1)	(2)	(3)	(4)	(5)
constant	4.954***	3.260***	4.567***	3.056***	3.490***
	(0.429)	(0.418)	(0.425)	(0.418)	(0.452)
safe × low	0.684**	0.820***	0.624**	0.769**	0.794**
	(0.318)	(0.318)	(0.306)	(0.310)	(0.314)
risky × low	−2.817***	−1.118***	−2.563***	−1.014**	−1.014**
	(0.375)	(0.405)	(0.376)	(0.406)	(0.412)
risky × high	2.937***	1.249***	3.045***	1.414***	1.500***
	(0.319)	(0.319)	(0.311)	(0.318)	(0.331)
utility		0.207***		0.197***	0.188***
		(0.026)		(0.026)	(0.026)
E[utility premium]			0.234***	0.174***	0.233***
			(0.053)	(0.053)	(0.057)
period					−0.087**
					(0.038)
R^2 overall	0.196	0.225	0.203	0.229	0.231

Robust standard errors in parentheses

700 observations, 70 principals

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: *Estimated effects on the principal's payment to the agent (Pay×3 treatment, first 10 periods)*

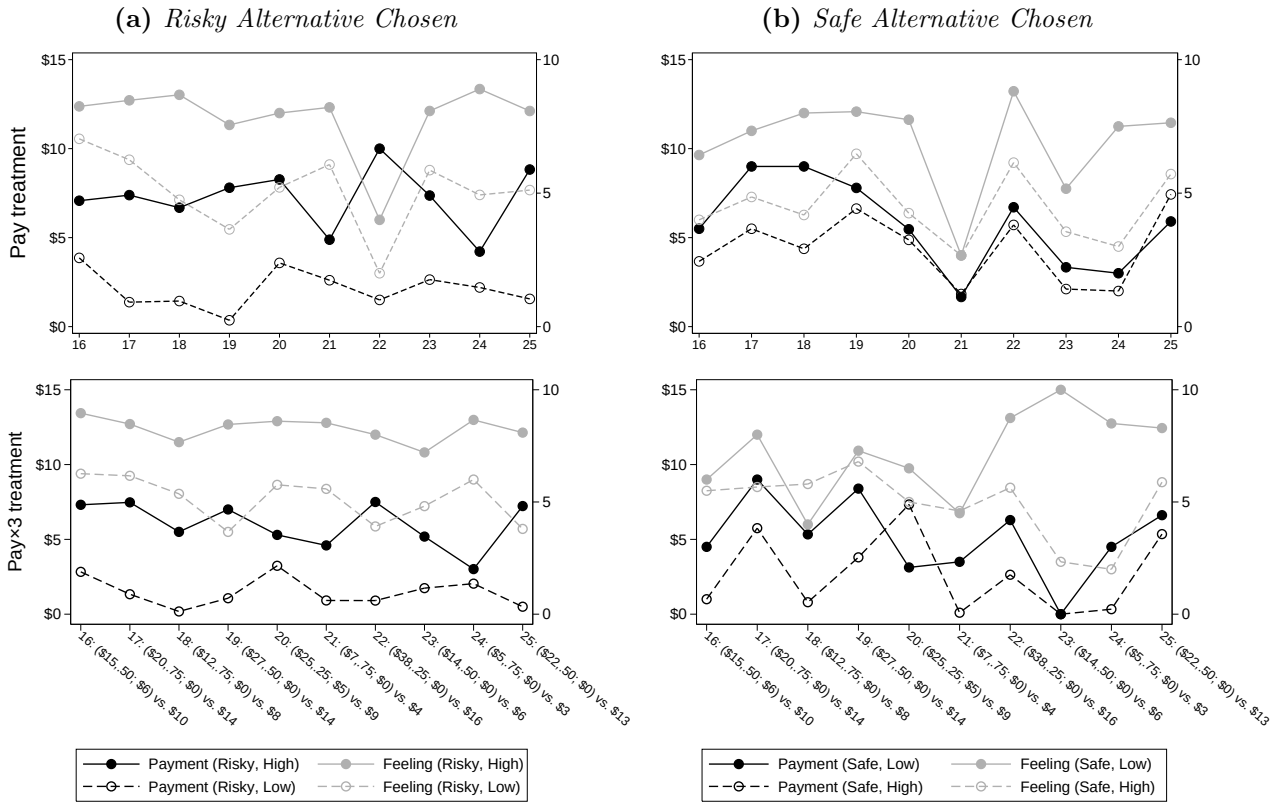
	(1)	(2)	(3)	(4)	(5)
constant	4.270***	1.944***	3.869***	1.758***	2.307***
	(0.531)	(0.560)	(0.509)	(0.541)	(0.611)
safe × low	0.978**	1.192***	0.803*	1.053**	1.069**
	(0.437)	(0.413)	(0.439)	(0.422)	(0.421)
risky × low	−2.771***	−0.445	−2.500***	−0.356	−0.386
	(0.543)	(0.563)	(0.532)	(0.556)	(0.553)
risky × high	3.417***	1.096**	3.482***	1.252**	1.330**
	(0.580)	(0.516)	(0.575)	(0.509)	(0.519)
utility		0.286***		0.273***	0.263***
		(0.044)		(0.043)	(0.043)
E[utility premium]			0.291***	0.214***	0.268***
			(0.078)	(0.077)	(0.080)
period					−0.103**
					(0.043)
R^2 overall	0.208	0.268	0.217	0.272	0.275

Robust standard errors in parentheses

550 observations, 55 principals

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 5: *Final* 10 *periods, Principals' average payment to their agents and feeling towards their agents' choice. The left vertical axis is the amount the principal transfers to the agent, and the right vertical axis is the principal's feeling towards the agents choice.*



as can be seen in Figure 7 of Appendix Section D.5.[28] This implies that when responding to the counterfactual comparison the principals focus on the agent and ignore the third party, i.e. they blame.

Robustness to Learning

Now that we have established a significant and sizable effect of counterfactual evaluations when the payments are costly, it is important to test whether this effect is persistent over time, or if subjects appreciate that in their environment the payment, being ex-post, has no effect, and should therefore set the costly transfer to zero. To test for this persistency in the Pay and Pay×3 treatments the number of decision periods was extended from 10 to 25. In the experimental environment, holding the agent responsible in the counterfactual comparison is a mistake in the sense that it does not reveal any information about hidden actions. Therefore, as the environment becomes more familiar,

[28] An analogous regression reported in Table 3 of Section 2.3 regarding the Allocate treatment, found no significant relationship between the outcomes and the payment to the third party.

the salient perturbation-the principal agent model-should become less salient, and subjects should become less sensitive to the counterfactual comparison. This is not what we see. While subjects do assign less to the agent over time, if anything the sensitivity to the counterfactual comparison is stronger in the final 15 periods of the Pay and Pay×3 treatments. A casual inspection of the graph of the payment and feeling towards the agent choice in Figure 5 shows the counterfactual effect maintains in the final 10 periods of the session. In Table 6 we report the results of a random effects regression estimated as described earlier for the respective Pay and Pay×3 treatments, but extended to all 25 periods of the experiment. We introduce the indicator variable $(Period > 10)$ which is equal to one when the decision period is between 11 and 25 and zero otherwise. As can be seen in Table 6, for each treatment the coefficient on $safe \times low$ remains significant. The coefficients pertaining to $(Period > 10)$ and its interactions indicate that this effect is somewhat stronger in the later periods. The negative coefficient on $(Period > 10)$ lowers the intercept, i.e. the payment to the agent in the counterfactually bad reference state where the safe alternative is chosen and the lottery yields the high outcome. The positive coefficient on $safe \times low \times (Period > 10)$ increases the payment to the agent in the counterfactually good state where the safe alternative is chosen and the lottery yields the low outcome.[29] These results maintain if we run the regression separately on the final 10 periods of the session, reported in Table 17 of Online Appendix Section D.

Comparative Statics

The comparative statics prediction of a negative correlation between the ratio of the average payments in each state and the probability the lottery yields a high outcome, discussed in Section 2.2 and supported in the Allocate treatment, had weaker support in the Pay and Pay×3 treatments. We performed a regression of the ratio $w(R,h)/w(S,\ell)$ on the probability the lottery yields a high outcome $P(h)$, analogous to that performed using the data from the Allocate treatment. For each of the two additional treatments this regression was performed with analytic weights for each period again proportional to the number of observations in the state with the fewest observations. In the Pay treatment, with 21 observations, the slope coefficient, though negative, was indistinguishable from zero ($p = 0.961$). In the Pay×3 treatment, with 20 observations, the slope coefficient was negative and nearly significant ($p = 0.13$). If we restrict the regressions from each treatment to periods where the minimum number of observations in each state is at least 10, the Pay treatment remains insignificant, but in the Pay×3 treatment the slope coefficient becomes highly significant,

[29] The difference in the coefficient values on $risky \times low$ and $risky \times high$ are also significant with a Wald test ($p < 0.01$). See Tables 15 & 16 in Online Appendix Section D for additional specifications

Table 6: *Estimated effects on the principal's payment to the Agent (Pay and Pay×3 treatments, all* 25 *periods)*

	Pay treatment			Pay×3 treatment		
	(1)	(2)	(3)	(4)	(5)	(6)
constant	3.240***	3.169***	3.601***	2.443***	1.864***	2.387***
	(0.389)	(0.425)	(0.453)	(0.484)	(0.544)	(0.632)
safe × low	1.083***	0.694**	0.718**	1.258***	0.871*	0.889*
	(0.283)	(0.302)	(0.305)	(0.332)	(0.459)	(0.461)
risky × low	−0.429	−1.208***	−1.208***	−0.000	−0.277	−0.302
	(0.314)	(0.443)	(0.446)	(0.429)	(0.542)	(0.545)
risky × high	1.047***	1.327***	1.412***	1.057***	1.031*	1.107**
	(0.216)	(0.320)	(0.333)	(0.322)	(0.527)	(0.535)
utility	0.216***	0.194***	0.186***	0.249***	0.274***	0.265***
	(0.019)	(0.026)	(0.026)	(0.033)	(0.043)	(0.043)
E[utility premium]	−0.005	0.171***	0.228***	0.014	0.204***	0.255***
	(0.029)	(0.055)	(0.062)	(0.043)	(0.075)	(0.079)
Period	−0.034**		−0.086**	−0.055***		−0.098**
	(0.014)		(0.039)	(0.018)		(0.043)
$(Period > 10)$		−1.841***	−1.239*		−2.061**	−1.292
		(0.527)	(0.655)		(0.891)	(0.956)
Period × $(Period > 10)$			0.028			0.021
			(0.045)			(0.048)
safe × low × $(Period > 10)$		0.709*	0.689*		0.669	0.656
		(0.401)	(0.399)		(0.537)	(0.535)
risky × low × $(Period > 10)$		2.206***	2.275***		1.913*	2.114**
		(0.652)	(0.651)		(1.012)	(1.010)
risky × high × $(Period > 10)$		1.535**	1.581**		2.889**	3.097**
		(0.709)	(0.721)		(1.216)	(1.235)
utility × $(Period > 10)$		0.130**	0.138***		0.122	0.140*
		(0.052)	(0.052)		(0.081)	(0.081)
E[utility premium] × $(Period > 10)$		−0.246***	−0.279***		−0.340**	−0.374**
		(0.066)	(0.071)		(0.145)	(0.151)
utility × risky × $(Period > 10)$		−0.148***	−0.154***		−0.237***	−0.258***
		(0.056)	(0.055)		(0.081)	(0.083)
E[utility premium] × risky × $(Period > 10)$		−0.122	−0.099		−0.170	−0.137
		(0.075)	(0.073)		(0.121)	(0.129)
R^2 overall	0.204	0.208	0.210	0.207	0.217	0.221

Robust standard errors in parentheses

Respectively 1750 & 1375 observations, 70 & 55 subjects

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

with $p < 0.0001$.

3 Conclusion

Principals in our experiment clearly exhibit unjustified blame: they adjust the payment to the agent according to the difference between the realized outcome and the outcome of the alternative the agent had not chosen, even though it is clear that he was not responsible for the outcome. The size of the blame effect is substantial relative to other possible influences like disappointment and the quality of the agent's decision.

Unjustified blame is clearly an operative factor when the agent chose the safe alternative: the principals made their payments to their agents contingent on a payoff-irrelevant event that the agents could not influence. The design of the Allocate treatment controlled for several confounding factors, which might otherwise have provided alternative explanations of the subjects' behavior, different from blame. The feature of having the principal assign money that they could not keep for themselves assured that the variation in payments cannot be explained by the principal's ability to pay, risk sharing motivations, or the principal's marginal utility of a transfer to the agent. Additionally, the assignment of payment to the random third party allowed us to rule out the possibility that with a low payoff, principals may simply be less willing to transfer to any party, either because of irritation with the payoff, or its influence on the perceived value of a transfer. The additional treatments that introduced a direct cost to the allocation task allowed us to rule out the possibility that this effect was an artifact driven by the ease of simply correlating their assigned payments to their subjective feeling, without any financial incentive to do otherwise.

If one takes as a benchmark of rational behavior the principle of holding others accountable only for the events for which they are responsible, then we may say that our subjects exhibited irrational behavior. Principals gave clear evidence that they were following a counterfactual evaluation based on events beyond the control of their agents, rather than following the merit and control principles. We claim that this blame is not irrational, and instead reflects deeply felt normative principles, which have wide application in society. We have already observed that blame was not blind, because principals took into account the quality of the decision of the agent, reducing blame when the ex-ante value of the chosen alternative was larger than the unchosen one, even when the final payoff was not. But the main support for our claim that blame is not irrational is the fact that the payment scheme implemented by our subjects was the same as the optimal payment in the

delegated expertise principal-agent problem, where the agent has to be paid a lower reward when the outcome is not favorable precisely in the counterfactual sense. In particular they have to be paid less if the unchosen alternative yields a high payoff, even when their choice induces a risk-free certain payment. The lower payment has to be paid even if it is clear to the principal, just as it was to our subjects, that the agents are not responsible for the outcome.

Though not optimal in our controlled experimental environment where all relevant actions are observed, blame may be optimal in more general environments. Blame may not be justified on the basis of what can be observed or inferred, but the common knowledge that it exists makes it a powerful incentive—a contract—and assures that more generally it will be justifiable: agents will have incentive to employ effort towards the mental and physical activities needed to benefit the principal, whether or not the principal understands or observes these activities. By doing so, blame implements a characteristic and counter-intuitive property of the optimal contract in the principal-agent model: that the payment is dependent on events that are outside the control of the agent, and in some cases, events that do not influence the principal's payoffs.

References

Baron, J., and J. C. Hershey (1988): "Outcome bias in decision evaluation," *Journal of Personality and Social Psychology*, 54(4), 569–579.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Bolton, P., and M. Dewatripont (2004): *Contract Theory*. MIT Press.

Charness, G. (2004): "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, 22(3), 665–688.

Charness, G., and D. I. Levine (2007): "Intention and Stochastic Outcomes: An Experimental study," *Economic Journal*, 117(522), 1051–1072.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), 817–869.

Converse, P. E. (1964): "The Nature of Belief Systems in Mass Publics," in *Ideology and Discontent*, ed. by D. E. Apter. Free Press of Glencoe.

Falk, A., E. Fehr, and U. Fischbacher (2008): "Testing theories of fairness–Intentions matter," *Games and Economic Behavior*, 62(1), 287–303.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fehr, E., and S. Gächter (2000): "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, E., and K. M. Schmidt (1999): "A Theory Of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114(3), 817–868.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

Gino, F., D. A. Moore, and M. H. Bazerman (2008): "No Harm, No Foul: The Outcome Bias in Ethical Judgments," Harvard Business School NOM Working Paper.

Harrison, G. W., and E. E. Ruström (2008): "Risk Aversion in the Laboratory," *Research in Experimental Economics*, 12, 41–196.

Hayek, F. A. (2011 (1960)): *The Constitution of Liberty*. The Unversity of Chicago Press.

Hirschman, A. O. (1970): *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.

Hölmstrom, B. (1979): "Moral Hazard and Observability," *Bell Journal of Economics*, 10, 74–91.

Holt, C., and S. Laury (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 1644–1655.

Jenter, D. C., and F. Kanaan (2011): "CEO Turnover and Relative Performance Evaluation," *Journal of Finance*, Forthcoming.

Kant, I. (1784): *Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Kleinig, J. (1971): "The Concept of Desert," *American Philosophical Quarterly*, 8(1), 71–78.

Mazzocco, P. J., M. D. Alicke, and T. L. Davis (2004): "On the Robustness of Outcome Bias: No Constraint by Prior Culpability," *Basic & Applied Social Psychology*, 26(2/3), 131–146.

Moulin, H. (2003): *Fair Division and Collective Welfare*. MIT Press.

Mowen, J. C., and T. H. Stone (1992): "An Empirical Analysis of Outcome Biases in Constituent Evaluations of Public Policy Decision Makers," *Journal of Public Policy & Marketing*, 11(1), 24–32.

Myerson, R. B. (1991): *Game Theory: Analysis of Conflict*. Harvard University Press.

Nelkin, D. K. (2004): "Moral Luck," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta. Stanford University.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281–1302.

Tan, H.-T., and M. G. Lipe (1997): "Outcome effects: The impact of decision process and outcome controllability," *Journal of Behavioral Decision Making*, 10(4), 315–325.

Wakker, P. P. (2008): "Explaining the Characteristics of the Power (CRRA) Utility Family," *Health Economics*, 17, 1329–1344.

Walster, E. (1966): "Assignment of responsibility for an accident," *Journal of Personality and Social Psychology*, 3(1), 73–79.

A Appendix: Experimental Sessions & Procedures

All experiments discussed herein were conducted at the Social and Behavioral Sciences Laboratory (SBSL) at the University of Minnesota using z-Tree experimental software (Fischbacher 2007). For each two-hour experimental session we first visited the last five minutes of a 200-500 student undergraduate principles of economics class and informed the students that they could participate in a two hour economic choice study conducted on the same day. They were also told that they would receive between \$8 and \$70 for their participation, with more than half of the subjects receiving between \$20 and \$40. Typically, 30-40 subjects were recruited and walked with the experimenter from their classroom to the laboratory. Upon arrival, students were randomly seated at individual computer carrel booths that isolated their screens from the view of other subjects and were given consent forms and payment receipts. Subjects were first informed only that they would participate in different decision environments–which we refer to as treatments from now on–after which they would be requested to respond to feedback questions and finally paid.[30] Subjects were told that we would be randomly selecting a decision period to count for payment and therefore it was important to make each decision as if it were the only decision they were paid for.[31] Next, the physical dies used to make the selection were displayed to them. After the preliminary overview of the session, the instructions were privately presented to subjects with a narrated Flash video (with headphones) streaming via a web browser and given a sheet presenting a summary of instructions, which they were given time to read over at the end of the video. Next, subjects were privately and permanently assigned (randomly) to one of two fixed roles, which we term here principal and agent. To familiarize participants with the user interface and reinforce the instructions, subjects participated in two unpaid periods of the experiment, with different parameters. To check for understanding, subjects were presented with five questions, where each incorrect response had customized feedback explaining why it was incorrect. Next, subjects participated in the main experimental treatments in their respective roles and then all subjects participated in a final preference elicitation task where they were presented with the same risky and safe alternatives the agent faced in the same order as the earlier experiments, and were instructed to choose between the alternatives for themselves. We

[30] The sessions with the Allocate treatment involved another experimental task not reported here, thus there were 10 periods of the Allocate treatment, 10 periods of the other experimental task, followed by 10 periods where subjects chose for the themselves. The behavior in the Allocate treatment did not differ significantly whether it preceded or followed the other experimental task, and thus the data was pooled together here.

[31] In the sessions with the Allocate treatment a 20-sided die was rolled to select one period from the twenty periods comprising the Allocate treatment and an experimental task not reported here, and a 10-sided die was rolled to select one period from the preference elicitation task. In the sessions with the Pay and Pay$\times$3 treatments a 4-sided and 10-sided die were rolled (rolling again if necessary) to select one period from the 35 periods that counted for payment.

term this preference elicitation task the Choose-For-Yourself experiment. In every session, after all tasks were completed the subjects typed open-ended response to feedback questions and then were paid. A total of 554 subjects participated.[32]

B Appendix: Blame and Delegated expertise

The game

There is a finite vector of assets $(A^i : i = 1, \ldots, n)$ with random returns $(Y^i : i = 1, \ldots, n)$; each Y^i taking values in a finite set. The distribution over returns is determined by the realization of a state of nature $\theta \in \Theta$, Θ a finite set, according to a given probability $\mu \in \Delta(\Theta)$. Also available is a safe asset S with fixed return y^0. The vector of returns of all assets is denoted $y = (y^i : i = 0, 1, \ldots, n) \in Y$. The probability of the return profile y given θ is denoted by $Q_\theta(y)$. An agent can choose an effort $e \in E$ and then observe a signal $x \in X$, X a finite set, drawn according to $P_{\theta e} \in \Delta(X)$. The effort e costs to the agent a utility cost $v(e)$. A non informative signal is available at minimum zero utility cost. Upon observing x the agent chooses an element $c \in \{0, 1, \ldots, n\} \equiv C$, that is one of the assets. A policy π adopted by the agent is a function $\pi : X \to \Delta(C)$. The principal can observe the realized return of all assets, and the agent's choice. A contract is a function $w : C \times Y \to R$ assigning to the agent a payment $w(c, y)$ at asset payoff y and choice c of the agent. A principal offers a contract w to the agent. After the choice c and outcome y he collects y^c and pays $w(c, y)$.

The time-line of the game is: first contract w is offered; the agent can accept or reject. If he accepts, he chooses a level of effort e and a policy π. Together they give the action $a \equiv (e, \pi)$, unobserved by the principal. Then θ is chosen according to μ and is not communicated to any of the two players; then x and y are determined given e and θ. The choice c with probability $\pi(\cdot; x)$ of the agent is implemented, and communicated to the principal, who gets the payment y^c and pays $w(y, c)$ to the agent. Then the game is over.

The choice of action a of the agent determines a distribution on C conditional on θ:

$$D^a_\theta(c) \equiv \sum_x P_{\theta e}(x)\pi(c; x) \tag{5}$$

[32] 304 subjects in the Allocate treatment, 140 subjects in the Pay treatment, and 110 subjects in the Pay×3 treatment.

and a joint probability $R(\cdot; a)$ on $Y \times C$, where

$$R(y, c; a) \equiv \sum_{\theta} \mu(\theta) Q_{\theta}(y) D_{\theta}^{a}(c). \tag{6}$$

The unconditional probability on Y is independent of the action of the agent:

$$\nu(y) \equiv \sum_{\theta} \mu(\theta) Q_{\theta}(y) \tag{7}$$

The probability R can be disintegrated as the product of the marginal on Y (independent of a) and a transition Γ from Y to C (which is dependent on a):

$$R(y, c; a) = \nu(y) \Gamma_{y}^{a}(c) \tag{8}$$

Note that the choice of the agent is really the choice of a correlation device $R(\cdot; a)$ over $Y \times C$ at the utility cost $v(e)$, using the policy π appropriately; and since the probability over Y is really beyond his reach he is really choosing the transition $y \to \Gamma_{y}^{a}$.

Optimal Contract Problem

In the standard reformulation of the problem, the principal is choosing the optimal contract w and action a of the agent subject to the incentive compatibility and individual rationality constraint of the agent. For a given utility G on net returns, the problem of the principal is

$$\max_{a,w} \sum_{y,c} R(y, c; a) G(y^{c} - w(c, y)) \tag{9}$$

subject to the IC constraint:

$$\forall a' = (e', \pi') : -v(e) + \sum_{y,c} R(y, c; a) U(w(c, y)) \geq -v(e') + \sum_{y,c} R(y, c; a') U(w(c, y)) \tag{10}$$

and the IR condition:

$$-v(e) + \sum_{y,c} R(y, c; a) U(w(c, y)) \geq 0 \tag{11}$$

where 0 is the value of the outside option. The constraint 10 includes the constraint that the agent chooses the desired effort, and that for any given signal chooses the desired asset.

When the wage compensation follows the control principle, that is $w(c, y) = \bar{w}(c)$, then if the cost of effort is strictly larger than zero for non zero effort the agent will provide zero effort. This is clear because for such a wage schedule the agent will face the problem

$$\max_a -v(e) + \sum_c \sum_y \nu(y)\Gamma_y^a(c)\bar{w}(c)$$

but any probability on the choice set C like $\sum_y \nu(y)\Gamma_y^a(c)$ can be achieved by the choice of a constant probability independent of x, hence the optimal choice of effort is the minimum cost of effort.

One risky asset

This is the simplest possible case, and is particularly interesting for us because it is the environment used in the experiment. Here $n = 1$, with one risky asset R in addition to the existing safe asset S. The states are bad and good, $\Theta \equiv \{\beta, \gamma\}$, while the effort choice is binary with no effort or effort $E \equiv \{ne, e\}$ (ne is "no effort"), with resulting signals being bad or good, $X \equiv \{b, g\}$. The payoffs for the risky asset are low or high, $Y^1 = \{\ell, h\}$, with $h > y^0 > \ell$ and the choice set for the agent is $C \equiv \{S, R\}$. We know already that either $w(R, h) \neq w(R, \ell)$ or $w(S, h) \neq w(S, \ell)$. We now want to show that they are both different, and there is a special order among the compensations, which is the same that we find in the data.

The $P_{\theta ne}$ is completely uninformative (the probability over signals is an arbitrary probability independent of θ), whereas $P_{\theta e}$ is informative. Denote $P_\mu(x)$ the total probability of the signal x, and $P(y|x)$ be the probability of the return y if the informative signal (when effort is provided) is x. Simple computations show that the wage to the agent without an outside option satisfies:

$$w(S, h) = w(R, \ell) = 0 \tag{12}$$

The two positive wages $w(S, \ell), w(R, h)$ are determined by the condition that the return to the choice of the non informative signal is equal to the maximum between the expected unconditional payment from the risky and the safe choice. To make the comparison with the payoff from the choice of the high effort, we rewrite these two payoffs as the payoffs of an agent that pays no effort cost, observes the informative signal but cannot condition his choice on the signal. We get that the

payoff from the choice of the non-informative signal is equal to:

$$\max\{(P_\mu(g)P(h|g) + P_\mu(b)P(h|b))w(R,h), (P_\mu(g)P(\ell|g) + P_\mu(b)P(\ell|b))w(S,\ell)\} \tag{13}$$

The two terms must be equal at the optimal contract. Also the net utility to the agent from the choice of the informative signal is:

$$P_\mu(g)P(h|g)w(R,h) + P_\mu(b)P(\ell|b)w(S,\ell) - v(e) \tag{14}$$

The incentive condition for the agent to choose R at g and S at b is:

$$P(hg)w(R,h) > P(lg)w(S,l);\ P(lb)w(S,l) > P(hb)w(R,h); \tag{15}$$

where for example $P(hg)$ is the probability of the event (h,g) at the high effort.

Equating (13) and (14) we find that $w(S,\ell), w(R,h)$ are the solution of the system:

$$P(\ell|b)w(S,\ell) - P(h|b)w(R,h) = \frac{v(e)}{P_\mu(b)} \tag{16}$$

$$-P(\ell|g)w(S,\ell) + P(h|g)w(R,h) = \frac{v(e)}{P_\mu(g)} \tag{17}$$

two positive numbers by condition 15 above. The wage is uniquely determined if the determinant $D \equiv P(\ell|b)P(h|g) - P(h|b)P(\ell|g)$ of the matrix defining the system of Equations (16) and (17) is not zero; more precisely

$$w(S,\ell) = \frac{1}{D}\left[P(h|g)\frac{v(e)}{P_\mu(b)} + P(h|b)\frac{v(e)}{P_\mu(g)}\right]$$

and

$$w(R,h) = \frac{1}{D}\left[P(\ell|g)\frac{v(e)}{P_\mu(b)} + P(\ell|b)\frac{v(e)}{P_\mu(g)}\right]$$

Comparative statics

We have seen that a positive payment is made only at (S,ℓ) and (R,h): which one is higher? Let us consider the case in which the signal is symmetric, $P_{\gamma e}(g) = P_{\beta e}(b)$ and $\mu(\beta) = 1/2$, so in particular the ex-ante probability of the good signal g and the bad signal b are the same. In this case the

ratio

$$\frac{w(R,h)}{w(S,\ell)} = \frac{P(\ell|g) + P(\ell|b)}{P(h|g) + P(h|b)} = \frac{P(\ell)}{P(h)}, \tag{18}$$

where $P(y)$ is the ex-ante probability of the outcome y. As intuitively clear, the wage for choosing the risky asset and high outcome is higher than the wage for choosing the safe asset and low outcome when the high outcome is relatively more rare than the low outcome.

C Appendix: Additional Material

C.1 Principal Choices

The preferences of principals towards the alternatives they faced in the first 10 periods are similar across experimental conditions. In Figure 6 the proportion of safe choices the principals made is graphed in each period, by experimental condition. As can be seen, in periods $2, 3$ and 10 the safe option is chosen slightly more often in the sessions involving the Pay treatment.

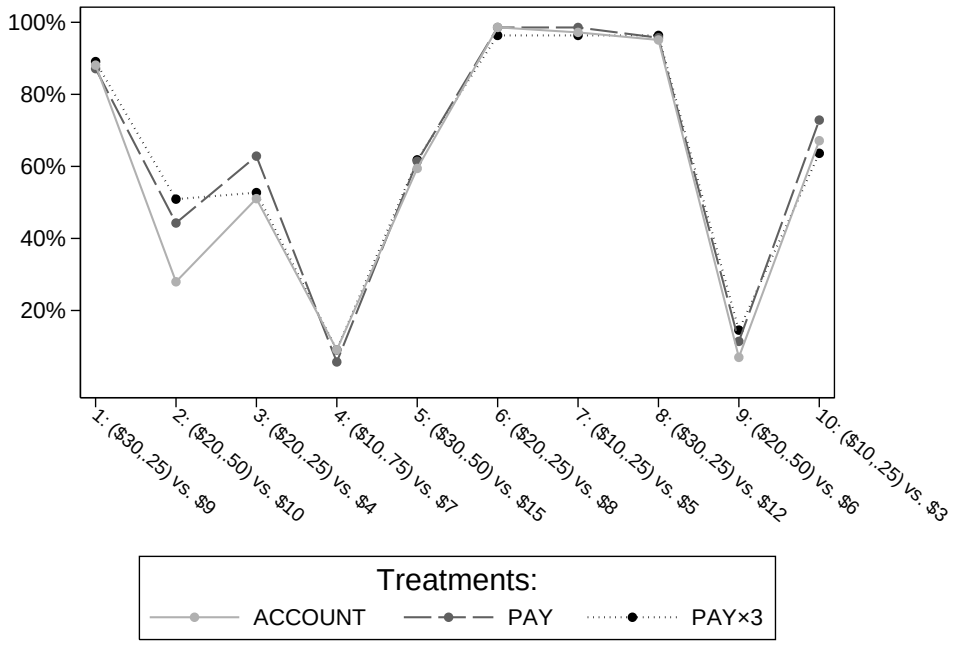


Figure 6: *The percentage of safe choices by the principal for each period, in each treatment*

The parameters of the choices over the ten decision periods do not provide a great variety of risky and safe alternative comparisons, and thus if principals have identical risk aversion and do not make mistakes, the choice data cannot be granular enough to identify risk-aversion parameters with the level of precision that a dedicated risk elicitation task would, such as the commonly used multiple-price list task of Holt and Laury (2002). If we adopt a discrete choice modeling approach, and view the principals' decisions for themselves as being generated by a binomial choice model with an underlying parametric utility function, then more information about risk aversion can be recovered from the frequency in which the alternatives are chosen. We explore three versions of a binomial choice model of the probability of choosing the safe option, all with an underlying

constant relative risk aversion (CRRA) utility function.[33] In two versions of the model, a latent index is parameterized by the risk aversion (α) parameter which determines the utility function $u(x) = x^{1-\alpha}/(1-\alpha)$, and the error parameter (σ), which determines the latent index, scaling the expected utility premium of choosing the safe option:

$$\text{Latent Index} = \frac{u(\text{c}) - pu(h)}{\sigma}.$$

The model is completed with a function, either logistic or gaussian, linking the index to the probability of choosing the safe option. In the third version of the binomial choice model, the probability of choosing the safe alternative takes the Luce Ratio structure, namely $u(\text{c})^{1/\sigma}/(u(\text{c})^{1/\sigma} + (pu(h))^{1/\sigma})$, where $u(\cdot)$ is a $CRRA$ utility function, and σ determines the error rate(Harrison and Ruström 2008).

We pool the data together from all three treatments to see if there is a significant difference in the estimated risk aversion, where the risk aversion parameter is $\alpha = \alpha_0 + \alpha_1 \mathbb{1}_{[Pay]} + \alpha_2 \mathbb{1}_{[Pay\times 3]}$. In Table 7 we present the results of the pooled maximum likelihood estimation for logistic and gaussian link functions, as well the Luce Ratio model. As we can see, the principals are nearly risk-neutral and higher levels of risk aversion in the Pay and Pay×3 treatments are not significant.

Table 7: *CRRA risk aversion MLE estimation for principals by link function (binomial choice model)*

	Probability Structure		
	Logistic	Gaussian	Luce Ratio
risk aversion (α)			
constant (α_0)	0.031*	0.039**	0.056***
	(0.017)	(0.018)	(0.019)
Pay treatment (α_1)	0.057*	0.055	0.056*
	(0.034)	(0.034)	(0.033)
Pay×3 treatment (α_2)	0.057	0.057	0.034
	(0.037)	(0.037)	(0.037)
error (σ)			
constant	1.051***	1.889***	0.187***
	(0.056)	(0.096)	(0.009)

Clustered standard errors in parentheses

2680 observations, 268 principals, 10 Periods

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[33] We choose CRRA as it is the most commonly used parametric family for fitting utility functions to data (Wakker 2008).

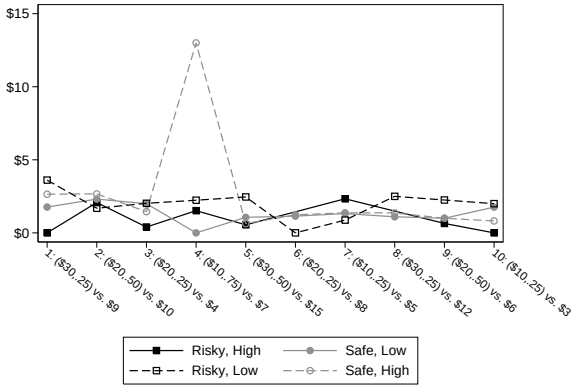
C.2 Figures and Tables

Table 8: *The binary decision problem for each period and the agents' choices for principals for each treatment*

	Alternatives		Agents' choices (safe)		
Period	Risky(R)	Safe(S)	Allocate	Pay	Pay × 3
1	($30,0.25; $0)	9	0.84	0.86	0.89
2	($20,0.5; $0)	10	0.18	0.24	0.20
3	($20,0.25; $0)	4	0.51	0.61	0.58
4	($10,0.75; $0)	6	0.06	0.03	0.04
5	($30,0.5; $0)	15	0.50	0.54	0.47
6	($20,0.25; $0)	8	0.95	0.96	0.95
7	($10,0.25; $0)	5	0.94	0.90	0.96
8	($30,0.25; $0)	12	0.94	0.91	0.87
9	($20,0.5; $0)	6	0.02	0.03	0.11
10	($10,0.25; $0)	3	0.72	0.73	0.67
11	($16,0.5; $0)	9	.	0.41	0.45
12	($20,0.25; $10)	11	.	0.06	0.07
13	($20,0.5; $0)	11	.	0.40	0.51
14	($15,0.75; $0)	10	.	0.04	0.05
15	($34,0.5; $0)	16	.	0.40	0.47
16	($15,0.5; $6)	10	.	0.19	0.15
17	($20,0.75; $0)	14	.	0.14	0.13
18	($12,0.75; $0)	8	.	0.19	0.15
19	($27,0.5; $0)	14	.	0.54	0.53
20	($25,0.25; $5)	9	.	0.29	0.20
21	($7,0.75; $0)	4	.	0.13	0.13
22	($38,0.25; $0)	16	.	0.90	0.76
23	($14,0.5; $0)	6	.	0.21	0.07
24	($5,0.75; $0)	3	.	0.09	0.15
25	($28,0.5; $0)	13	.	0.26	0.62

Figure 7: *Principal's payment to the third party (first* 10 *periods)*

(a) *Pay treatment*



(b) *Pay×3 treatment*