Theoretical Economics 11 (2016), 1145–1179 1555-7561/20161145

Rewards and punishments: Informal contracting through social preferences

Sylvain Chassang
Department of Economics, New York University

Christian Zehnder
Department of Organizational Behavior, University of Lausanne

This paper develops a positive model of informal contracting in which rewards and punishments are not determined by an ex ante optimal plan but instead express the ex post moral sentiments of the arbitrating party. We consider a subjective performance evaluation problem in which a principal can privately assess the contribution of an agent to the welfare of a broader group. In the absence of formal contingent contracts, the principal chooses ex post transfers that maximize her social preferences. We characterize the incentives induced by the principal's preferences, contrast them with ex ante optimal contracts, and derive novel testable predictions about the way externalities are internalized in informal settings.

Keywords. Informal contracts, social preferences, fairness, intent-based justice, subjective performance evaluation, no punishment without guilt.

JEL classification. C72, D23, D86, K13, K42, L20.

1. Introduction

This paper develops a positive model of informal contracting in which rewards and punishments are not determined by an ex ante optimal contract but instead express the ex post moral sentiments of the arbitrating party. We consider a subjective performance evaluation problem in which the principal can privately assess the contribution of an agent to the welfare of a broader group.[1] The agent's actions affect both his outcome and that of the group. We assume that the principal cannot commit to transfer schemes,

Sylvain Chassang: chassang@nyu.edu
Christian Zehnder: christian.zehnder@unil.ch

An early version of this paper was previously circulated under the title "A theory of informal justice." We are indebted to Dilip Abreu, Roland Benabou, Niko Matoushek, and Wolfgang Pesendorfer for very helpful discussions. We are grateful to Armin Falk, Bob Gibbons, and Klaus Schmidt as well as seminar participants at Bern, Bonn, Chicago Booth, Columbia (Political Economy and Law), Georgetown, Lausanne, Maastricht, MIT Sloan, Montreal, NBER, Princeton, and Uppsala for many useful comments. Chassang gratefully acknowledges the hospitality of the University of Chicago Booth School of Business as well as support from the Alfred P. Sloan Foundation. Zehnder gratefully acknowledges support from the Swiss National Science Foundation (Project 26051096).

[1] Throughout we refer to the agent as "he" and to the principal as "she."

DOI: 10.3982/TE2063

but instead implements transfers that maximize her social preferences ex post. This may be because the principal does not have commitment power, is not physically present at the ex ante stage, or because specifying fully contingent contracts is simply too complicated. We show that social preferences impose plausible restrictions on patterns of rewards and punishments, that are not captured by existing models of informal incentives such as relational contracts.[2] These restrictions have novel implications for the internalization of different types of externalities.

The game form we use is straightforward. A principal can transfer resources between an active agent, player A, and a broader group modeled as a single passive player P. Player A takes a private action $a \in \{C, D\}$ that induces stochastic payoffs for himself and player P. Action C is a pro-social action that increases the expected payoff of player P at the expense of player A. The principal observes realized payoffs, as well as an imperfect signal of player A's behavior. The principal's evaluation is subjective in the sense of Baker et al. (1994): circumstantial evidence of player A's behavior is available, but is not usable by an external court.[3] Transfers have an efficiency cost: the cost to the transferring party exceeds the value transferred to the receiving party. These costs reflect specificities in the transferred resources, promotions, and decision rights. The principal has no commitment power, and chooses ex post transfers that maximize her social preferences. These transfers give rise to an informal incentive scheme that in turn determines player A's behavior.

Since the principal determines transfers ex post, social preferences are crucial to our model. Ex post, a principal exclusively concerned with efficiency would never impose costly transfers. In deterministic decision problems, the principal's preferences coincide with the inequity-averse preferences suggested by Fehr and Schmidt (1999).[4] To deal with the stochastic nature of our environment we extend their model to accommodate two forms of uncertainty: (i) exogenous uncertainty over outcomes given player A's action, and (ii) endogenous uncertainty over player A's behavior. Motivated by experimental evidence, our specification places weight on both ex post (or allocative) fairness, and ex ante (or procedural) fairness. Preferences over ex ante fairness make this a nonexpected-utility model and the principal need not be consequentialist (Machina 1989): the perceived fairness of an unequal realized allocation depends on the fairness of the underlying lottery that generated that outcome.

We show that two qualitatively distinct *modes of informal justice* can arise, depending on the weight that the principal places on ex ante versus ex post fairness. A large weight on ex post fairness implies that rewards and punishments follow what we refer to as *outcome-based justice*: transfers depend only on outcomes, ignoring all side information; there is no punitive justice, in the sense that transfers at most compensate

[2] See, for instance, Green and Porter (1984), Bull (1987), MacLeod and Malcomson (1989), Baker et al. (1994, 2002), Levin (2003).

[3] Whether the principal's signal is public or private plays no role in our setting. In richer contracting environments MacLeod (2003) emphasizes the value of cross-checking mechanisms to elicit correlated information from the agent and the principal when signals are private.

[4] We pick Fehr–Schmidt's model of social preferences because of its parsimony and tractability. Alternative models include Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Levine (1998), Charness and Rabin (2002), Benabou and Tirole (2006).

for realized inequality, and informal incentives induce a generically unique pure strategy equilibrium. A large weight on ex ante fairness, in contrast, implies that rewards and punishments follow *intent-based justice*: transfers depend both on payoff outcomes and on any available side information; punitive transfers going above and beyond realized inequality are possible; finally, there may be multiple equilibria, and equilibrium may require mixing by player A.

Under intent-based justice, the principal only imposes punishments on the agent if her posterior belief that the agent chose non-pro-social action D is sufficiently high: because the principal cares about ex ante fairness, she is unwilling to punish an agent she believes has behaved in a pro-social manner. Rewards and punishments reflect actual changes in the principal's belief over the agent's action. This contrasts significantly with existing models of formal and informal contracting (including the seminal work of Holmström (1979), Harris and Raviv (1979), Green and Porter (1984), or Bull (1987)) in which the agent's behavior is known in equilibrium, so that rewards and punishments are conditioned on noise and do not reflect changes in posterior beliefs. We refer to this property as the *no-punishment-without-guilt* restriction and show that it has significant consequences on the way different externalities are internalized. Negative externalities induce mixed strategy equilibria in which externalities are partially internalized. Positive externalities induce multiple pure strategy equilibria, under which externalities are either fully internalized or not internalized at all.[5]

Finally, we outline efficiency properties of informal justice. Outcome-based justice is efficient conditional on transfers and guarantees a minimum share of the efficient surplus. However, it remains bounded away from first-best efficiency even as side information becomes precise. Intent-based justice admits a most pro-social equilibrium that approaches efficiency as information becomes precise. However, intent-based justice may admit other equilibria achieving an arbitrarily low share of the efficient surplus. In this sense, intent-based justice is potentially more efficient, but also less robust than outcome-based justice.

Our work is related to relational contracting models that also place restrictions on rewards and punishments in the absence of ex ante formal contracts (see, for instance, Green and Porter 1984, Bull 1987, MacLeod and Malcomson 1989, Baker et al. 1994, 2002, Levin 2003).[6] Relational contracts restrict incentives by requiring them to be subgame perfect in an appropriate repeated game. Our approach imposes that rewards and punishments maximize the ex post social preferences of the relevant decision-maker.[7]

[5] Positive and negative externalities are defined in reference to exogenously given status quo expected payoffs. In a negative externality environment, action C leaves passive player P at her status quo payoff, while action D brings player P below her status quo payoff. In a positive externality environment, action D leaves player P at her status quo payoff, while action C brings player P above her status quo payoff.

[6] For more recent work on relational contracts, see Chassang (2010), Board (2011), Halac (2012), Fong and Li (2010), Li and Matouschek (2013).

[7] Compte and Postlewaite (2015) also explore the idea that emotions place restrictions on informal incentives by studying a repeated game in which play is conditioned on emotional states rather than the full history of past observables. In their framework, emotions are purely informational states that do not affect preferences.

Our analysis is related to Fudenberg and Tirole (1990), who study renegotiation in a principal–agent problem. In their model the principal has no commitment power, and contracts are renegotiated at an interim stage occurring after the agent's decision, but before outcomes are realized. At this interim stage, it is Pareto improving for the principal to offer the agent insurance. Hence, there cannot exist an equilibrium in which the agent puts effort with probability 1. Renegotiation would lead to perfect insurance, thereby removing all incentives for effort. The interim contracts offered by the principal at the renegotiation stage can be interpreted as maximizing social preferences that weighs both the principal's profits and the agent's expected utility.

Because the principal's social preferences play a central role in our framework, this paper belongs to the growing behavioral contracting literature. One strand of this literature takes contracts as given and contrasts their implications when agents are selfish and when agents have social preferences.[8] Another line of research investigates optimal contracting in the presence of agents with social preferences.[9] We argue that social preferences define informal contracting heuristics, and study the patterns of punishments and rewards they induce, as well as their effectiveness in sustaining efficient play.

The paper is structured as follows. Section 2 describes our general framework. Section 3 outlines benchmark properties of ex ante optimal contracts. Sections 4, 5, and 6 characterize patterns of rewards and punishments, as well as equilibrium behavior under intent-based justice. Section 7 deals with the case of outcome-based justice. Section 8 discusses our modeling choices as well as directions for further research. Unless mentioned otherwise, proofs are collected in Appendix A and Appendix S.B, available in a supplementary file on the journal website, http://econtheory.org/supp/2063/supplement.pdf. Appendix S.A (in the supplementary file) discusses the robustness of our findings to various modeling assumptions and develops a simple model of endogenous contract completeness.

2. Framework

We first describe our model, and then provide relevant examples. Two key features are that (i) the principal cannot (or does not) commit to an ex ante contract and (ii) the principal chooses ex post transfers that maximize her social preferences.

2.1 *The formal model*

Players, payoffs, and information A principal privately assesses the contribution of an active player A to the welfare of a broader group modeled as a passive player P. Player

[8] For instance, several studies show that generous fixed wages induce fair-minded workers to increase non-enforceable effort provision (Fehr et al. 1997, Falk and Gaechter 2002, Charness 2004). Other work suggests that explicit incentives can reduce effort by crowding out pro-social motivation (Gneezy and Rustichini 2000, Benabou and Tirole 2003, Ellingsen and Johannesson 2008, Falk and Kosfeld 2006).

[9] Experimental work shows that when agents have social preferences, non-enforceable bonus payments are a powerful motivator (Fehr et al. 2007), linear payment rules may be optimal (Englmaier and Wambach 2010), and rigid ex ante contracts limit shading (Hart and Moore 2008, Fehr et al. 2011).

P makes no strategic decisions, but her welfare enters the social preferences of the principal.

Formally, player A takes a private action $a \in \{C, D\}$ that affects her payoff and that of passive player P. The principal observes consequences $z = (u, x) \in Z = U \times X$, which are a sufficient statistic for the players' payoffs $u \equiv (u_A, u_P)$, but also include a payoff-irrelevant signal x, informative of player A's behavior. Except when explicitly mentioned otherwise, we assume that $Z \subset \mathbb{R}^k$ is compact, with a nonempty interior.

We make a few technical assumptions. Let $\mathcal{L}$ denote the restriction of the Lebesgue measure to Z and let $f(z|a)$ denote the common-knowledge density of observable outcomes z given action a against $\mathcal{L}$. We assume that for all $a \in \{C, D\}$, density $f(z|a)$ is bounded below by some value $\underline{h} > 0$. We refer to (Z, f) as the *environment* and to its restriction to payoffs $(U, f_{|U})$ as the *payoff environment*.

Assumption 1. *The log-likelihood ratio* $\log(f(z|D)/f(z|C))$ *has no mass points under* $\mathcal{L}$*:*

$$\forall l \in \mathbb{R}, \quad \mathcal{L}\left(\left\{z \text{ s.t. } \log\left(\frac{f(z|D)}{f(z|C)}\right) = l\right\}\right) = 0.$$

This assumption helps ensure that optimal transfer policies are unique.

Given $a \in \{C, D\}$, let $\neg a$ denote the alternative action. We assume that payoffs (u_A, u_P) are centered in the following way.

Assumption 2. *We have*

$$\begin{aligned} &\textit{(centering)} \quad \forall i \in \{A, P\}, \exists a \in \{C, D\} \quad \textit{s.t.} \quad \mathbb{E}[u_i|a] \geq 0 \geq \mathbb{E}[u_i|\neg a] \\ &\textit{(conflict)} \quad \mathbb{E}[u_A|D] > \mathbb{E}[u_A|C] \quad \textit{and} \quad \mathbb{E}[u_P|D] < \mathbb{E}[u_P|C]. \end{aligned} \tag{1}$$

Centering lets us interpret payoffs u_A and u_P as departures from either player's outside option, i.e., from the counterfactual payoff each would obtain if the other player was absent. *Conflict* restricts attention to cases where additional incentives are needed to internalize externalities, and labels C as a pro-social action creating value for player P at the expense of player A.

The purpose of Assumption 2 is only to reduce the number of cases covered in the analysis. Our model of informal justice remains well defined when Assumption 2 does not hold.

Transfers After observing state z, the principal can implement a transfer $T_z \in [-T_{\max}, T_{\max}]$ between the two players. This results in an incentive scheme $T : z \mapsto T_z$ contingent on state z. Transfer T_z has a dead-weight loss $\lambda|T_z|$, with $\lambda > 0$, accruing to the transferring party.[10] We assume for simplicity that $T_{\max} \geq \max_z |u_A - u_P|$. By convention, a positive transfer corresponds to a transfer from player A to player P. Let us denote by

$$u_A^T \equiv u_A - T - \lambda T^+ \quad \text{and} \quad u_P^T \equiv u_P + T - \lambda T^-$$

[10] We think of the cost of transfers as arising from specificities in the resources being transferred from one party to the other. Appendix S.A shows that our analysis is robust to perturbations to social preferences and to the cost of transfers.

player A's and P's payoffs net of transfers, and denote by $u^T \equiv (u^T_A, u^T_P)$ the net payoff profile.[11]

The principal's problem We denote by $\pi \in \Delta(\{C, D\})$ mixed strategies of player A, and interpret mixed strategies as pure strategies played by a population of players. Given a belief $\pi \in \Delta(\{C, D\})$, the principal chooses a transfer scheme T to maximize social preferences

$$V(\pi, T) = \sum_{a \in \{C,D\}} \pi(a)(\delta \mathbb{E}[\Phi(u^T)|a] + (1-\delta)\Phi(\mathbb{E}[u^T|a])), \tag{2}$$

where $\Phi(u) \equiv u_A + u_P - \alpha|u_A - u_P|$, $\alpha \geq 0$, and $\delta \in [0, 1]$.

These social preferences capture inequity aversion, preferences for ex ante as well as ex post fairness, and betrayal aversion. Inequity aversion is captured by utility function $\Phi(u) = u_A + u_P - \alpha|u_A - u_P|$, a simplified version of Fehr and Schmidt (1999)'s model. When $\alpha > 0$, the principal not only values utilitarian efficiency, but also dislikes inequity. Our interpretation of u_A and u_P as departures from reference status quo payoffs plays a role because of the inequity term $|u_A - u_P|$. Since the principal evaluates equity by comparing deviations from outside options, shifts in reference payoffs change the principal's preferences over transfers. We assume throughout the paper that $\alpha > \lambda/(2 + \lambda)$. In the opposite case, the efficiency loss of transfers outweighs their impact on equity and the principal sets transfers identically equal to 0 (see Lemma S.A.1 in the supplementary Appendix).

Preferences over ex post and ex ante fairness are captured by parameter δ. Term $\mathbb{E}[\Phi(u^T)|a]$ evaluates the fairness of realized allocations and captures ex post (or allocative) fairness. Term $\Phi(\mathbb{E}[u^T|a])$ evaluates the fairness of expected payoffs conditional on action a and captures ex ante (or procedural) fairness. This class of preferences is motivated by experimental evidence from Bolton et al. (2005), Charness and Levine (2007), and others.[12] As Machina (1989) highlights—using the example of a mother deciding how to allocate a piece of candy between her two children and preferring a lottery over any certain outcome—preferences for ex ante fairness cannot be captured by an expected utility model. As a result, the principal in our model is not consequentialist.[13]

Betrayal aversion (see Bohnet and Zeckhauser 2004) follows from the fact that the principal's preferences $V(\pi, T)$ over transfers T given π differentiate strategic uncertainty π over actions and nonstrategic uncertainty $f(z|a)$ over outcomes, given actions. Using the notation of (2), terms corresponding to ex ante fairness sum to $\sum_a \pi(a)\Phi(\mathbb{E}[u|a])$, rather than $\Phi(\mathbb{E}[u|\pi])$.[14] Interpreting mixed strategies as pure strategies played by a population, the principal evaluates the fairness of each individual relationship between given players A and P, rather than fairness at the population level:

[11] By convention, $T^+ = \max\{T, 0\}$ and $T^- = \max\{-T, 0\}$.

[12] See in particular Cushman et al. (2009), Schächtele et al. (2011), or Krawczyk and Le Lec (2010). When π is a point mass at C or D these preferences coincide with models of choice over social lotteries proposed by Krawczyk (2011) and axiomatized by Saito (2013).

[13] Given a realization of payoffs, her preferences over transfers ex post depend on the counterfactual distribution of potential payoffs. See Fudenberg and Levine (2012) for a recent discussion.

[14] Appendix S.A shows that folding together strategic and objective uncertainty yields very different predictions. In particular, the principal's transfers no longer depend on signal x.

a situation in which player A takes an action that benefits him with 50% chance and benefits the other player with 50% chance feels very different from a situation in which 50% of players A in the population take the action that benefits them, while 50% of players A take the action that benefits the other player.[15] This modeling choice captures the evidence of Bohnet and Zeckhauser (2004) and Bohnet et al. (2008), which shows that people treat fixed lotteries over social outcomes differently when the uncertainty is determined by the strategic behavior of others and when it is determined by objective uncertainty.

Solution concept Because the principal cares about ex ante fairness, the ex post transfer scheme T she implements depends on player A's expected behavior π. Since behavior π itself depends on the transfer scheme T, we have a game in which transfers T determine behavior π, and behavior π determines transfers T. As was previously noted, the principal is not consequentialist and we use Bayes Nash equilibrium as our solution concept.

Definition 1. For any distribution $\pi \in \Delta(\{C, D\})$, a pair (π, T) is a Bayes Nash equilibrium if and only if $T \in \arg\max_T V(\pi, T)$ and, for all $a \in \{C, D\}$ such that $\pi(a) > 0$, $\mathbb{E}(u_A^T|a) \geq \mathbb{E}(u_A^T|\neg a)$.

For pure strategies $\pi \in \{C, D\}$ we impose on Bayes Nash equilibria (π, T^π) the requirement that transfer scheme T^π be the limit of optimal transfer schemes $T^{\widehat{\pi}}$ for full-support distributions $\widehat{\pi} \in \Delta(\{C, D\})$ approaching π.[16]

Although the principal cares about ex ante fairness, transfers are being chosen ex post: the principal's transfers in state z depend on what transfers she would have implemented in state $z' \neq z$.[17] Our model is thus related to the framework of psychological game theory (Geanakoplos et al. 1989), in which a player's utility depends on outcomes, beliefs about others' actions, and others' beliefs. In our approach, the principal's transfer choice not only depends on her belief about the agent's action, but also on her beliefs about her own counterfactual behavior at different states.

2.2 *Motivation*

Our model builds on two core assumptions. First, the principal may not be able to (or may choose not to) commit to an ex ante contract. Second, the principal will decide on ex post transfers on the basis of her social preferences. We describe four settings in which these assumptions apply.

[15] If purification is obtained by considering a continuous but concentrated distribution of preferences, our model of preferences may be interpreted as valuing type-by-type fairness.

[16] Lemma 3 (see Section 4) establishes the existence of such limit transfers for all π. Note that the refinement holds trivially when π is fully mixed. Imposing this trembling-hand condition allows us to select a unique transfer scheme for all π (see Lemma 3). Without this refinement, optimal transfer problem (2) can admit a continuum of solutions that need not be monotonic in likelihood ratio $f(z|C)/f(z|D)$.

[17] In principle this could lead to multiple personal equilibria à la Kőszegi and Rabin (2006), among which we select the principal's preferred personal equilibrium.

High fixed costs make ex ante formal contracting unlikely to be used at the lower echelons of organizations. Consider the manager of a sales team. A salesperson may exert negative or positive externalities on her team by poaching customers or by providing expertise in dealing with questions and support issues. While this incentive problem is too small to justify formal contracting, the manager can still provide incentives through the allocation of tasks, resources, and promotions. It is plausible that she will do so according to some sense of fairness.[18]

Even at the higher echelons of organizations, where the scale of incentive problems might justify legal costs, formal contracting may be limited by bounded rationality. Consider a senior executive arbitrating between two divisions of a company. Providing incentives to each division head may well justify using formal contracts. For instance, to avoid market competition across divisions, a firm may consider formally excluding certain sales from the division heads' performance evaluation.[19] However, the senior executive may be unaware of other incentive problems. For instance, one division may have the opportunity to support the development of the other in a new geographic market. If those circumstances are not anticipated when a formal contract is written, the senior executive is mechanically left to make discretionary promotion decisions on the basis of her social preferences.

A third example is parental discipline. Parents do not typically commit to formal contracts, and frequently make disciplining decisions based on their intuitive moral judgement. Intuitively, considerations of ex post and ex ante fairness play a significant role. If a child breaks the toy of a sibling, the parents' response will not be fully driven by outcomes (ex post fairness), but will also depend on their perception of the child's intents (ex ante fairness).

Finally, in some situations the principal is not even present ex ante. One example is lay juries arbitrating a civil lawsuit. Instructions given to juries can offer significant freedom in the assignment of guilt and damages.[20] The decisions of the jury must then express the ex post moral judgement of its members. In fact, in civil cases the jury instructions clarify that the standard of proof for circumstantial evidence—*preponderance of the evidence*—requires jury members to place posterior likelihood roughly greater than a half on the reprehensible act having happened.[21] Regardless of circumstantial evidence, juries are not instructed to condemn defendants whom they believe to be innocent. This coincides with the no-punishment-without-guilt property that we emphasize throughout the paper and contrasts with existing contracting models.

[18] The psychology literature on dispute settlement in organizations (see, for instance, Folger and Konovsky 1989, Greenberg 1990, Lind et al. 1993, or Konovsky 2000) emphasizes the importance of ex ante (or procedural) fairness in the workings of organizations.

[19] Volkswagen Group recently experienced such internal competition between its Skoda, Volkswagen, and Audi divisions (Hawranek 2010), leading to the firing of Skoda's ambitious chairman.

[20] This is illustrated by the common concern that juries have excessive leeway in specifying damages. See for instance the recent (challenged) award of $23.6B in damages by a Florida jury (Robles 2014).

[21] For instance, see the jury instructions (available on the American Bar Association's website) for gender discrimination case Velez v. Novartis (2010). Instructions distinguish damages awarded because of disparate impact of policies on men and women regardless of intent (ex post fairness), versus damages awarded because of disparate treatment, which captures intent, and authorizes punitive damages (ex ante fairness).

3. The ex ante optimal benchmark

As a preliminary to our characterization of behavior under informal justice, we highlight properties of optimal contract T^* when the principal can commit ex ante. Contract T^* solves

$$\begin{aligned} &\max_{\pi\in\Delta(\{C,D\}),T} V(\pi,T) \\ &\text{with } \pi, T \quad \text{s.t.} \quad \forall a\in\{C,D\}, \quad \pi(a)>0 \quad \Rightarrow \quad \mathbb{E}[u_A^T|a]\geq\mathbb{E}[u_A^T|\neg a]. \end{aligned} \tag{P1}$$

To state continuity and genericity results, the space of integrable functions from Z to $\mathbb{R}$ (or $\mathbb{R}^2$) is endowed with the L_1 norm $\|\cdot\|_1$: for any integrable function $g:Z\to\mathbb{R}$, $\|g\|_1=\int_{z\in Z}|g(z)|\,\mathrm{d}z$. Since observable state z is a sufficient statistic for payoffs u, we can apply the L_1 norm to payoffs by viewing u as a mapping from Z to $\mathbb{R}^2$.

Contract T^* satisfies the following intuitive properties.[22]

(i) *Pure and unique behavior.* For generic payoff structures (i.e., for an open and dense set of payoff structures u in the L_1 topology), optimal ex ante contracts implement a unique pure action $a\in\{C,D\}$.

(ii) *Use of information.* Optimal ex ante contracts condition transfers on, and only on, realized payoff differences u_A-u_P and likelihood ratio $f(z|D)/f(z|C)$.

(iii) *Punishment without guilt.* Optimal ex ante contracts can exhibit punishment without guilt: on the equilibrium path, player A may be penalized by positive transfers $T_z>0$ even though $\pi(C)=1$, i.e., there is common knowledge that he took pro-social action C (see the example below).

(iv) *Punitive justice.* Optimal ex ante contracts can exhibit punitive justice, i.e., transfers $T_z>0$ that more than compensate for realized inequality, so that in some state z, $u_A>u_P$ but $u_A^T<u_P^T$.

An example

The following numerical example helps us illustrate the mechanics of our model. Public stochastic state z takes values in $Z=\{-1,1\}$, with payoffs taking the form

$$u_A=-[z+\gamma] \quad \text{and} \quad u_P=5[z+\gamma],$$

where $\gamma\in\{-\frac{1}{2},\frac{1}{2}\}$ is a fixed parameter used to perform comparative statics.

Given player A's action $a\in\{C,D\}$, the distribution of z is given by

z	-1	1
prob$(z\|C)$	$\frac{1}{4}$	$\frac{3}{4}$
prob$(z\|D)$	$\frac{3}{4}$	$\frac{1}{4}$

[22] The proofs are straightforward and are omitted for concision.

We further assume that $\delta = 0$ (i.e., the principal cares only about ex ante fairness), $\alpha \geq 1$ (inequity matters), and $\lambda \in (0, 2)$ (transfer costs are moderate).

Note that parameter γ merely shifts the payoffs of each player by a constant. When $z + \gamma > 0$ ($z + \gamma < 0$), player A has a positive (negative) externality on player P. Note that C is a pro-social action increasing the likelihood of states favorable to player P.[23]

Lemma 1. *In the environment of the example, any optimal contract* (π^*, T^*) *solving (P1) is such that: the following statements hold:*

(i) For every $\lambda \in (0, 2)$, $\alpha \in [1, +\infty)$, *and* $\gamma \in \{-\frac{1}{2}, \frac{1}{2}\}$, *the optimal contract implements action* $a^* = C$.

(ii) If $\gamma = -\frac{1}{2}$, *the optimal transfer scheme satisfies* $T^*(z = -1) > 0$.

The proof is given in Appendix S.B. Points (i) and (ii) imply that player A may be punished on the equilibrium path, even if it is common knowledge that player A took action C (punishment-without-guilt).

We now assume that the principal cannot commit ex ante, so that behavior π and transfers T are jointly determined in equilibrium. Varying the weight on ex ante versus ex post fairness yields two distinct classes of equilibria: *outcome-based justice* and *intent-based justice*.

4. Intent-based justice

We begin with the case of intent-based justice, i.e., we assume that the weight on ex ante fairness is sufficiently high: $\delta < \lambda/(\alpha(2 + \lambda))$. We characterize in this section the principal's optimal transfer scheme T^π given belief π, and turn to equilibrium predictions in Sections 5 and 6.

Take a distribution of behavior $\pi \in \Delta(\{C, D\})$ as given. Let $f_\pi(z) \equiv \sum_{a \in \{C,D\}} \pi(a) \times f(z|a)$ denote the induced distribution over observables $z \in Z$, and for all $a \in \{C, D\}$, define posterior beliefs

$$\pi(a|z) \equiv \frac{\pi(a) f(z|a)}{\sum_{\widehat{a} \in \{C,D\}} \pi(\widehat{a}) f(z|\widehat{a})}.$$

For concision, we use the notation $\Sigma u_z \equiv u_A + u_P$ and $\Delta u_z \equiv u_A - u_P$. Given transfers T, we have $\Sigma u_z^T = \Sigma u_z - \lambda |T_z|$ and $\Delta u_z^T = \Delta u_z - (2 + \lambda) T_z$. Noting that $\pi(a) f(z|a) = \pi(a|z) f_\pi(z)$, the principal's value function over transfer schemes can be expressed as

$$\begin{aligned} V(\pi, T) = \int_{z \in Z} &(\Sigma u_z - \lambda |T_z|) f_\pi(z) \, \mathrm{d}z \\ &- \delta\alpha \int_{z \in Z} |\Delta u_z - (2 + \lambda) T_z| f_\pi(z) \, \mathrm{d}z \\ &- (1 - \delta)\alpha \sum_{a \in \{C,D\}} \left| \int_{z \in Z} [\Delta u_z - (2 + \lambda) T_z] \pi(a|z) f_\pi(z) \, \mathrm{d}z \right|. \end{aligned} \tag{3}$$

[23] Note that this example does not satisfy the restriction on likelihood ratios imposed by Assumption 1. As a result, optimal transfer schemes need not be unique, but this does not affect the analysis in this example.

The three terms in the principal's value function respectively trade off minimizing the efficiency cost of transfers, minimizing ex post outcome inequality (allocative fairness), and minimizing ex ante payoff inequality (procedural fairness). Note that the space of transfer functions $T \in [-T_{\max}, T_{\max}]^Z$ is convex and that $V(\pi, T)$ is concave in T.

The next lemma shows that optimal transfers T^π do not reverse existing payoff asymmetries: net of transfers, action C continues to generate inequality in favor of player P, while action D generates inequality in favor of player A.

Lemma 2. *For any $\pi \in \Delta(\{C, D\})$, there exists an optimal transfer policy T^π. In addition, any optimal transfer policy T^π satisfies*

$$\mathbb{E}[\Delta u^{T^\pi}|D] \geq 0 \geq \mathbb{E}[\Delta u^{T^\pi}|C].$$

This lets us sign the term corresponding to ex ante fairness in expression (3), and simplifies the optimization problem. For any pair of multipliers $\mu = (\mu_C, \mu_D) \geq 0$, define the Lagrangian

$$\begin{aligned} L(\mu, z, T_z) \equiv -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| + (1-\delta)\alpha(2+\lambda)[\pi(D|z) - \pi(C|z)]T_z \\ - \mu_D \pi(D|z)T_z + \mu_C \pi(C|z)T_z. \end{aligned} \tag{4}$$

Optimal transfer schemes can be characterized as follows.

Lemma 3 (Characterization). *For every distribution π in the interior of $\Delta(\{C, D\})$ and every $\delta \in [0, 1]$, there exists a unique optimal transfer scheme T^π. It takes the form*

$$T_z^\pi = \arg \max_{T_z \in [-T_{\max}, T_{\max}]} L(\mu, z, T_z)$$

for a vector $\mu = (\mu_C, \mu_D) \geq 0$ such that $\max\{\mu_C, \mu_D\} \leq (1-\delta)\alpha(2+\lambda)$.

There exist unique transfer schemes T^C and T^D such that, under the L_1 norm, $\lim_{\pi \to C} T^\pi = T^C$ and $\lim_{\pi \to D} T^\pi = T^D$.

Inspection of (4) yields that T_z^π depends only on realized inequality Δu_z and on the posterior likelihood ratio $\pi(D|z)/\pi(C|z) = \pi(D)f(z|D)/(\pi(C)f(z|C))$, so that transfers can depend on informative payoff-irrelevant signals x—a feature shared with optimal ex ante contracts. However, unlike ex ante optimal contracts, transfers under informal justice also depend on prior beliefs π.

The uniqueness of optimal transfers relies on Assumption 1. It guarantees that the set of values z such that Lagrangian $L(\mu, z, T_z)$ admits multiple maximizers T_z has measure zero.

Proposition 1 (Rewards and punishments). *(i) There is no punishment without guilt and no reward without virtue:*

$$T_z^\pi > 0 \quad \Rightarrow \quad \pi(D|z) \geq \frac{1}{1-\delta}\left(\frac{\lambda}{\alpha(2+\lambda)} - \delta\right) > 0 \quad (\textit{no punishment without guilt})$$

$$T_z^\pi < 0 \quad \Rightarrow \quad \pi(C|z) \geq \frac{1}{1-\delta}\left(\frac{\lambda}{\alpha(2+\lambda)} - \delta\right) > 0 \quad (\textit{no reward without virtue}).$$

(ii) *For any interior $\pi \in \Delta(\{C, D\})$, there exist thresholds $-1 \leq h_-^{\max} < h_-^{\Delta} < h_+^{\Delta} < h_+^{\max} \leq 1$ such that transfer policy T^π takes the form*

$$T_z^\pi = \begin{cases} 0 & \textit{if } \pi(D|z) - \pi(C|z) \in (h_-^\Delta, h_+^\Delta) \\ -T_{\max} & \textit{if } \pi(D|z) - \pi(C|z) < h_-^{\max} \\ T_{\max} & \textit{if } \pi(D|z) - \pi(C|z) > h_+^{\max} \\ (\Delta u_z)^+/(2+\lambda) & \textit{if } \pi(D|z) - \pi(C|z) \in (h_+^\Delta, h_+^{\max}) \\ -(\Delta u_z)^-/(2+\lambda) & \textit{if } \pi(D|z) - \pi(C|z) \in (h_-^{\max}, h_-^\Delta).^{24} \end{cases}$$

Point (i) establishes that the principal punishes (resp. rewards) only if she holds a sufficient posterior belief that the agent took action D (resp. action C)—there is no punishment without guilt (no reward without virtue). We call this mode of informal justice *intent based*.

Partial proof. We prove the no-punishment-without-guilt property when the principal cares only about ex ante fairness (i.e., $\delta = 0$, as in our numerical example), and refer the reader to Appendix A for a complete treatment.

It follows from Lemma 3 that there exists $\mu = (\mu_C, \mu_D) \geq 0$ such that for any $\pi \in \Delta(\{C, D\})$ and $z \in Z$, optimal transfers T_z^π satisfy $T_z^\pi \in \arg\max_{T_z \in [-T_{\max}, T_{\max}]} L(\mu, z, T_z)$ for

$$L(\mu, z, T_z) = -\lambda|T_z| + [\alpha(2+\lambda) - \mu_D]\pi(D|z)T_z - \underbrace{[\alpha(2+\lambda) - \mu_C]}_{\geq 0}\pi(C|z)T_z.$$

Hence, if optimal transfers T_z^π are strictly positive, it must be that $[\alpha(2+\lambda) - \mu_D] \times \pi(D|z) \geq \lambda$. Since $\mu_D \geq 0$, this implies that $\pi(D|z) \geq \lambda/(\alpha(2+\lambda))$. The principal must place a sufficiently high posterior weight on guilt to implement punitive transfers. □

Point (ii) shows that transfers take a threshold form dependent on the precision $\pi(D|z) - \pi(C|z)$ with which player A's behavior is inferred ex post. Transfers follow three distinct regimes:

1. For sufficiently extreme posterior beliefs $\pi(D|z) - \pi(C|z)$, the magnitude of transfers is $T_{\max}$: transfers more than compensate for realized inequality; there is *punitive justice*.[25]

2. For sufficiently strong, but less extreme posterior beliefs, transfers are implemented through selective fairness. If the principal tends to believe that player A took action D, she implements equalizing transfers when realized payoffs are in favor of A, but not if P is favored. This is consistent with findings of Henrich et al.

[24] At the limit where $\pi(C) = 1$ or $\pi(D) = 1$, limit transfers respectively, take the form $T_z^C = -(1/(2+\lambda))\Delta u_z^- \mathbf{1}_{f(z|C)/f(z|D) \geq \theta}$ with θ such that $\mathbb{E}[\Delta u^{T^C}|C] = 0$ and $T_z^D = (1/(2+\lambda))\Delta u_z^+ \mathbf{1}_{f(z|D)/f(z|C) \geq \theta}$ with θ such that $\mathbb{E}[\Delta u^{T^D}|D] = 0$.

[25] This region may be empty if $h_+^{T_{\max}} = 1$ and $h_-^{T_{\max}} = -1$. See Lemma S.B.1 in Appendix S.B for sufficient conditions ensuring that this region is not empty in equilibrium.

(2006) showing that perceived misbehavior is often punished by withdrawing informal social protection.

3. For middling beliefs, transfers do not improve the principal's sense of fairness enough to compensate efficiency costs. She avoids transfers altogether.

5. Intent-based justice in equilibrium

Under intent-based justice, incentives provided by the principal and the behavior of player A are jointly determined. We first provide a general characterization of equilibrium behavior and then study how the set of equilibria changes across negative and positive externality environments.

5.1 *Existence and structure of equilibria*

Since strategy profiles (π, T) live in a continuous high dimensional space, the existence of equilibrium requires a proof. We know that given π, there exists a unique optimal transfer scheme T^π (Lemma 3). Let us denote by $\Gamma(\pi) \equiv \mathbb{E}[u_A^{T^\pi}|C] - \mathbb{E}[u_A^{T^\pi}|D]$ player A's incentives to take pro-social action C under transfer scheme T^π. Recall that $f(\cdot|a)$ denotes the distribution of states $z \in Z$ conditional on action $a \in \{C, D\}$. The following continuity property holds (see Appendix S.B for a proof).

Lemma 4. *Transfer T^π and mapping $\Gamma(\pi)$ are continuous in π and f under the L_1 norm.*

The existence of equilibrium follows from the continuity of incentives with respect to π. In particular, equilibria under intent-based justice are characterized by the zeros of Γ. One useful implication is that there exists a most pro-social equilibrium $(\overline{\pi}, T^{\overline{\pi}})$ characterized by

$$\overline{\pi} = \arg \max_{\pi \in \Delta(\{C,D\})} \{\pi(C)|\Gamma(\pi) \geq 0\}.$$

It will be instructive to consider the perfect monitoring limit for environments (Z, f).

Definition 2 (Perfect monitoring). Consider a sequence $(Z_n, f_n)_{n\in\mathbb{N}}$ of environments, all consistent with the same payoff environment $(U, f_{|U})$. We say that this sequence of environments approaches perfect monitoring if and only if

$$\forall \kappa > 0, \quad \lim_{n\to\infty} \text{prob}_{f_n}\left(\frac{f_n(z|D)}{f_n(z|C)} > \kappa \middle| D\right) = 1 \quad \text{and} \quad \lim_{n\to\infty} \text{prob}_{f_n}\left(\frac{f_n(z|C)}{f_n(z|D)} > \kappa \middle| C\right) = 1.$$

As we approach the perfect monitoring limit, with arbitrarily high probability the principal obtains an arbitrarily strong signal of which action was taken.

5.2 *Negative externalities*

Take as given a payoff environment $(U, f_{|U})$. We refer to it as a negative externality environment whenever

$$\mathbb{E}[u_A|C] = \mathbb{E}[u_P|C] = 0, \qquad \mathbb{E}[u_A|D] > 0 > \mathbb{E}[u_P|D] \tag{5}$$

$$\text{and} \quad \mathbb{E}[u_A + u_P|C] - \mathbb{E}[u_A + u_P|D] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|D] \tag{6}$$

hold. Condition (5) states that pro-social action C delivers status quo payoffs in expectation, while D generates a negative externality for player P. Condition (6) is automatically satisfied when action C generates a higher expected sum of payoffs than action D since condition (1) implies that $\mathbb{E}[\Delta u|D] \geq 0$. The following result holds.

Proposition 2 (Homogeneous response to negative externalities). *Take as given a payoff environment $(U, f_{|U})$ satisfying (5) and (6).*[26]

(i) For any environment (Z, f) consistent with $(U, f_{|U})$, all equilibria (π, T) are such that $\pi(C) < 1$.

(ii) Consider environments $(Z_n, f_n)_{n\geq 0}$ consistent with $(U, f_{|U})$, approaching perfect monitoring. For n sufficiently large, all equilibria (π_n, T_n) satisfy $\pi_n(C) > 0$.

This result suggests that organizations should be relatively homogeneous in their informal response to negative externalities: negative externalities are never fully internalized ($\pi(C) < 1$), but they are always partially internalized provided that monitoring is sufficiently accurate ($\pi(C) > 0$). The result that negative externalities are never fully internalized is intuitive. If C occurred with probability 1, no-punishment-without-guilt would imply that the principal does not impose transfers. The result that negative externalities are partially internalized is less obvious, but can be illustrated using the example introduced in Section 3.

Example. Set $\gamma = -\frac{1}{2}$. Action C is a status quo action ($\mathbb{E}[u_A|C] = \mathbb{E}[u_P|C] = 0$), and D induces a negative externality on player P ($\mathbb{E}[u_A|D] = 1 > \mathbb{E}[u_P|D] = -5$).

For any $\pi(D) \in (0, 1)$, the transfers chosen by the principal solve

$$\max_{T_{-1}, T_1} \pi(C)\left[-\frac{1}{4}\lambda|T_{-1}| - \frac{3}{4}\lambda|T_1| - \alpha(2+\lambda)\left|\frac{1}{4}T_{-1} + \frac{3}{4}T_1\right|\right]$$
$$+ \pi(D)\left[-\frac{3}{4}\lambda|T_{-1}| - \frac{1}{4}\lambda|T_1| - \alpha\left|6 - (2+\lambda)\left(\frac{3}{4}T_{-1} + \frac{1}{4}T_1\right)\right|\right].$$

For $\pi(D)$ sufficiently close to 1, the second term dominates and is maximized by setting $T_{-1} \geq 0$, $T_1 \geq 0$ and (since $\alpha > \lambda/(2+\lambda)$), $6 - (2+\lambda)(\frac{3}{4}T_{-1} + \frac{1}{4}T_1) = 0$. Under this

[26] Note that while the assumption of purely negative externalities is knife-edge, the continuity of transfer schemes with respect to f (Lemma 4) implies that Proposition 2 continues to hold when payoffs are perturbed and conditions (5) and (6) hold only approximately.

constraint, the first term is minimized by setting $T_1 = 0$ and $T_{-1} = 8/(2+\lambda)$. For such transfers, player A's expected payoffs for actions C and D are

$$\mathbb{E}[u_A^T|C] = -\frac{2+2\lambda}{2+\lambda} \quad \text{and} \quad \mathbb{E}[u_A^T|D] = -\frac{4+5\lambda}{2+\lambda}.$$

Hence $\mathbb{E}[u_A^T|C] > \mathbb{E}[u_A^T|D]$, which implies that there is no equilibrium such that $\pi(D) = 1$. Externalities are always internalized (in this case, Lemma S.B.2 in Appendix S.B shows there exists a unique mixed-strategy equilibrium).

In our sales-team example this result suggests that managers who value ex ante fairness will never be able to fully prevent salespeople from poaching clients. No-punishment-without-guilt implies that the manager is willing to sanction only when she has a sufficiently strong posterior belief that a salesperson poached customers. Some residual misbehavior must subsist for the manager to be willing to enforce transfers. ◊

5.3 *Positive externalities*

We say that a given payoff environment $(U, f_{|U})$ is a positive externality environment if D is a status quo action, while C generates positive value for player P at a cost to player A. Formally, we assume that

$$\mathbb{E}[u_A|D] = \mathbb{E}[u_P|D] = 0, \qquad \mathbb{E}[u_P|C] > 0 > \mathbb{E}[u_A|C] \tag{7}$$

$$\text{and} \quad \mathbb{E}[u_A + u_P|C] - \mathbb{E}[u_A + u_P|D] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|C]. \tag{8}$$

Note that one can transition between negative and positive externality environments by shifting payoffs with a constant. This occurs when setting $\gamma = \frac{1}{2}$ rather than $-\frac{1}{2}$ in our numerical example. The following proposition holds.

Proposition 3 (Heterogeneous response to positive externalities). *Take as given a payoff environment $(U, f_{|U})$ satisfying (7) and (8).*

(i) For any environment (Z, f) consistent with payoff environment $(U, f_{|U})$, there exists an equilibrium (π, T) such that $\pi(D) = 1$ and T is identically equal to zero.

(ii) Consider environments $(Z_n, f_n)_{n\geq 0}$ consistent with $(U, f_{|U})$, approaching perfect monitoring. For n sufficiently large, there exists an equilibrium (π_n, T_n) such that $\pi_n(C) = 1$.

Positive externality environments are consistent with multiple pure strategy equilibria in which externalities are fully internalized or not internalized at all.[27] Establishing that $\pi(D) = 1$ is part of an equilibrium is straightforward. If $\pi(D) = 1$, condition (7) implies that there is no inequality in expected payoffs and the principal's optimal policy is to implement zero transfers. As a result, player A has no incentives to take action C, and $\pi(D) = 1$ is indeed an equilibrium. The existence of an equilibrium such that $\pi(C) = 1$ is less obvious, but can be illustrated using the example of Section 3.

[27] Note that these two pure strategy equilibria exist for information structures $f(z|a)$ in a neighborhood of the perfect monitoring limit and, therefore, are insensitive to local perturbations in the signalling structure.

Example. Set $\gamma = \frac{1}{2}$. Action D is a status quo action ($\mathbb{E}[u_A|D] = \mathbb{E}[u_P|D] = 0$), and C induces a positive externality on player P ($\mathbb{E}[u_A|C] = -1 < \mathbb{E}[u_P|C] = 5$).

As $\pi(C)$ approaches 1, there is inequality in expected payoffs that the principal seeks to correct. We know from Proposition 1 that such transfers concentrate rewards in states z that have the highest likelihood ratio (for action C vs. action D). For $\pi(C)$ sufficiently close to 1, $T(z = -1) = 0$ and $T(z = 1) = -8/(2 + \lambda)$ turn out to be optimal transfers. The induced expected payoffs for player A are

$$\mathbb{E}[u_A^T|C] = -1 + \frac{6}{2+\lambda} \quad \text{and} \quad \mathbb{E}[u_A^T|D] = \frac{2}{2+\lambda}.$$

Since $\lambda < 2$, it follows that $\mathbb{E}[u_A^T|C] > \mathbb{E}[u_A^T|D]$. Hence, playing C is indeed part of an equilibrium.

This result suggests that organizations may be heterogeneous in their informal response to positive externalities, yielding different "firm cultures." Some organizations may end up in a low-cooperation equilibrium, where good outcomes are assigned to luck; some organizations may end up in a high-cooperation equilibrium, where good outcomes are assigned to good behavior. ♢

6. Intent-based justice and information

We now study the effectiveness of intent-based justice as a function of the quality of information available to the principal. For this, we fix the payoff environment and evaluate the ability of intent-based justice to provide incentives for pro-social behavior as the information available to the principal varies. Specifically we fix payoff environment $(U, f_{|U})$ and assume throughout the rest of this section that

$$\mathbb{E}[u_A + u_P|C] - \mathbb{E}[u_A + u_P|D] > \frac{\lambda}{2+\lambda}\mathbb{E}[|u_A - u_P||C].$$

This ensures that regardless of the full specification of environment (Z, f) (including side signals x), the optimal ex ante contract induces action C.

The next proposition relates the quality of information available, the principal's willingness to punish, and the effectiveness of intent-based justice in inducing pro-social behavior. It is useful to define $\hbar \equiv (2 - \lambda/(\alpha(2 + \lambda)))(\lambda/(\alpha(2 + \lambda)) - \delta)^{-1} > 0$ and $\Psi \equiv \mathbb{E}[u_A|D] - \mathbb{E}[u_A|C] + 1/(2 + \lambda)\mathbb{E}[\Delta u|C]$. Note that Ψ is strictly positive in negative externality environments since in that case $\mathbb{E}[\Delta u|C] = 0$.[28]

Proposition 4. *(i) For any $\pi \in \Delta(\{C, D\})$, the probability that the principal punishes player A conditional on action D being taken satisfies*

$$\text{prob}(T^\pi > 0|D) \leq \hbar\frac{\pi(D)}{\pi(C)}\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle|D\right].$$

[28] When $\Psi > 0$, the agent's baseline payoff $\mathbb{E}[u_A|D]$ conditional on action D is higher than the payoff $\mathbb{E}[u_A|C] - 1/(2 + \lambda)\mathbb{E}[\Delta u|C]$ she would obtain under action C, even including rewards $-1/(2 + \lambda)\mathbb{E}[\Delta u|C]$ sufficient to remove expected payoff inequality between players. An implication of this (see the proof of Proposition 4 for details) is that whenever $\Psi > 0$, all equilibrium transfer schemes that induce player A to take action C with positive probability must involve punishment with positive probability.

The most cooperative equilibrium $(\overline{\pi}, T^{\overline{\pi}})$ *satisfies*

$$\overline{\pi}(C) \leq 1 - \frac{\Psi}{\hbar T_{\max}(1 + \mathbb{E}[\frac{f(z|D)}{f(z|C)}|D]) + \Psi}.$$

(ii) *Consider environments* $(Z_n, f_n)_{n \geq 0}$ *approaching perfect monitoring. As* n *goes to infinity,* $\overline{\pi}_n$ *converges to* C *and* $V(T^{\overline{\pi}_n}, \overline{\pi}_n)$ *converges to the value obtained under the optimal ex ante contract, i.e.,*

$$\lim_{n \to \infty} V(T_n^{\text{ex ante}}, C) - V(T^{\overline{\pi}_n}, \overline{\pi}_n) = 0.$$

Point (i) emphasizes that beliefs and information constrain the principal's willingness to punish. When the expected likelihood ratio $\mathbb{E}[f(z|D)/f(z|C)|D]$ is finite, so that we are bounded away from perfect monitoring, the principal is unwilling to punish player A unless she has a sufficiently high prior belief that he took action D.[29] When punishments are necessary to sustain cooperation (which is the case when $\Psi > 0$), this implies an upper bound on the highest amount of cooperation that can occur as a function of the quality of information available to the principal.

In turn, when the environment approaches perfect monitoring, intent-based justice can sustain arbitrarily high levels of pro-social behavior and approaches the efficiency of ex ante optimal contracts.[30] This finding adds to the argument that social preferences may serve to encode useful social heuristics.[31] We emphasize that point (ii) applies only to the most pro-social equilibrium $(\overline{\pi}, T^{\overline{\pi}})$. Some equilibria may yield an arbitrarily low share of first-best surplus, even as information becomes arbitrarily good. This is for instance the case in the $(\pi = D, T = 0)$ laissez-faire equilibrium that arises in positive externality environments (see Proposition 3).

7. Outcome-based justice

We now treat the case where the principal places weight $\delta > \lambda/(\alpha(2 + \lambda))$ on ex post fairness. It becomes optimal for the principal to equalize payoffs for each outcome realization.

Proposition 5. *For all behavior distributions* $\pi \in \Delta(\{C, D\})$, *the optimal transfer scheme is*

$$T_z^{\pi} = \frac{\Delta u_z}{2 + \lambda} \equiv T_z^O.$$

[29] This prediction is consistent with the experimental evidence of Fudenberg et al. (2012) who find that noise increases leniency in the repeated prisoners' dilemma with imperfect public monitoring.

[30] In the negative externality case, this implies that there exist equilibria such that $\pi(C)$ approaches 1 as the environment approaches perfect monitoring.

[31] In a related spirit, Andreoni and Samuelson (2006) show that social preferences support cooperation in a finitely repeated prisoners' dilemma; Carmichael and MacLeod (2003) and MacLeod (2007) argue that social preferences may be rationalized as encoding for efficient equilibrium play in a Nash demand game with sunk costs.

Generically with respect to payoff functions $u : Z \rightarrow \mathbb{R}^2$ (i.e., for an open and dense set under the L_1 topology), there exists a unique equilibrium and it is in pure strategies.

This mode of informal justice essentially follows talionic law: "an eye for an eye, a tooth for a tooth." We refer to it as being *outcome based* and denote by T^O the corresponding transfer scheme. We emphasize two corollaries. The first contrasts patterns of punishment and reward under outcome-based justice to those obtained under intent-based justice.

Corollary 1. *(i) Transfers T^O_z depend on payoffs u_z, and not on signal x.*

(ii) No punitive damages are awarded.

(iii) There can be punishment-without-guilt, i.e., $T^O_z > 0$ even though $\pi(C) = 1$.

The next corollary notes that from an efficiency perspective, outcome-based justice implies a version of the rotten kid theorem (Becker 1974, Ray and Ueda 1996), adjusted for transfer costs.

Corollary 2. *Consider a^* an equilibrium action by player A under transfer scheme T^O. We have that*

$$\begin{aligned}\mathbb{E}[u_A^{T^O} + u_P^{T^O} | a^*] &= \max_{a \in \{C,D\}} \mathbb{E}[u_A^{T^O} + u_P^{T^O} | a] \\ &= \max_{a \in \{C,D\}} \mathbb{E}[u_A + u_P | a] - \frac{\lambda}{2+\lambda} \mathbb{E}[|u_A - u_P| | a].\end{aligned}$$

Indeed, since transfers T^O equalize outcomes realization by realization, the payoff of player A is proportional to the sum of payoffs. Hence player A will take the action maximizing total payoffs.[32] However, because outcome-based justice does not exploit potentially valuable side information x, it makes excessive use of costly transfers, and as a result, informal incentives derived from outcome-based justice remain bounded away from ex ante optimal payoffs (with no expected inequality) by an amount $\lambda/(2+\lambda) \min_{a \in \{C,D\}} \mathbb{E}[|u_A - u_P| | a] - |\mathbb{E}[u_A - u_P | a]|$, even when information becomes arbitrarily good.

8. Discussion

We conclude with a discussion of our modeling choices and of potential applications.

Modeling choices

Our specification of social preferences was motivated by experimental evidence and simplicity. However, our approach extends to other social preferences. We discuss below the pros and cons of alternative specifications.

[32] This result relies on the observability of realized payoffs. With unobserved payoffs (say through private effort) we would not obtain efficient behavior from the agent.

One possibility would be to use a type-based model with "altruistic" and "selfish" types (e.g., Levine 1998, Benabou and Tirole 2006). This specification would also lead to a Bayesian game potentially consistent with multiple equilibria. Its main theoretical advantage is that the principal's preferences may satisfy standard expected utility. In addition, the narrative that people are altruistic or selfish and that we reward them based on our beliefs sounds plausible. The main disadvantage is that the principal must care about unobserved, payoff-irrelevant types and her belief over the distribution of such types matters for the structure of equilibria. This additional unobserved degree of freedom makes inference and prediction difficult. Note also that shifts in payoffs turning negative externalities into positive externalities have no consequences in type-based models, since the signalling value of actions is unaffected.

Our model fits broadly into the framework of psychological game theory (Geanakoplos et al. 1989, Battigalli and Dufwenberg 2009). A simplification of our model is that preferences depend only on first order beliefs and outcomes, rather than on the full hierarchy of beliefs. A corresponding limitation is that we fail to capture some subtle phenomena discussed in the literature on intention-based fairness (see, e.g., Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006). In particular, the principal's optimal transfer scheme is independent of options that are available but not chosen by player A.

Applications

In principle our analysis lets us identify from observable data whether intent- or outcome-based justice is being used. Under intent-based justice (and not under outcome-based justice), rewards and punishments depend on side information and punitive transfers may be used. Under intent-based justice, the nature of equilibria should depend on whether we are in a positive externality or negative externality environment. Positive externalities are associated with multiple pure strategy equilibria. Negative externalities are associated with a smaller range of mixed-strategy equilibria. This prediction speaks directly to the literature on persistent productivity differences across seemingly similar enterprises (Gibbons and Henderson 2012). Our model suggests that these productivity differentials may be driven by heterogeneity in the way firms internalize positive externalities, rather than negative ones.

Multiplicity of equilibria under intent-based justice raises concerns with the use of jury trials. Our theory suggests that the way justice is delivered could be heterogenous, and exhibits large inefficiencies compared to ex ante optimal rules. Such concerns are somewhat alleviated by Propositions 2 and 4: multiplicity is limited in negative externality environments, and intent-based justice becomes approximately efficient in environments where evidence is very informative.

A final practical contribution of our model is to identify informational environments in which ex ante contracting is most valuable (see Appendix S.A for a simple model of endogenous contract completeness). Under intent-based justice, ex ante contracting is least valuable when informative signals are available: strong posterior beliefs increase the principal's willingness to implement transfers ex post. The opposite holds under outcome-based justice. Since outcome-based justice does not exploit side information, ex ante contracting is most valuable when signals are informative.

Appendix A: Proofs

A.1 *Proofs for Sections 2 and 3*

Lemma 5. *Whenever $\alpha < \lambda/(2+\lambda)$, the optimal transfer scheme is identically equal to zero, regardless of behavior distribution π.*

Proof. We denote by 0 the transfer function identically equal to zero. Consider an alternative transfer function $T \neq 0$. Using the fact that for any $(a,b) \in \mathbb{R}^2$, $|a| - |b| \leq |a-b|$ and $|a+b| \leq |a| + |b|$, it follows that

$$V(\pi,T) - V(\pi,0) \leq -\lambda \int_{z\in Z} |T_z| f_\pi(z)\,\mathrm{d}z + \delta\alpha(2+\lambda)\int_{z\in Z} |T_z| f_\pi(z)\,\mathrm{d}z + (1-\delta)\alpha(2+\lambda)\int_{z\in Z} |T_z| f_\pi(z)\,\mathrm{d}z.$$

Hence, $V(\pi,T) - V(\pi,0) < 0$ and the optimal transfer policy is identically equal to zero. □

A.2 *Proofs for Section 4*

Proof of Lemma 2. We first show that optimal transfer schemes exist. Let $\mathcal{M}_{T_{\max}}$ denote the set of measurable functions $T : Z \to \mathbb{R}$ such that $\sup_{z\in Z} |T_z| \leq T_{\max}$. For any $\pi \in \Delta(\{C,D\})$, consider a sequence of transfer functions $(T_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to+\infty} V(\pi, T_n) = \sup_{T\in\mathcal{M}_{T_{\max}}} V(\pi,T)$. Theorem 1a of Komlós (1967) implies that there exists a transfer function $T_\infty \in \mathcal{M}_{T_{\max}}$ such that for every $N \in \mathbb{N}$, T_∞ is the limit in the L_1 sense of convex combinations of $(T_k)_{k\geq N}$. By concavity and continuity of $V(\pi,\cdot)$ under the L_1 norm, it follows that $V(\pi,T_\infty) = \sup_{T\in\mathcal{M}_{T_{\max}}} V(\pi,T)$. Hence, the principal's optimization problem admits a solution.

We show that any solution T to the original optimization problem $\max_T V(\pi,T)$ must satisfy $\mathbb{E}[\Delta u^T|C] \leq 0 \leq \mathbb{E}[\Delta u^T|D]$. First, it cannot be optimal to have $\mathbb{E}[\Delta u^T|a] > 0$ for all $a \in \{C,D\}$ or $\mathbb{E}[\Delta u^T|a] < 0$ for all $a \in \{C,D\}$. Imagine that $\forall a \in \{C,D\}$, $\mathbb{E}[\Delta u^T|a] > 0$. Optimal transfers then solve

$$\max_{T\in\mathcal{M}_{T_{\max}}} \int_Z \{-\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| + (1-\delta)\alpha(2+\lambda)T_z\} f_\pi(z)\,\mathrm{d}z.$$

Since $\lambda > \delta\alpha(2+\lambda)$, this implies that for all z, $T_z \geq 0$. However, this contradicts the assumption that $\mathbb{E}[\Delta u^T|C] > 0$ since $\mathbb{E}[\Delta u|C] \leq 0$. The assumption that $\forall a \in \{C,D\}$, $\mathbb{E}[\Delta u^T|a] < 0$ yields a similar contradiction.

To deal with the remaining cases, we temporarily impose that π be in the interior of $\Delta(\{C,D\})$. Assume that $\mathbb{E}[\Delta u^T|C] \geq 0 \geq \mathbb{E}[\Delta u^T|D]$ with one inequality being strict. The third term of expression (3) simplifies and transfer scheme T must solve

$$\max_{T\in\mathcal{M}_{T_{\max}}}\int_Z\{-\lambda|T_z|-\delta\alpha|\Delta u_z-(2+\lambda)T_z|+(1-\delta)\alpha(2+\lambda)[\pi(C|z)-\pi(D|z)]T_z\}f_\pi(z)\,\mathrm{d}z$$

under constraints

$$\begin{aligned}-\mathbb{E}[\Delta u|C]+\frac{2+\lambda}{\pi(C)}\int_z\pi(C|z)T_zf_\pi(z)\,\mathrm{d}z\le 0;&\quad\left(\mu_C\frac{\pi(C)}{2+\lambda}\right)\\ \mathbb{E}[\Delta u|D]-\frac{2+\lambda}{\pi(D)}\int_z\pi(D|z)T_zf_\pi(z)\,\mathrm{d}z\le 0;&\quad\left(\mu_D\frac{\pi(D)}{2+\lambda}\right),\end{aligned}$$

where $\mu=(\mu_C,\mu_D)\ge 0$ are associated Lagrangian multipliers. A solution T is such that for all z, T_z solves

$$\begin{aligned}&\max_{T_z\in[-T_{\max},T_{\max}]}L(\mu,z,T_z)\\ &\quad=-\lambda|T_z|-\delta\alpha|\Delta u_z-(2+\lambda)T_z| \qquad (9)\\ &\qquad+\underbrace{[(1-\delta)\alpha(2+\lambda)[\pi(C|z)-\pi(D|z)]+\mu_D\pi(D|z)-\mu_C\pi(C|z)]}_{\equiv\gamma_z}T_z,\end{aligned}$$

with $\mu_D\times\mathbb{E}[\Delta u^T|D]=0$ and $\mu_C\times\mathbb{E}[\Delta u^T|C]=0$.

Since $\lambda>\delta\alpha(2+\lambda)$, the first two terms of (9) are minimized at $T_z=0$, implying that for all z, $\gamma_zT_z\ge 0$. Term γ_z can be written as

$$\gamma_z=\left[(1-\delta)\alpha(2+\lambda)-\frac{\mu_C+\mu_D}{2}\right][\pi(C|z)-\pi(D|z)]-\frac{\mu_C-\mu_D}{2}.$$

Let $\kappa\equiv(1-\delta)\alpha(2+\lambda)-(\mu_C+\mu_D)/2$. Assume that $\kappa>0$. Since $T_z>0$ only if $\gamma_zT_z\ge 0$, there exists $\theta>0$ such that for all $z\in Z$, $T_z(f(z|C)-\theta f(z|D))\ge 0$. This implies that

$$\int_Z T_zf(z|C)\,\mathrm{d}z\ge\theta\int_Z T_zf(z|D)\,\mathrm{d}z. \qquad (10)$$

However, $\mathbb{E}[\Delta u^T|C]\ge 0$ implies $\int_Z T_zf(z|C)\,\mathrm{d}z\le 0$ and $\mathbb{E}[\Delta u^T|D]\ge 0$ implies $\int_Z T_zf(z|D)\,\mathrm{d}z\ge 0$. Furthermore, one of these inequalities must be strict, which contradicts (10).

Assume that $\kappa\le 0$. This implies that $(\mu_C+\mu_D)/2\ge(1-\delta)\alpha(2+\lambda)$. If $\mu_C>0$ and $\mu_D>0$, then $\mathbb{E}[\Delta u^T|C]=\mathbb{E}[\Delta u^T|D]=0$ and point (ii) holds. Consider the case where $\mu_D=0$ so that $\mu_C\ge 2(1-\delta)\alpha(2+\lambda)$, $\mathbb{E}[\Delta u^T|C]=0$, and $\mathbb{E}[\Delta u^T|D]<0$. Since $\pi(C|z)-\pi(D|z)\in(-1,1)$, we necessarily have that $\gamma_z=\kappa[\pi(C|z)-\pi(D|z)]-\mu_C/2\le 0$. Hence, for all z, $T_z\le 0$, which contradicts $\mathbb{E}[\Delta u^T|D]<0$. Inversely, consider the case where $\mu_C=0$ so that $\mu_D\ge 2(1-\delta)\alpha(2+\lambda)$, $\mathbb{E}[\Delta u^T|D]=0$, and $\mathbb{E}[\Delta u^T|C]>0$. Since $\pi(C|z)-\pi(D|z)\in(-1,1)$, we necessarily have that $\gamma_z=\kappa[\pi(C|z)-\pi(D|z)]+\mu_D/2\ge 0$. Hence, for all z, $T_z\ge 0$, which contradicts $\mathbb{E}[\Delta u^T|C]>0$. This rules out the case where $\mathbb{E}[\Delta u^T|C]\ge 0\ge\mathbb{E}[\Delta u^T|D]$ with one inequality holding strictly.

A similar reasoning rules out configurations such that $\mathbb{E}[\Delta u^T|D]\ge 0$ and $\mathbb{E}[\Delta u^T|C]\ge 0$, as well as $\mathbb{E}[\Delta u^T|D]\le 0$ and $\mathbb{E}[\Delta u^T|C]\le 0$, with one inequality holding strictly. This concludes the proof of point (ii) when π is interior.

If π is not interior so that $\pi(a)=1$ for $a\in\{C,D\}$, transfer T^{π} is defined as the limit (if it exists) of schemes $T^{\widehat{\pi}}$ for $\widehat{\pi}$ interior and converging to π. The existence of such a limit is proven in Lemma 3. The fact that it also satisfies condition $\mathbb{E}[\Delta u^{T^{\pi}}|D]\geq\mathbb{E}[\Delta u^{T^{\pi}}|C]$ follows from the continuity of expectations under the L_1 norm. □

Proof of Lemma 3. Consider the case where π is in the interior of $\Delta(\{C,D\})$. We know from Lemma 2 that we can restrict our attention to transfer functions T such that $\mathbb{E}[\Delta u^T|D]\geq 0\geq\mathbb{E}[\Delta u^T|C]$. The principal's optimization problem becomes

$$\max_{T\in\mathcal{M}_{T_{\max}}}\int_Z\{-\lambda|T_z|-\delta\alpha|\Delta u_z-(2+\lambda)T_z|+(1-\delta)\alpha(2+\lambda)[\pi(D|z)-\pi(C|z)]T_z\}f_\pi(z)\,\mathrm{d}z$$

under constraints

$$-\mathbb{E}[\Delta u|D]+\frac{2+\lambda}{\pi(D)}\int_Z\pi(D|z)T_zf_\pi(z)\,\mathrm{d}z\leq 0;\quad\left(\mu_D\frac{\pi(D)}{2+\lambda}\right)$$
$$\mathbb{E}[\Delta u|C]-\frac{2+\lambda}{\pi(C)}\int_Z\pi(C|z)T_zf_\pi(z)\,\mathrm{d}z\leq 0;\quad\left(\mu_C\frac{\pi(C)}{2+\lambda}\right),$$

where $\mu=(\mu_D,\mu_C)\geq 0$ are Lagrange multipliers. A solution to this problem is such that for all z, T_z solves

$$\begin{aligned}\max_{T_z\in[-T_{\max},T_{\max}]}&L(\mu,z,T_z)\\ =-\lambda|T_z|&-\delta\alpha|\Delta u_z-(2+\lambda)T_z|\\ &+\underbrace{[(1-\delta)\alpha(2+\lambda)[\pi(D|z)-\pi(C|z)]-\mu_D\pi(D|z)+\mu_C\pi(C|z)]}_{\equiv\gamma_z}T_z,\end{aligned}\tag{11}$$

with $\mu_D\times\mathbb{E}[\Delta u^T|D]=0$ and $\mu_C\times\mathbb{E}[\Delta u^T|C]=0$. Let us show that $\max\{\mu_C,\mu_D\}\leq(1-\delta)\alpha(2+\lambda)$. We first show that

$$\kappa\equiv(1-\delta)\alpha(2+\lambda)-(\mu_D+\mu_C)/2\geq 0.$$

Since, $\lambda>\delta\alpha(2+\lambda)$, for every T_z, $\gamma_zT_z\geq 0$. Term γ_z can be rewritten as

$$\gamma_z=\kappa[\pi(D|z)-\pi(C|z)]-\frac{\mu_D-\mu_C}{2}.$$

Assume that $\kappa<0$. Then there exists $\theta>0$ such that for all $z\in Z$, $T_z(f(z|D)-\theta f(z|C))\leq 0$. This implies that

$$\int_{z\in Z}T_z[f(z|D)-\theta f(z|C)]\,\mathrm{d}z\leq 0\quad\Rightarrow\quad\int_{z\in Z}T_zf(z|D)\,\mathrm{d}z\leq\theta\int_{z\in Z}T_zf(z|C)\,\mathrm{d}z.\tag{12}$$

We distinguish three cases: $\mu_D>0$ and $\mu_C>0$, $\mu_D=0$ and $\mu_C>0$, and $\mu_D>0$ and $\mu_C=0$. Let us begin with the case in which $\mu_D>0$ and $\mu_C>0$. This implies that

$$\int_{z\in Z}T_zf(z|C)\,\mathrm{d}z\leq 0\leq\int_{z\in Z}T_zf(z|D)\,\mathrm{d}z,$$

with one inequality being strict. Of course this contradicts inequality (12). If $\mu_D = 0$ and $\mu_C > 2(1-\delta)\alpha(2+\lambda)$, then

$$\gamma_z = \kappa[\pi(D|z) - \pi(C|z)] + \frac{\mu_C}{2} > 0.$$

It follows that for all $z \in Z$, $T_z \geq 0$, which contradicts $\mu_D > 0$. A similar reasoning shows that $\mu_C > 0$. Altogether, this implies that we must have $\kappa \geq 0$.

We now show that $\mu_C \leq (1-\delta)\alpha(2+\lambda)$. Term γ_z can be written as

$$\gamma_z = [2(1-\delta)\alpha(2+\lambda) - \mu_C - \mu_D]\pi(D|z) - (1-\delta)\alpha(2+\lambda) + \mu_C.$$

We know from the previous argument that the first term is necessarily positive. If we had $\mu_C > (1-\delta)\alpha(2+\alpha)$, then we would have that $\gamma_z > 0$ for all $z \in Z$, which implies that for all $z \in Z$, $T_z \geq 0$, and contradicts $\mu_C > 0$. A symmetric reasoning shows that $\mu_D \leq (1-\delta)\alpha(2+\lambda)$.

Altogether, this also implies that $\kappa > 0$. Indeed, assume that $\kappa = 0$. Term γ_z becomes independent of z, and the optimal transfer scheme has a constant sign. This implies that either $\mu_C = 0$ or $\mu_D = 0$, thereby implying that $(\mu_C + \mu_D)/2 \leq \frac{1}{2}(1-\delta)\alpha(2+\lambda)$ and $\kappa > 0$—a contradiction.

To prove uniqueness we use Propositions 1 and 5 proven below. Proposition 5 implies the result when $\delta > \lambda/(\alpha(2+\lambda))$. Consider now the setting where $\delta < \lambda/(\alpha(2+\lambda))$. Assume that there are two distinct solutions T^1 and T^2 to the principal's optimization problem $\max_T V(\pi, T)$, both taking the threshold form described in Proposition 1. By Assumption 1, if the two transfer schemes are different, they must use different thresholds. By concavity of $V(\pi, \cdot)$, it follows that for every $\rho \in [0,1]$, $\rho T^1 + (1-\rho)T^2$ is also optimal. However, such convex combinations do not take the threshold form described in Proposition 1. This is a contradiction and it follows that there must exist a unique solution to the principal's problem.

We now deal with the case where π is a pure strategy. For simplicity we treat the case where $\pi(C) = 1$. We show that for any sequence of interior $\widehat{\pi}$ converging to pure strategy C, $T^{\widehat{\pi}}$ converges to a unique transfer scheme T^C. If $\delta > \lambda/(\alpha(2+\lambda))$ the result is immediate since by Proposition 5, $T_z^{\widehat{\pi}} = \Delta u_z/(2+\lambda)$ for any interior $\widehat{\pi}$. Consider now the case where $\delta < \lambda/(\alpha(2+\lambda))$. By Proposition 1 and using the fact that

$$\widehat{\pi}(D|z) - \widehat{\pi}(C|z) = \left(\widehat{\pi}(D)\frac{f(z|D)}{f(z|C)} - \widehat{\pi}(C)\right) \Big/ \left(\widehat{\pi}(D)\frac{f(z|D)}{f(z|C)} + \widehat{\pi}(C)\right),$$

transfers $T^{\widehat{\pi}}$ can be expressed as

$$T_z^{\widehat{\pi}} = \begin{cases} 0 & \text{if } f(z|D)/f(z|C) \in (\widehat{\theta}_-^{\Delta}, \widehat{\theta}_+^{\Delta}) \\ -T_{\max} & \text{if } f(z|D)/f(z|C) < \widehat{\theta}_-^{\max} \\ T_{\max} & \text{if } f(z|D)/f(z|C) > \widehat{\theta}_+^{\max} \\ \Delta u_z^+/(2+\lambda) & \text{if } f(z|D)/f(z|C) \in (\widehat{\theta}_+^{\Delta}, \widehat{\theta}_+^{\max}) \\ -\Delta u_z^-/(2+\lambda) & \text{if } f(z|D)/f(z|C) \in (\widehat{\theta}_-^{\max}, \widehat{\theta}_-^{\Delta}) \end{cases}$$

for $(\widehat{\theta}^{\max}_-, \widehat{\theta}^{\Delta}_-, \widehat{\theta}^{\max}_+, \widehat{\theta}^{\Delta}_+)$ in the support of $f(z|D)/f(z|C)$. The set of transfer schemes defined by such thresholds is compact under the L_1 norm, and as $\widehat{\pi}$ approaches C we can extract a subsequence converging to a transfer scheme $\widehat{T}^C$ taking a similar threshold form. This limit scheme must solve $\max_T V(C, T)$, i.e., solve,

$$\max_T \mathbb{E}[-\lambda|T| - \delta\alpha|\Delta u - (2+\lambda)T||C] - (1-\delta)\alpha|\mathbb{E}[\Delta u|C] - (2+\lambda)\mathbb{E}[T|C]|. \tag{13}$$

Any scheme solving (13) is such that $T_z \in \{0, \Delta u_z/(2+\lambda)\}$, T_z takes a constant sign, and $\mathbb{E}[\Delta u|C] - (2+\lambda)\mathbb{E}[T|C] = 0$. The only such transfer policy taking a threshold form is the policy T^C defined by

$$T^C_z = \begin{cases} -\frac{\Delta u_z^-}{2+\lambda} & \text{if } \frac{f(z|C)}{f(z|D)} \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where θ is chosen so that $\mathbb{E}[\Delta u^{T^C}|C] = 0$. Since all converging subsequences converge to T^C, it follows that $T^{\widehat{\pi}}$ converges to T^C under the L_1 norm for any sequence of values $\widehat{\pi}$ approaching C. The case where $\pi(D) = 1$ is essentially identical. □

Proof of Proposition 1. We begin with point (i). By Lemma 3, transfers T_z maximize

$$\begin{aligned} L(z, \mu, T_z) = {} & -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ & + \underbrace{[(1-\delta)\alpha(2+\lambda) - \mu_D]}_{\geq 0}\pi(D|z)T_z - \underbrace{[(1-\delta)\alpha(2+\lambda) - \mu_C]}_{\geq 0}\pi(C|z)T_z. \end{aligned}$$

Hence, since $\mu_D \geq 0$, $T_z > 0$ implies that $(1-\delta)\alpha(2+\lambda)\pi(D|z) \geq \lambda - \delta\alpha(2+\lambda)$, which implies the no-punishment-without-guilt bound on $\pi(D|z)$. A similar reasoning implies the no-reward-without-virtue bound on $\pi(C|z)$.

We now turn to point (ii). Rearranging expression (9), $T^\pi_z \in [-T_{\max}, T_{\max}]$ maximizes the Lagrangian

$$\begin{aligned} L(z, \mu, T_z) &= -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ &\qquad + [(1-\delta)\alpha(2+\lambda)[\pi(D|z) - \pi(C|z)] - \mu_D\pi(D|z) + \mu_C\pi(C|z)]T_z \\ &= -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ &\qquad + \left[\left((1-\delta)\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right)(\pi(D|z) - \pi(C|z)) - \frac{\mu_D - \mu_C}{2}\right]T_z. \end{aligned}$$

Since $(1-\delta)\alpha(2+\lambda) - (\mu_D + \mu_C)/2 > 0$, $L(\mu, z, T_z)$ exhibits increasing differences in T_z and $\pi(D|z) - \pi(C|z)$. The particular form of T^π_z and the existence of thresholds $-1 \leq h^{\max}_- \leq h^{\Delta}_- \leq h^{\Delta}_+ \leq h^{\max}_+ \leq 1$ follow from the fact that L is piecewise linear and necessarily attains its maximum at either 0, $\Delta u_z/(2+\lambda)$, $T_{\max}$ or $-T_{\max}$.

We now show that necessarily, $h^{\max}_- < h^{\Delta}_- < h^{\Delta}_+ < h^{\max}_+$. Transfers T_z maximize

$$\begin{aligned} L(z, \mu, T_z) = {} & -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ & + \left[\left((1-\delta)\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right)(\pi(D|z) - \pi(C|z)) - \frac{\mu_D - \mu_C}{2}\right]T_z. \end{aligned}$$

Since $\delta < \lambda/(\alpha(2+\lambda))$, term $-\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z|$ is strictly minimized at $T_z = 0$ with left and right derivatives ∇_- and ∇_+ such that $\nabla_- > 0 > \nabla_+$. It follows that $T_z > 0$ if and only if

$$\pi(D|z) - \pi(C|z) \geq \left(\frac{\mu_D - \mu_C}{2} - \nabla_+\right) \Big/ \left((1-\delta)\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right) \equiv h_+^\Delta.$$

Similarly $T_z < 0$ if and only if

$$\pi(D|z) - \pi(C|z) \leq \left(\frac{\mu_D - \mu_C}{2} - \nabla_-\right) \Big/ \left((1-\delta)\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right) \equiv h_-^\Delta.$$

Note that

$$h_+^\Delta - h_-^\Delta = \frac{-\nabla_+ + \nabla_-}{(1-\delta)\alpha(2+\lambda) - \frac{\mu_D+\mu_C}{2}} > 0.$$

We now establish that $-1 < h_-^\Delta$ and $h_+^\Delta < 1$. Indeed, if we had $h_+^\Delta \geq 1$, then there would be no state z such that $T_z > 0$, which would imply that $\mu_D = 0$. However, in that case, for z such that $\pi(D|z) - \pi(C|z)$ approaches 1, $L(\mu, z, T_z)$ takes the form

$$L(\mu, z, T_z) \simeq -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| + [(1-\delta)\alpha(2+\lambda)]T_z.$$

This expression is strictly maximized at $T_z > 0$, which is a contradiction. Hence it must be that $h_+^\Delta < 1$. Similar proofs show that $-1 < h_-^\Delta$, $h_-^{\max} < h_-^\Delta$, and $h_+^\Delta < h_+^{\max}$, using the fact that aversion to ex post inequality $|\Delta u_z - (2+\lambda)T_z|$ imposes additional costs when implementing transfers above and beyond realized inequality. Note that we may have $h_-^{\max} = -1$ or $h_+^{\max} = 1$.

Limit schemes T^π as π approaches a C or D were derived in the proof of Lemma 3. □

A.3 *Proofs for Section 5*

Proof of Proposition 2. We first show that there exists no equilibrium such that $\pi(C) = 1$. Indeed if $\pi(C) = 1$, the principal's optimal transfer scheme maximizes

$$-\lambda\mathbb{E}[|T_z||C] - \delta\alpha\mathbb{E}[|\Delta u_z - (2+\lambda)T_z||C] - (1-\delta)\alpha(2+\lambda)|\mathbb{E}[T_z|C]|. \tag{14}$$

Since $\delta < \lambda/(\alpha(2+\lambda))$, expression (14) is maximized by transfer scheme $T \equiv 0$. Under this transfer scheme, player A's expected payoffs satisfy $\mathbb{E}[u_A|C] < \mathbb{E}[u_A|D]$, so that his best response is to play D. Hence, there cannot be an equilibrium such that $\pi(C) = 1$.

Consider environments $(Z_n, f_n)_{n\in\mathbb{N}}$ consistent with payoff structure $(U, f_{|U})$ and approaching perfect monitoring. For a given environment (Z, f), the optimal transfer scheme given $\pi(D) = 1$ takes the form

$$T_z^D = \begin{cases} \frac{\Delta u_z^+}{2+\lambda} & \text{if } \frac{f(z|D)}{f(z|C)} \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where θ is such that $\mathbb{E}[\Delta u^{T^D}|D]=0$. Hence, we have that

$$\mathbb{E}[u_A^{T^D}|D]=\mathbb{E}\left[\frac{u_A^{T^D}+u_P^{T^D}}{2}\Big|D\right]=\mathbb{E}\left[\frac{u_A+u_P}{2}\Big|D\right]-\mathbb{E}\left[\frac{\lambda}{2}|T^D|\Big|D\right]$$
$$=\mathbb{E}\left[\frac{u_A+u_P}{2}\Big|D\right]-\frac{\lambda}{2(2+\lambda)}\mathbb{E}[\Delta u|D].$$

In turn, using the fact that $\mathbb{E}[u_A|C]=\mathbb{E}[u_P|C]$, we have that

$$\mathbb{E}[u_A^{T^D}|C]=\mathbb{E}[u_A-(1+\lambda)T_z^D|C]$$
$$=\mathbb{E}\left[\frac{u_A+u_P}{2}\Big|C\right]-\frac{1+\lambda}{2+\lambda}\int_Z\Delta u_z^+\mathbf{1}_{f(z|D)/f(z|C)\geq\theta}f(z|C)\,\mathrm{d}z.$$

Let us show that as n grows large, the corresponding threshold θ_n grows arbitrarily large as well. Indeed, for any $s>0$, define

$$H(s)\equiv\int_{z\in Z}\Delta u_z f(z|D)\,\mathrm{d}z-\int_{z\in Z}\Delta u_z^+\mathbf{1}_{f(z|D)/f(z|C)>s}f(z|D)\,\mathrm{d}z.$$

The term $H(s)$ is increasing in s and threshold θ is defined by the equation $H(\theta)=0$. We now show that for any $s>0$, as n grows large, $H(s)<0$. Indeed

$$H(s)\leq-\int_{z\in Z}(u_A-u_P)^-f_n(z|D)\,\mathrm{d}z+\int_{z\in Z}(u_A-u_P)^+\mathbf{1}_{f_n(z|D)/f_n(z|C)<s}f_n(z|D)\,\mathrm{d}z$$
$$\leq-\int_{z\in Z}(u_A-u_P)^-f_n(z|D)\,\mathrm{d}z+\underbrace{T_{\max}\times\operatorname{prob}\left(\frac{f_n(z|D)}{f_n(z|C)}<s\Big|D\right)}_{\to0\text{ as }n\to\infty},$$

where we used the assumption that $\max_z|\Delta u_z|\leq T_{\max}$. This implies that θ_n must grow arbitrarily large as n goes to infinity.

Noting that

$$\int_{z\in Z}\Delta u_z^+\mathbf{1}_{f_n(z|D)/f_n(z|C)\geq\theta_n}f_n(z|C)\,\mathrm{d}z=\int_{z\in Z}\Delta u_z^+\mathbf{1}_{f_n(z|D)/f_n(z|C)\geq\theta_n}\frac{f_n(z|C)}{f_n(z|D)}f_n(z|D)\,\mathrm{d}z$$
$$\leq\frac{T_{\max}}{\theta_n},$$

it follows that as n grows large, $\mathbb{E}[u_A^{T^D}|C]$ converges to $\mathbb{E}[(u_A+u_P)/2|C]$. Hence, whenever $\mathbb{E}[u_A+u_P|C]-\mathbb{E}[u_A+u_P|D]>-\lambda/(2+\lambda)\mathbb{E}[\Delta u|D]$, for n large enough, $\mathbb{E}_{f_n}[u_A^{T_D}|C]-\mathbb{E}_{f_n}[u_A^{T_D}|D]>0$. This contradicts the existence of an equilibrium such that $\pi(D)=1$. □

Proof of Proposition 3. Point (i) is immediate. If $\pi(D)=1$, the principal's optimal transfer scheme maximizes

$$-\lambda\mathbb{E}[|T_z||C]-\delta\alpha\mathbb{E}[|\Delta u_z-(2+\lambda)T_z||C]-(1-\delta)\alpha(2+\lambda)|\mathbb{E}[T_z|C]|. \tag{15}$$

Since $\delta < \lambda/(\alpha(2+\lambda))$, expression (15) is maximized for a transfer scheme $T \equiv 0$. Player A's expected payoffs satisfy $\mathbb{E}[u_A|C] < \mathbb{E}[u_A|D]$ and her best response is to play D. Hence, (π, T) with $\pi(D) = 1$ and $T = 0$ is an equilibrium.

Consider environments $(Z_n, f_n)_{n\in\mathbb{N}}$ consistent with payoff structure $(U, f_{|U})$ and approaching perfect monitoring.

For a given environment (Z, f), the optimal transfer scheme given $\pi(C) = 1$ takes the form

$$T_z^C = \begin{cases} -\frac{\Delta u_z^-}{2+\lambda} & \text{if } \frac{f(z|C)}{f(z|D)} \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where θ is chosen so that $\mathbb{E}[\Delta u^{T^C}|C] = 0$. Hence player A's payoffs conditional on actions C and D are

$$\begin{aligned} \mathbb{E}[u_A^{T^C}|C] &= \mathbb{E}\left[\frac{u_A^{T^C} + u_P^{T^C}}{2}\Big|C\right] = \mathbb{E}\left[\frac{u_A + u_P}{2}\Big|C\right] - \mathbb{E}\left[\frac{\lambda}{2}|T^C|\Big|C\right] \\ &= \mathbb{E}\left[\frac{u_A + u_P}{2}\Big|C\right] - \frac{\lambda}{2(2+\lambda)}\mathbb{E}[\Delta u|C]. \end{aligned}$$

In turn, using the fact that $\mathbb{E}[u_A|D] = \mathbb{E}[u_P|D]$, we have that

$$\mathbb{E}[u_A^{T^C}|D] = \mathbb{E}[u_A - T^C|D] = \mathbb{E}\left[\frac{u_A + u_P}{2}\Big|D\right] + \frac{1}{2+\lambda}\int_Z \Delta u_z^- \mathbf{1}_{f(z|C)/f(z|D)\geq\theta} f(z|D)\,\mathrm{d}z.$$

Let us show that as n grows large, threshold θ_n grows arbitrarily large as well. Indeed, for any $s > 0$, define

$$H(s) \equiv \int_{z\in Z} \Delta u_z f(z|C)\,\mathrm{d}z + \int_{z\in Z} \Delta u_z^- \mathbf{1}_{f(z|C)/f(z|D)>s} f(z|C)\,\mathrm{d}z.$$

The term $H(s)$ is decreasing in s and threshold θ is defined by the equation $H(\theta) = 0$. We now show that for any $s > 0$, as n grows large, $H(s) > 0$. Indeed

$$\begin{aligned} H(s) &\geq \int_{z\in Z} (u_A - u_P)^+ f_n(z|C)\,\mathrm{d}z - \int_{z\in Z} (u_A - u_P)^- \mathbf{1}_{f_n(z|C)/f_n(z|D)<s} f_n(z|C)\,\mathrm{d}z \\ &\geq \int_{z\in Z} (u_A - u_P)^+ f_n(z|C)\,\mathrm{d}z + \underbrace{T_{\max} \times \operatorname{prob}\left(\frac{f_n(z|C)}{f_n(z|D)} < s\Big|C\right)}_{\to 0 \text{ as } n\to\infty}. \end{aligned}$$

This implies that θ_n must grow arbitrarily large as n goes to infinity.

Noting that

$$\begin{aligned} \int_{z\in Z} \Delta u_z^- \mathbf{1}_{f_n(z|C)/f_n(z|D)\geq\theta_n} f_n(z|D)\,\mathrm{d}z &= \int_{z\in Z} \Delta u_z^+ \mathbf{1}_{f_n(z|D)/f_n(z|C)\geq\theta_n} \frac{f_n(z|C)}{f_n(z|D)} f_n(z|D)\,\mathrm{d}z \\ &\leq \frac{T_{\max}}{\theta_n}, \end{aligned}$$

it follows that as n grows large, $\mathbb{E}[u_A^{T^C}|D]$ converges to $\mathbb{E}[(u_A+u_P)/2|D]$. Hence, whenever $\mathbb{E}[u_A+u_P|C]-\mathbb{E}[u_A+u_P|D] > -\lambda/(2+\lambda)\mathbb{E}[\Delta u|C]$, for n large enough, $\mathbb{E}_{f_n}[u_A^{T_C}|C]-\mathbb{E}_{f_n}[u_A^{T_C}|D] > 0$. This implies that there exists an equilibrium such that $\pi(D)=1$. □

A.4 *Proofs for Section 6*

Proof of Proposition 4. We begin with point (i). It follows from Lemma 3 that for all $z \in Z$, transfer T_z^π must solve

$$\max_{T_z} -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ + \underbrace{[(1-\delta)\alpha(2+\lambda)[\pi(D|z)-\pi(C|z)] - \mu_D\pi(D|z) + \mu_C\pi(C|z)]}_{\equiv\gamma_z} T_z,$$

with $\max\{\mu_C,\mu_D\} \le (1-\delta)\alpha(2+\lambda)$. We have that $T_z > 0$ if and only if $\gamma_z \ge \lambda - \delta\alpha(2+\lambda) > 0$. Coefficient γ_z satisfies

$$\begin{aligned}\gamma_z &= [2(1-\delta)\alpha(2+\lambda)-\mu_C-\mu_D]\pi(D|z) - (1-\delta)\alpha(2+\lambda)+\mu_C \\ &\le 2(1-\delta)\alpha(2+\lambda)\pi(D|z).\end{aligned}$$

Hence, a necessary condition to have $T_z > 0$ is that

$$\pi(D|z) \ge \frac{\lambda-\delta\alpha(2+\lambda)}{2(1-\delta)\alpha(2+\lambda)} \quad\Longleftrightarrow\quad \frac{f(z|D)}{f(z|C)} \ge \frac{1}{\hbar}\frac{\pi(C)}{\pi(D)}.$$

Using the Bienaymé–Chebyshev inequality, this implies that

$$\text{prob}(T_z^\pi > 0|D) \le \hbar\frac{\pi(D)}{\pi(C)}\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\Big|D\right]. \tag{16}$$

Inequality (16) implies bounds on the frequency with which action C can be sustained in equilibrium. Take π and the induced transfer scheme T^π as given. Player A chooses to cooperate if and only if $\mathbb{E}[u_A^{T^\pi}|C] \ge \mathbb{E}[u_A^{T^\pi}|D]$. Briefly dropping the π superscript on transfers, this is equivalent to

$$\int_{z\in Z}\left(u_A - \frac{1+\lambda}{2+\lambda}T_z^+ + T_z^-\right)f(z|C)\,\mathrm{d}z \ge \int_{z\in Z}\left(u_A - \frac{1+\lambda}{2+\lambda}T_z^+ + T_z^-\right)f(z|D)\,\mathrm{d}z.$$

Hence, action C is incentive compatible if and only if,

$$\underbrace{-\int_{z\in Z}\frac{1+\lambda}{2+\lambda}T_z^+[f(z|C)-f(z|D)]\,\mathrm{d}z}_{\equiv K_0} + \underbrace{\int_{z\in Z}T_z^-[f(z|C)-f(z|D)]\,\mathrm{d}z}_{\equiv K_1} \\ \ge \int_{z\in Z}u_A[f(z|D)-f(z|C)]\,\mathrm{d}z.$$

We establish upper bounds on K_0 and K_1:

$$\begin{aligned} K_0 &\leq \frac{1+\lambda}{2+\lambda}\int_{z\in Z} T_z^+ f(z|D)\,\mathrm{d}z \leq T_{\max}\,\mathrm{prob}(T_z > 0|D) \\ &\leq T_{\max}\hbar\frac{\pi(D)}{\pi(C)}\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\Big|D\right]. \end{aligned}$$

From Lemma 3 we know that $\int_Z [\Delta u_z + (2+\lambda)(T_z^- - T_z^+)] f(z|C)\,\mathrm{d}z \leq 0$. This implies that

$$K_1 \leq -\frac{1}{2+\lambda}\mathbb{E}[\Delta u|C] + \int_{z\in Z} T_z^+ f(z|C)\,\mathrm{d}z.$$

Using Bienaymé–Chebyshev results and noting that $\mathbb{E}[f(z|D)/f(z|C)|C] = 1$, we obtain that

$$K_1 \leq -\frac{1}{2+\lambda}\mathbb{E}[\Delta u|C] + T_{\max}\frac{\pi(D)}{\pi(C)}\hbar.$$

Altogether, this implies that a necessary condition for action C to be incentive compatible is

$$\frac{\pi(D)}{\pi(C)}\hbar T_{\max}\left(1+\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\Big|D\right]\right) \geq \mathbb{E}[u_A|D] - \mathbb{E}[u_A|C] + \frac{1}{2+\lambda}\mathbb{E}[\Delta u|C].$$

This concludes the proof of point (i).

We now turn to the proof of point (ii). Fix some interior value $\pi(C) \in (0,1)$. We denote $T^{\pi,n}$ the corresponding transfer scheme in environment (Z_n, f_n). We first establish the following property of transfer schemes $T^{\pi,n}$ as n grows large: for all $\epsilon > 0$, there exists $N > 0$ large enough such that for all $n \geq N$,

$$\begin{aligned} |\mathbb{E}[\Delta u^{T^{\pi,n}}|C]| &\leq \epsilon \\ |\mathbb{E}[\Delta u^{T^{\pi,n}}|D]| &\leq \epsilon \\ \mathrm{prob}(z \text{ s.t. } T_z^{\pi,n} \notin [-\Delta u_z^-, 0]|a = C) &\leq \epsilon \\ \mathrm{prob}(z \text{ s.t. } T_z^{\pi,n} \notin [0, \Delta u_z^+]|a = D) &\leq \epsilon. \end{aligned}$$

Consider the principal's value function $V(a,T)$. Transfer schemes $\widehat{T}^D$ and $\widehat{T}^C$ (which may differ from schemes T^C and T^D defined in footnote 24) respectively solve $\max_T V(D,T)$ and $\max_T V(C,T)$ if and only if

$$\begin{aligned} |\mathbb{E}[\Delta u^{\widehat{T}^C}|C]| &= 0 \\ |\mathbb{E}[\Delta u^{\widehat{T}^D}|D]| &= 0 \\ \mathrm{prob}(z \text{ s.t. } \widehat{T}_z^C \notin [-(\Delta u_z)^-, 0]|a = C) &= 0 \\ \mathrm{prob}(z \text{ s.t. } \widehat{T}_z^D \notin [0, (\Delta u_z)^+]|a = D) &= 0. \end{aligned} \tag{17}$$

Furthermore keeping distribution over payoffs (u_A, u_P) constant, one can pick respective solutions $\widehat{T}^C$ and $\widehat{T}^D$ that are independent of side information x and of index n.

For any T, the principal's value function is

$$V(\pi, T) = \sum_{a\in\{C,D\}} \pi(a)V(a, T).$$

Let $T^{CD,n}$ be defined by

$$T_z^{CD,n} = \begin{cases} \widehat{T}_z^C & \text{if } \frac{f_n(z|C)}{f_n(z|D)} \geq 2 \\ \widehat{T}_z^D & \text{if } \frac{f_n(z|D)}{f_n(z|C)} \geq 2 \\ 0 & \text{otherwise.} \end{cases}$$

For any $a \in \{C, D\}$ (denoting by $\neg a$ the other action) we have that

$$\begin{aligned} V(a, T^{CD,n}) &\geq V(a, \widehat{T}^a) - [\lambda + \alpha(2+\lambda)] \int_Z |T^{CD,n} - \widehat{T}^a| f_n(z|a)\,\mathrm{d}z \\ &\geq V(a, \widehat{T}^a) - 2T_{\max}[\lambda + \alpha(2+\lambda)]\,\mathrm{prob}\left(\frac{f_n(z|a)}{f_n(z|\neg a)} < 2 \middle| a\right). \end{aligned}$$

By optimality of $T^{\pi,n}$, $V(\pi, T^{\pi,n}) \geq V(\pi, T^{CD,n})$. Using the fact that as n grows to infinity, $\mathrm{prob}(f_n(z|a)/f_n(z|\neg a) < 2|a)$ goes to zero, we obtain that

$$\liminf_{n\to\infty} \sum_{a\in\{C,D\}} \pi(a)V(a, T^{\pi,n}) \geq \sum_{a\in\{C,D\}} \pi(a)V(a, \widehat{T}^a).$$

Hence, since $\pi(C) \in (0, 1)$ is fixed, for n sufficiently large, we have that $T^{\pi,n}$ must approximately solve $\max_T V(C, T)$ and $\max_T V(D, T)$, which implies that there exists N sufficiently large such that for all $n \geq N$,

$$\begin{aligned} |\mathbb{E}[\Delta u^{T^{\pi,n}}|C]| &\leq \epsilon \\ |\mathbb{E}[\Delta u^{T^{\pi,n}}|D]| &\leq \epsilon \\ \mathrm{prob}(z \text{ s.t. } T_z^{\pi,n} \notin [-\Delta u_z^-, 0]|a = C) &\leq \epsilon \\ \mathrm{prob}(z \text{ s.t. } T_z^{\pi,n} \notin [0, \Delta u_z^+]|a = D) &\leq \epsilon. \end{aligned}$$

Otherwise, one could extract sequences of transfer schemes $T^{\pi,n}$ converging to solutions of $\max_T V(D, T)$ and $\max_T V(C, T)$ that do not satisfy (17).

Since $u_A^T = (u_A^T + u_P^T)/2 + \Delta u^T/2$, player A's choice under transfer scheme $T^{\pi,n}$ solves

$$\max_{a\in\{C,D\}} \mathbb{E}(u_A^{T^{\pi,n}}|a) = \frac{1}{2}\mathbb{E}(u_A + u_P - \lambda|T^{\pi,n}||a) + \frac{1}{2}\mathbb{E}[\Delta u^{T^{\pi,n}}|a].$$

We know that for any $\epsilon > 0$, $|\mathbb{E}[\Delta u^{T^{\pi,n}}|a]| \leq \epsilon$ for n large enough. Furthermore, by assumption, $\mathbb{E}[u_A + u_P|C] - \lambda/(2+\lambda)\mathbb{E}[|u_A - u_P||C] > \mathbb{E}[u_A + u_P|D]$. Altogether, this implies that for n large enough, transfer $T^{\pi,n}$ induces the agent to take action C. By continuity of mapping Γ_n, this implies that $\overline{\pi}_n(C) \geq \pi(C)$. Since π was chosen arbitrarily, this implies that $\lim_{n\to\infty} \overline{\pi}_n(C) = 1$. Since $T^{\overline{\pi}_n,n}$ solves $\max_T V(\overline{\pi}_n, T)$, it follows that transfers approach ex ante efficiency as n goes to infinity. □

A.5 *Proofs for Section 7*

Proof of Proposition 5. Transfer scheme T^O is the unique minimizer of the term

$$-\lambda \int_{z\in Z} |T_z| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_{z\in Z} |\Delta u_z - (2+\lambda)T_z| f_\pi(z)\,\mathrm{d}z$$

in expression (3). In addition, for this transfer policy, we have by construction that $\mathbb{E}[\Delta u^{T^O}|D] = \mathbb{E}[\Delta u^{T^O}|C] = 0$, which implies that the term

$$-(1-\delta)\alpha \sum_{a\in\{C,D\}} \left| \int_{z\in Z} [\Delta u_z - (2+\lambda)T_z]\pi(a|z) f_\pi(z)\,\mathrm{d}z \right|$$

in objective function (3) is also minimized.[33] Hence, T^O is indeed the unique minimizer of $V(\pi, \cdot)$ for every distribution π.

We now establish that generically with respect to payoffs, $\mathbb{E}[u_A^{T^O}|C] - \mathbb{E}[u_A^{T^O}|D] \neq 0$. By continuity it follows that if $\mathbb{E}[u_A^{T^O}|C] - \mathbb{E}[u_A^{T^O}|D] \neq 0$ for payoff functions u_A, u_P, then $\mathbb{E}[\hat{u}_A^{T^O}|C] - \mathbb{E}[\hat{u}_A^{T^O}|D] \neq 0$ for all sufficiently close payoff functions of $\hat{u}_A$, $\hat{u}_P$. Inversely, if $\mathbb{E}[u_A^{T^O}|C] - \mathbb{E}[u_A^{T^O}|D] = 0$, pick $\epsilon > 0$, and keeping the distribution over $z \in Z$ constant, consider the modified payoffs $\hat{u}_A$ and $\hat{u}_P$ defined by

$$\begin{aligned} \hat{u}_A(z) &= u_A(z) + \epsilon, & \hat{u}_P(z) &= u_P(z) + \epsilon & &\text{if } f(z|C) \geq f(z|D) \\ \hat{u}_A(z) &= u_A(z) - \epsilon, & \hat{u}_P(z) &= u_P(z) - \epsilon & &\text{if } f(z|C) < f(z|D). \end{aligned}$$

By construction, it follows that $\mathbb{E}[\hat{u}_A^{T^O}|C] - \mathbb{E}[\hat{u}_A^{T^O}|D] > 0$, which concludes the proof. □

Proof of Corollary 2. Under transfer scheme T^O, by construction $u_A^{T^O} = u_P^{T^O}$, so that $u_A^{T^O} = (u_A^{T^O} + u_P^{T^O})/2 = (u_A + u_P - \lambda|T^O|)/2$. Hence, any action a^O that solves $\max_{a\in\{C,D\}} \mathbb{E}(u_A^{T^O}|a)$ must also maximize $S(a, T^O) = \mathbb{E}(u_A + u_P - 2\underline{u} - \lambda|T^O||a)$. □

References

Andreoni, James and Larry Samuelson (2006), "Building rational cooperation." *Journal of Economic Theory*, 127, 117–154. [1161]

Baker, George, Robert Gibbons, and Kevin J. Murphy (1994), "Subjective performance measures in optimal incentive contracts." *The Quarterly Journal of Economics*, 109, 1125–1156. [1146, 1147]

Baker, George, Robert Gibbons, and Kevin J. Murphy (2002), "Relational contracts and the theory of the firm." *The Quarterly Journal of Economics*, 117, 39–84. [1146, 1147]

Battigalli, Pierpaolo and Martin Dufwenberg (2009), "Dynamic psychological games." *Journal of Economic Theory*, 144, 1–35. [1163]

[33] Note that term $\int_Z (u_A + u_P) f_\pi(z)\,\mathrm{d}z$ is independent of T.

Becker, Gary (1974), "A theory of social interactions." *Journal of Political Economy*, 82, 1063–1093. [1162]

Benabou, Roland and Jean Tirole (2003), "Intrinsic and extrinsic motivation." *Review of Economic Studies*, 70, 489–520. [1148]

Benabou, Roland and Jean Tirole (2006), "Incentives and prosocial behavior." *American Economic Review*, 96, 1652–1678. [1146, 1163]

Board, Simon (2011), "Relational contracts and the value of loyalty." *American Economic Review*, 101, 3349. [1147]

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser (2008), "Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States." *American Economic Review*, 98, 294–310. [1151]

Bohnet, Iris and Richard Zeckhauser (2004), "Trust, risk and betrayal." *Journal of Economic Behavior & Organization*, 55, 467–484. [1150, 1151]

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels (2005), "Fair procedures: Evidence from games involving lotteries." *The Economic Journal*, 115, 1054–1076. [1150]

Bull, Clive (1987), "The existence of self-enforcing implicit contracts." *The Quarterly Journal of Economics*, 102, 147–159. [1146, 1147]

Carmichael, Lorne H. and W. Bentley MacLeod (2003), "Caring about sunk costs: A behavioral solution to holdup problems with small stakes." *Journal of Law, Economics, and Organization*, 19, 106. [1161]

Charness, Gary (2004), "Attribution and reciprocity in an experimental labor market." *Journal of Labor Economics*, 22, 665–688. [1148]

Charness, Gary and David I. Levine (2007), "Intention and stochastic outcomes: An experimental study." *The Economic Journal*, 117, 1051–1072. [1150]

Charness, Gary and Matthew Rabin (2002), "Understanding social preferences with simple tests." *The Quarterly Journal of Economics*, 117, 817–869. [1146]

Chassang, Sylvain (2010), "Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts." *American Economic Review*, 100, 448–465. [1147]

Compte, Olivier and Andrew Postlewaite (2015), "Plausible cooperation." *Games and Economic Behavior*, 91, 45–59. [1147]

Cushman, Fiery, Anna Dreber, Ying Wang, and Jay Costa (2009), "Accidental outcomes guide punishment in a "trembling hand" game." *PLoS ONE*, 4, e6699. [1150]

Dufwenberg, Martin and Georg Kirchsteiger (2004), "A theory of sequential reciprocity." *Games and Economic Behavior*, 47, 268–298. [1146, 1163]

Ellingsen, Tore and Magnus Johannesson (2008), "Pride and prejudice: The human side of incentive theory." *American Economic Review*, 98, 990–1008. [1148]

Englmaier, Florian and Achim Wambach (2010), "Optimal incentive contracts under inequity aversion." *Games and Economic Behavior*, 69, 312–328. [1148]

Falk, Armin and Urs Fischbacher (2006), "A theory of reciprocity." *Games and Economic Behavior*, 54, 293–315. [1146, 1163]

Falk, Armin and Simon Gaechter (2002), "Reputation and reciprocity—consequences for the labour relation." *Scandinavian Journal of Economics*, 104, 1–26. [1148]

Falk, Armin and Michael Kosfeld (2006), "The hidden cost of control." *American Economic Review*, 96, 1611–1630. [1148]

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger (1997), "Reciprocity as a contract enforcement device: Experimental evidence." *Econometrica*, 65, 833–860. [1148]

Fehr, Ernst, Oliver D. Hart, and Christian Zehnder (2011), "Contracts as reference points—experimental evidence." *American Economic Review*, 101, 493–525. [1148]

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt (2007), "Fairness and contract design." *Econometrica*, 75, 121–154. [1148]

Fehr, Ernst and Klaus M. Schmidt (1999), "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114, 817–868. [1146, 1150]

Folger, Robert and Mary A. Konovsky (1989), "Effects of procedural and distributive justice on reactions to pay raise decisions." *Academy of Management Journal*, 32, 115–130. [1152]

Fong, Yuk-fai and Jin Li (2010), "Relational contracts, efficiency wages, and employment dynamics." Northwestern University, Kellogg School of Management. [1147]

Fudenberg, Drew and David K. Levine (2012), "Fairness, risk preferences and independence: Impossibility theorems." *Journal of Economic Behavior & Organization*, 81, 606–612. [1150]

Fudenberg, Drew, David Rand, and Anna Dreber (2012), "Slow to anger and fast to forgive: Cooperation in an uncertain world." *American Economic Review*, 102, 720–749. [1161]

Fudenberg, Drew and Jean Tirole (1990), "Moral hazard and renegotiation in agency contracts." *Econometrica*, 58, 1279–1319. [1148]

Geanakoplos, John, David Pearce, and Ennio Stacchetti (1989), "Psychological games and sequential rationality." *Games and Economic Behavior*, 1, 60–79. [1151, 1163]

Gibbons, Robert and Rebecca Henderson (2012), "What do managers do?: Exploring persistent performance differences among seemingly similar enterprises." Harvard Business School. [1163]

Gneezy, Uri and Aldo Rustichini (2000), "Pay enough or don't pay at all." *Quarterly Journal of Economics*, 115, 791–810. [1148]

Green, Edward J. and Robert H. Porter (1984), "Noncooperative collusion under imperfect price information." *Econometrica*, 52, 87–100. [1146, 1147]

Greenberg, Jerald (1990), "Organizational justice: Yesterday, today, and tomorrow." *Journal of Management*, 16, 399. [1152]

Halac, Marina (2012), "Relational contracts and the value of relationships." *American Economic Review*, 102, 750–779. [1147]

Harris, Milton and Artur Raviv (1979), "Optimal incentive contracts with imperfect information." *Journal of Economic Theory*, 20, 231–259. [1147]

Hart, Oliver D. and John H. Moore (2008), "Contracts as reference points." *Quarterly Journal of Economics*, 123, 1–48. [1148]

Hawranek, Dietmar (2010), "Czech headache: Skoda spells trouble for parent company Volkswagen." *Der Spiegel*. [1152]

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker (2006), "Costly punishment across human societies." *Science*, 312, 1767. [1156, 1157]

Holmström, Bengt (1979), "Moral hazard and observability." *The Bell Journal of Economics*, 10, 74–91. [1147]

Komlós, Janos (1967), "A generalization of a problem of Steinhaus." *Acta Mathematica Hungarica*, 18, 217–229. [1164]

Konovsky, Mary A. (2000), "Understanding procedural justice and its impact on business organizations." *Journal of Management*, 26, 489. [1152]

Kőszegi, Botond and Matthew Rabin (2006), "A model of reference-dependent preferences." *The Quarterly Journal of Economics*, 121, 1133–1165. [1151]

Krawczyk, Michal W. (2011), "A model of procedural and distributive fairness." *Theory and Decision*, 70, 111–128. [1150]

Krawczyk, Michal W. and Fabrice Le Lec (2010), "Give me a chance! An experiment in social decision under risk." *Experimental Economics*, 13, 500–511. [1150]

Levin, Jonathan (2003), "Relational incentive contracts." *American Economic Review*, 93, 835–857. [1146, 1147]

Levine, David K. (1998), "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1, 593–622. [1146, 1163]

Li, Jin and Niko Matouschek (2013), "Managing conflicts in relational contracts." *American Economic Review*, 103, 2328–2351. [1147]

Lind, E. Allan, Carol T. Kulik, Maureen Ambrose, and Maria V. de Vera Park (1993), "Individual and corporate dispute resolution: Using procedural fairness as a decision heuristic." *Administrative Science Quarterly*, 38, 224–251. [1152]

Machina, Mark J. (1989), "Dynamic consistency and non-expected utility models of choice under uncertainty." *Journal of Economic Literature*, 27, 1622–1668. [1146, 1150]

MacLeod, W. Bentley (2003), "Optimal contracting with subjective evaluation." *American Economic Review*, 93, 216–240. [1146]

MacLeod, W. Bentley (2007), "Can contract theory explain social preferences?" *American Economic Review*, 97, 187–192. [1161]

MacLeod, W. Bentley and James M. Malcomson (1989), "Implicit contracts, incentive compatibility, and involuntary unemployment." *Econometrica*, 57, 447–480. [1146, 1147]

Rabin, Matthew (1993), "Incorporating fairness into game theory and economics." *American Economic Review*, 83, 1281–1302. [1146, 1163]

Ray, Debraj and Kaoru Ueda (1996), "Egalitarianism and incentives." *Journal of Economic Theory*, 71, 324–348. [1162]

Robles, Frances (2014), "Jury awards $23.6 billion in Florida smoking case." *The New York Times*. [1152]

Saito, Kota (2013), "Social preferences under risk: Equality of opportunity versus equality of outcome." *American Economic Review*, 103, 3084–3101. [1150]

Schächtele, Simeon, Tobias Gerstenberg, and David Lagnado (2011), "Beyond outcomes: The influence of intentions and deception." Working Paper, UCL. [1150]

Velez v. Novartis (2010) "04-cv-09194." U.S. District Court, Southern District of New York (Manhattan). Available at www.americanbar.org/content/dam/aba/administrative/labor_law/meetings/2011/eeo/029.pdf. [1152]

Co-editor George J. Mailath handled this manuscript.

Submitted 2015-1-7. Final version accepted 2015-11-24. Available online 2015-12-7.