# Rewards and Punishments: Informal Contracting through Social Preferences[*]

Sylvain Chassang — Princeton University

Christian Zehnder[†] — University of Lausanne

October 7, 2014.

**Abstract**

This paper develops a novel 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, subjective performance evaluation, incomplete contracts, heuristics.

[*] 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, Klaus Schmidt as well as seminar participants at Bern, Bonn, Chicago Booth, Columbia, 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 No. 26051096).

[†] Chassang: chassang@princeton.edu, Zehnder: christian.zehnder@unil.ch.

# 1 Introduction

This paper develops a novel 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, but instead implements transfers that maximize her social preferences ex post. This may be because the principal does not have commitment power, because she is not physically present at the ex ante stage, or because specifying fully contingent contracts requires excessive cognitive bandwidth. We show that social preferences impose intuitively plausible restrictions on patterns of rewards and punishments, not captured by existing models of informal incentives such as relational contracts.[2] These restrictions have novel implications about the way different types of externalities are internalized.

The game-form we use is straightforward. A principal can transfer resources between an active agent, referred to as player $A$, and a broader group modeled as a single passive player $P$. Player $A$ takes a private action $a \in \{C, D\}$ which induces stochastic payoffs for himself and passive player $P$. Action $C$ can be thought of as a pro-social action that increases the expected payoff of player $P$ at the expense of player $A$. The principal observes realized payoffs to the different players, 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): circumstancial evidence of player $A$'s behavior is available to the principal, but is not usable by an external court.[3] Transfers between players have an efficiency cost: the cost to the transferring party exceeds

[1] Throughout we refer to the agent as "he", and to the principal as "she." In our model, the broader group is modeled as a passive player that takes no action and matters only through the principal's social preferences.

[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.

the value transferred to the receiving party. These costs are modeled in a reduced-form way, and reflect inefficiencies in the reallocation of 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 which in turn determines player $A$'s behavior. Given a specification of social preferences, we refer to the corresponding transfer rule as a *mode of informal justice*.

Our main modeling assumption is that the principal determines transfers ex post based on her sense of fairness. Social preferences are crucial to our model, because a principal exclusively concerned with efficiency would never impose costly transfers. In deterministic decision problems, our specification of the principal's social preferences coincides 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 model of social preferences places weight on both ex post (or allocative) fairness, and ex ante (or procedural) fairness. Allowing for preferences over ex ante fairness makes this a non-expected-utility model and the principal need not be consequentialist (Machina, 1989): the fairness of an unequal realized allocation depends on the fairness of the underlying lottery which 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. When the principal places a high weight on ex post fairness, rewards and punishments follow what we refer to as *outcome-based justice*: transfers depend only on payoff outcomes, ignoring all side information; there is no punitive justice, in the sense that transfers at most compensate for realized inequality; and informal incentives induce a generically unique pure strategy

[4] We pick the Fehr-Schmidt model as the basis for the principal's preferences mainly because of its parsimony and tractability. Other models of social preferences include: Bolton and Ockenfels (2000) (inequity aversion); Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006) (fairness of intentions); Levine (1998) (type-based reciprocity); Charness and Rabin (2002) (preferences for social welfare); Benabou and Tirole (2006), Ellingsen and Johannesson (2007) (concerns for social reputation or self-respect).

equilibrium. In the complementary case where the principal places a high weight on ex ante fairness, rewards and punishment 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, our model implies novel and plausible restrictions on informal incentives which we refer to as *no-punishment-without-guilt*. Specifically, we show that 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 to have behaved in a pro-social manner. As a result, rewards and punishments reflect actual changes in the principal's belief over the action taken by the agent. 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 punishment are conditioned on noise and do not reflect changes in posterior beliefs. We show that the no-punishment-without-guilt restriction has significant consequences on the way positive and negative externalities are internalized. Negative externalities induce mixed strategy equilibria in which externalities are partially but never fully internalized. Positive externalities induce multiple pure strategy equilibria, under which externalities are either fully internalized or not at all.[5]

Finally, we outline efficiency properties of different modes of informal justice. We show that outcome-based justice leads to efficient decision-making conditional on transfers, and that it guarantees a minimum share of the efficient surplus. However, it remains bounded away from first-best efficiency even as side information becomes arbitrarily precise. In contrast, intent-based justice admits a most pro-social equilibrium that approaches efficiency

[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.

as information becomes arbitrarily precise. However, even as information becomes precise, 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 closely related to Fudenberg and Tirole (1990) who study the effect of renegotiation in a canonical principal-agent problem. Because of moral hazard, the optimal ex ante contract with commitment must expose the agent to some of the risk. Fudenberg and Tirole (1990) drop the assumption that the principal has commitment power, and allow for renegotiation at an interim stage occurring *after* the agent's action is taken, but *before* outcomes are realized. At this interim stage, it is Pareto improving for the principal to offer the agent insurance. As a result, there cannot exist an equilibrium in which the agent puts effort with probability one: renegotiation would lead to perfect insurance, thereby removing all incentives for effort. Although Fudenberg and Tirole (1990) do not assume that the principal has social preferences, interim contracts offered by the principal under renegotiation constraints can be interpreted as maximizing social preferences taking the form of an appropriately weighted average of the principal's profits and the agent's expected utility. Our work can be viewed as embracing this behavioral interpretation of Fudenberg and Tirole (1990), explicitly using a rich class of social preferences that reflect behavioral evidence accumulated in recent years. By using actual social preferences as the motive for ex post transfers, our model makes novel predictions on the way different externalities are internalized, and avoids fragility to the timing of renegotiation from which Fudenberg and Tirole (1990) suffers. In their model, when outcomes are known at the time of renegotiation, there is no scope for insurance, and renegotiation has no bite on the ex ante optimal contract. In our case, social preferences insure that there is a motive for redistribution regardless of the information available.

Even though we consider a one-shot game, our work shares a common motivation with the relational contracting literature. In the absence of formal ex ante contracts, the relational contracts approach places plausible restrictions on rewards and punishments available for incentive provision by requiring them to be subgame perfect in an appropriate repeated game (see for instance Green and Porter (1984), Bull (1987), MacLeod and Malcomson (1989), Baker et al. (1994, 2002), Levin (2003)).[6] Our approach also places restrictions on punishments and rewards by imposing that they be optimal from the perspective of a decision-maker with social preferences. We show that this yields novel yet plausible restrictions on patterns of rewards and punishments, with non-obvious implications for the way different types of externalities are internalized.[7]

Because the principal's social preferences play a central role in our framework, this paper contributes to a growing literature at the intersection of contract theory and behavioral economics. 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] Our paper embraces the idea that social preferences in and of themselves define informal contracting heuristics that may be quite effective at sustaining efficient play. This echoes Andreoni and Samuelson (2006) who show that social preferences successfully support cooperation in a finitely repeated prisoners' dilemma, as well as Carmichael and MacLeod (2003) and MacLeod (2007) who argue that social preferences may be rationalized as encoding for efficient equilibrium play in a Nash demand game with sunk costs.

[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 (2010) 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 the full history of past observables. In their framework, emotions are purely informational states that do not affect preferences.

Note that in contrast to repeated games models of informal contracting, in our setting, equilibrium multiplicity is not a precondition for informal incentives to arise, but rather an outcome of interest reflecting the mechanics of informal justice in particular environments.

[8] For instance, several studies show that generous fixed-wages induce fair-minded workers to increase non-enforceable effort provision (Fehr et al., 1993, 1997, Fehr and Gächter, 1998, Falk and Gaechter, 2002, Hannan et al., 2002, Charness, 2004, Charness et al., 2004). Other work suggests that explicit incentive contracts can reduce effort by crowding out pro-social motivation (Gneezy and Rustichini, 2000, Fehr and Falk, 2002, Benabou and Tirole, 2003, Ellingsen and Johannesson, 2008, Falk and Kosfeld, 2006).

[9] For example, it has been shown that in the presence of fair-minded agents non-enforceable bonus payments are a powerful motivator (Fehr et al., 2007), linear payment rules may be optimal (Englmaier and Wambach, 2010), and rigid contracts that fix the terms ex ante may limit counterproductive behavior (Hart and Moore, 2008, Fehr et al., 2011).

The paper is structured as follows. Section 2 introduces the model and illustrates our main qualitative points with a detailed example. Section 3 describes our general framework, while Sections 4 and 5 characterize patterns of informal justice as a function of the principal's social preferences. Section 6 discusses our modeling choices as well as challenges for further research. Online Appendix A provides several extensions. Proofs are contained in Online Appendix B.

## 2 An Example

Because we emphasize ex ante (or procedural) fairness as a determinant of the principal's social preferences, ours is necessarily a non-expected utility model. For this reason, the mechanics of equilibrium are a little unusual and we delineate them using a particularly simple example. This example is sufficiently detailed to capture the main novel predictions of our approach, making it a suitable summary of the paper. We generalize the analysis in subsequent sections.

**Motivation.** Two assumptions are at the core of our model. The first is that in many environments, the relevant principal may not be able to, or may choose not to, commit to an ex ante contingent contract. The second is that in the absence of an ex ante contract, the principal will decide on ex post transfers on the basis of her social preferences. To motivate these assumptions, we describe four economically relevant settings that our model hopes to capture.

Note that even when formal contracts are available in principle, they also carry high fixed costs. Hence contracting or not should be regarded as an endogenous decision. When studying such settings, a model like ours is needed to correctly specify what would happen under the counterfactual in which contracts are not used.

Parental discipline is an important and natural example of informal contracting. Parents do not typically commit to formal contracts specifying the way they will respond to a child's behavior, but they frequently make disciplining decisions based on their intuitive moral judgement. Intuitively, considerations of both ex post and ex ante fairness play a significant role. If a child breaks the toy of a sibling, some transfer may be implemented regardless of intent – this corresponds to ex post fairness. In addition, the parents' response may also depend on their perception of the child's intentions, which is captured by ex ante fairness. Keeping circumstantial evidence constant, a child expected to behave may get away scot-free, whereas a child expected to misbehave may get additional punishment.

Formal commitment to ex ante contracts is also unlikely to be used at the lower echelons of organizations. Since writing and implementing formal contracts carries high fixed costs, they are only economically viable if the scale of the incentive problems they seek to correct is sufficiently large. Imagine a manager responsible for a team of sales-people. A particular sales-person may exert negative or positive externalities on his team by poaching customers, or by providing expertise in dealing with clients' questions and support issues. This is too small an incentive problem to justify the legal costs required to contract. However, the manager can still implement transfers across sales-people through her allocation of tasks, her assignment of resources, or her decision to sponsor promotions. In the absence of contracts, it is plausible that she will do so according to some sense of fairness, captured here by social preferences. Indeed, 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. For instance, if the manager believes a sales-person increased his performance by poaching the customers of an other, fairness concerns may cause her to direct new leads to the slighted sales-person.

Even at the higher echelon of organizations, where the scale of incentive problems could in principle justify legal costs, the use of ex ante contracts may be limited by bounded-rationality constraints. Consider for instance the problem of a senior executive arbitrating between two divisions of a company. At this scale, providing correct 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.[10] However, the senior executive may also be unaware of other incentive problems that can arise. For instance, one division may have the opportunity to support the development of the other in a new geographic market in which it is already implanted. If those circumstances are not anticipated at the time a formal contract is written, then the senior executive is left to decide how they should affect ex post promotion decisions on the basis of her ex post social preferences.

Finally, it may well be that the relevant principal in charge of implementing transfers is not actually present at an ex ante stage before actions are taken. One such example is lay juries arbitrating on a lawsuit. While jury members are given instructions regarding legal procedure, these instructions offer the jury significant freedom in the assignment of guilt and damages.[11] Hence, at this ex post stage, the decisions of the jury necessarily express the moral judgement of its members. In fact, jury instructions reinforce this view. They clarify that in civil cases, 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. This coincides with the no-punishment-without-guilt property that we emphasize throughout the paper, as well as with classic models of jury behavior (Feddersen and Pesendorfer, 1998). This also contrasts with existing formal and informal contracting models (Holmström (1979), Harris and Raviv (1979), Green and Porter (1984), Bull (1987)): *juries are not instructed to condemn defendants whom they believe to be innocent* (in spite of circumstantial evidence). Furthermore, proposed jury instructions available on the American Bar Association's website (Velez v. Novartis, 2010) can arguably be interpreted as expressing social preferences placing weight on both ex post and ex ante fairness. In this gender discrimination case, jury instructions distinguish damages to be awarded because of disparate impact of policies on men and women regardless of intent (this

[10] 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.

[11] This is illustrated by the rising concern that juries have excessive leeway in specifying damages. See for instance the recent (challenged) award of 23.6 Billion dollars in damages by a Florida jury (Robles, 2014).

is ex post fairness), versus damages awarded because of disparate treatment, which captures intent, and authorizes punitive damages (this is ex ante fairness).

## 2.1 Setup

We now provide a formal description of our framework, and illustrate it using the example of a mid-level manager responsible for a group of sales-people.

**Players.** We consider a subjective performance evaluation problem in which a principal (the manager) can privately assess the contribution of an agent (a sales-person) to the welfare of a broader group (the sales-team). We model this broader group as a representative passive player $P$ with no strategic decisions to make, but whose welfare enters the social preferences of the principal. The principal cares about both efficiency and fairness, which she can trade-off using costly ex post transfers between players. We assume that outcome evaluations are subjective and cannot be verified by outside parties (such as courts). The principal cannot commit to an ex ante contract, and the transfers she chooses must be optimal from the perspective of her ex post social preferences.

**Actions and payoffs.** Player $A$ takes a private action $a \in \{C, D\}$ which controls the distribution of a public stochastic state $z \in Z = \{-1, 0, 1\}$:

| $z$ | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $\text{prob}(z\|C)$ | $\nu$ | $1-2\nu$ | $\nu$ |
| $\text{prob}(z\|D)$ | $1$ | $0$ | $0$ |

where $\nu \in (0, 1/4]$. Player $A$ and $P$'s respective payoffs take the following form

$$u_A = -[z+\gamma] \quad \text{and} \quad u_P = 2[z+\gamma],$$

where $\gamma \in \{0, 1\}$ is a fixed parameter used to perform comparative statics across environments. Note that parameter $\gamma$ merely shifts the payoffs of each player by a constant.

Payoffs $u_A$ and $u_P$ are interpreted as departures from either player's outside option, i.e. from the counterfactual payoff they would obtain if the other player was absent. When $z + \gamma > 0$, player $A$ has a positive externality on player $P$, and conversely, when $z + \gamma < 0$, player $A$ has a negative externality on $P$. Note that $C$ is a pro-social action that increases the likelihood with which states favorable to player $P$ arise. Because of noise term $\nu > 0$, this is an imperfect monitoring framework and outcome $z = -1$ can happen even if player $A$ takes action $C$. Payoffs are designed to satisfy the following properties:

- for all values $\gamma \in \{0, 1\}$, $\mathbb{E}(u_A + u_P|C) > \mathbb{E}(u_A + u_P|D)$:
  action $C$ maximizes the sum of payoffs;

- for all values $\gamma \in \{0, 1\}$, $\mathbb{E}(u_A|C) < \mathbb{E}(u_A|D)$ and $\mathbb{E}(u_P|C) > \mathbb{E}(u_P|D)$:
  there is a conflict of interest — in the absence of additional incentives, player $A$ would take action $D$;

- for all values $\gamma \in \{0, 1\}$ there exists a "status quo" action, i.e. an action $a \in \{C, D\}$ such that $\mathbb{E}(u_A|a) = \mathbb{E}(u_P|a) = 0$, corresponding to the expected payoffs each player would obtain if the other party was absent.

Negative externality environments correspond to $\gamma = 0$. The status quo action is the efficient action $C$. Action $D$ benefits player $A$ and imposes a negative expected externality on player $P$. For instance, in our managerial example, $C$ would correspond to neutral behavior, while $D$ would correspond to poaching customers that other sales-people have been cultivating.

Positive externality environments correspond to $\gamma = 1$. The status quo action is the inefficient action $D$. Action $C$ is costly to player $A$ but yields a positive expected externality for player $P$. In our example, $D$ would now be neutral behavior, while $C$ would correspond to sharing expertise on a particularly difficult case.

Note that player $A$ does not have social preferences and maximizes his expected payoffs. To keep our model simple, only the principal has social preferences, which we describe now.[12]

**The Principal's problem.** The realized state $z$ (a sufficient statistic for payoffs $u_A, u_P$) is observed by a principal who can implement a transfer $T_z \in [-T_{\max}, T_{\max}]$ between the two players. This transfer may, for example, be implemented by reassigning responsibilities, tasks, or promotions. 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 $\lambda \in (0, \frac{1}{2})$, accruing to the transferring party.[13] 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

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

player $A$ and $P$'s payoffs net of transfers, and by $u^T \equiv (u_A^T, u_P^T)$ the corresponding profile of payoffs.[14] We denote by $\pi \in \Delta(\{C, D\})$ mixed strategies of player $A$. We 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

$$V(\pi, T) = \sum_{a \in \{C,D\}} \pi(a) \left( \mathbb{E}[u_A^T + u_P^T | a] - \alpha \left| \mathbb{E}[u_A^T - u_P^T | a] \right| \right). \tag{1}$$

These preferences can be viewed as a variant of Fehr and Schmidt (1999)'s model of social preferences applied to ex ante payoffs. Whenever $\alpha = 0$ the principal simply values ex ante utilitarian efficiency. When $\alpha > 0$, the principal values ex ante utilitarian efficiency and dislikes ex ante inequity. This model of preferences captures preferences for ex ante

[12] Endowing the agent with social preferences would not change our qualitative predictions: arguably, the agent's payoffs could already include his social preferences. In addition, it is plausible that an a priori neutral principal should be more sensitive to fairness concerns than directly interested parties.

[13] We think of the cost of transfers as arising from specificities in the resources being transferred from one party to the other. To focus on the novel aspects of our model, we do not endogenize the cost of transfers. A similar reduced-form assumption is frequent in the literature on optimal regulation. See for instance Laffont and Tirole (1986) or Laffont and Martimort (2009).

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

(or procedural) fairness: what matters to the principal is equality of opportunities.[15] 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 — this is not an expected utility model and the principal is not consequentialist: given a realization of payoffs, her preferences over transfers ex post depend on the counterfactual distribution of potential payoffs.[16] Such preferences for ex ante (or procedural) fairness are consistent with experimental evidence from Bolton et al. (2005), Charness and Levine (2007) or Krawczyk and Le Lec (2010).

Furthermore, the principal's preferences differentiate strategic uncertainty arising from mixed strategies $\pi$ over actions, and non-strategic uncertainty arising from lotteries $f(z|a)$ over outcomes, given actions. Under our interpretation of mixed strategies as pure strategies played by a population, this essentially assumes that the principal evaluates the fairness of individual relationships between a given player $A$ and a given player $P$, rather than fairness at the population level. From the perspective of the principal, a situation in which player $A$ takes an action that benefits him with 50% chance and benefits the other player with 50% chance is very different from a situation in which 50% of player $A$s in the population take the action that benefits them, while 50% of player $A$s take the action that benefits the other player. This insures that even when there is mixing at the population level, the principal cares about the action taken by a given player A. Appendix A discusses the implications of alternative modeling decisions.

From an empirical perspective, this modeling choice is consistent with recent findings by Bohnet and Zeckhauser (2004) and Bohnet et al. (2008) showing that people treat lotteries over social outcomes differently when the uncertainty is determined by strategic play by others, and when it is determined by objective uncertainty, even if the experimental design guarantees that the two lotteries have the same distribution over social consequences.

[15] To focus on the main novel points of the paper, the example here assumes that the principal cares solely about ex ante fairness. Subsequent sections allow for a broader set of social preferences that also value ex post fairness.

[16] See Fudenberg and Levine (2012) for a recent discussion.

## 2.2 Ex ante contracting

As a benchmark, we describe the optimal ex ante contract when the principal has commitment power. For any action $a \in \{C, D\}$, let us denote by $\neg a$ the alternative action. The optimal ex ante transfer scheme $T : z \in Z \mapsto T_z \in [-T_{\max}, T_{\max}]$ solves

$$\max_{a \in \{C,D\}, T} V(a, T) \;\Big|\; \mathbb{E}\left[u_A^T | a\right] \geq \mathbb{E}\left[u_A^T | \neg a\right]. \tag{P1}$$

**Fact 1** *Any optimal contract $(a^*, T^*)$ solving P1 satisfies the following properties:*

*(i) for every $\lambda \in (0, 1/2)$, $\alpha \in (0, +\infty)$, and $\gamma \in \{0, 1\}$ the optimal contract implements action $a^* = C$;*

*(ii) if $\gamma = 0$ the optimal transfer scheme satisfies $T^*(z = -1) > 0$.*

In words, point $(i)$ observes that given our specific parameter restrictions, the optimal ex ante contract implements pro-social action $C$ regardless of social preferences $\alpha$, and regardless of whether parameter $\gamma$ defines an environment with positive or negative externalities. Point $(ii)$ states that if the environment is one with negative externalities ($\gamma = 0$), the principal punishes player $A$ whenever state $z = -1$ occurs. Note that this occurs, although in equilibrium there is common knowledge that player $A$ took action $C$. In our managerial example, when customer poaching is a concern, the optimal ex ante contract would require sales-person $A$ to be penalized following performance that is unusually high compared to that of others. Punishment must occur on the equilibrium path, even though the principal is certain that the low performance of others is not due to customer poaching. This property, which we name *punishment-without-guilt*, holds in most models of formal and informal contracting, including Green and Porter (1984)'s model of oligopoly pricing. Firms trigger price wars following poor demand realizations even though there is common knowledge in equilibrium that all firms cooperated. Punishment-without-guilt no longer holds in the model of informal justice we analyze below.

## 2.3 Informal Justice

We drop the assumption that the principal is able to commit to contracts ex ante and instead we study the transfer scheme that she implements when maximizing her ex post social preferences.

**Informal justice in equilibrium.** Because the principal cares about ex ante fairness, the ex post transfer scheme $T$ she implements depends on player $A$'s expected behavior $\pi$. For instance, when the principal's prior is that player $A$ picked selfish action $D$, she will be enclined to implement positive transfers from $A$ to $P$. Since behavior $\pi$ itself depends on the incentives provided by transfer scheme $T$, we now have a game in which transfers $T$ determine behavior $\pi$, and behavior $\pi$ determines transfers $T$. As was previously noted, the principal is not consequentialist and we use Bayes Nash equilibrium as our solution concept.

**Definition 1 (solution concept)** *For any distribution $\pi \in \Delta(\{C, D\})$, a pair $(\pi, T)$ 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 require that transfer scheme $T^\pi$ be the limit of optimal transfer schemes $T^{\widehat{\pi}}$ for full-support distributions $\widehat{\pi} \in \Delta(\{C, D\})$ approaching $\pi$.*[17]

One immediate difference with optimal ex ante contracts is that under informal justice, incentives will be provided only if the deadweight loss of transfers $\lambda$ is not too large compared to inequity aversion parameter $\alpha$.

**Fact 2** *Whenever $\alpha < \frac{\lambda}{2+\lambda}$, transfer scheme $T$ is identically equal to 0 and the unique equilibrium behavior is $D$.*

Whenever $\alpha < \frac{\lambda}{2+\lambda}$, even though the principal would have liked to commit to non-zero transfers ex ante (Fact 1), ex post — taking actions as given — the deadweight loss from transfers overwhelms the fairness benefits they generate. We assume that $\alpha > \frac{\lambda}{2+\lambda}$ for the rest of the paper.

[17] Lemma 2 establishes the existence of such limit transfers.

We now characterize the nature of incentives provided under informal justice. We first establish that transfers must satisfy *no-punishment-without-guilt.* This restriction on incentives will have implications for the way different externalities are internalized.

Given prior $\pi \in \Delta(\{C, D\})$, let us denote by $\pi(a|z)$ the principal's posterior belief that action $a$ was taken, given $z$.

**Fact 3 (no punishment without guilt)** $\forall z \in Z$,

$$T_z^\pi > 0 \Rightarrow \pi(D|z) > \frac{\lambda}{\alpha(2+\lambda)} \qquad \textit{(no punishment without guilt),}$$

$$T_z^\pi < 0 \Rightarrow \pi(C|z) > \frac{\lambda}{\alpha(2+\lambda)} \qquad \textit{(no reward without virtue).}$$

In words, the principal will only punish the agent when she holds a sufficiently high posterior belief that the agent took action $D$ — there is no punishment without guilt. Inversely, the principal only rewards the agent when she holds a sufficiently high posterior belief that the agent took action $C$ — there is no reward without virtue.

This restriction on patterns of punishments and rewards has different implications depending on the nature of externalities.

**Negative externalities.** Consider the case of negative externalities ($\gamma = 0$). Action $C$ yields status quo expected payoffs (i.e. $\mathbb{E}[u_A|C] = \mathbb{E}[u_P|C] = 0$). Action $D$ imposes a negative externality on $P$ (i.e. $\mathbb{E}[u_P|D] < 0 < \mathbb{E}[u_A|D]$). The following result holds.

**Fact 4 (homogeneous response to negative externalities)** *There exists a unique equilibrium. Player A's behavior $\pi$ is characterized by*

$$\pi(C) = \frac{\alpha(2+\lambda) - \lambda}{2\nu\lambda + \alpha(2+\lambda) - \lambda}.$$

This result suggests that organizations may be homogeneous in their informal response to negative externalities: negative externalities are systematically, but never fully, internalized.

Clarifying why $\pi(C) = 1$ and $\pi(D) = 1$ are not consistent with an equilibrium is instructive. Imagine that $\pi(C) = 1$. No-punishment-without-guilt implies that the principal will not implement any punishment. Furthermore, since this is a negative externality environment, there is no expected inequality conditional on action $C$. Hence, the principal is also unwilling to reward the agent. As a result, the principal provides no incentives, and $C$ cannot be an optimal response for the agent.

Imagine now that $\pi(D) = 1$. Since there is large inequality in expectation, the principal will be willing to punish the agent. Optimal transfers satisfy $T_{-1} = \frac{3}{2+\lambda}$ and $T_0, T_1 \leq 0$ with $(1-\nu)T_0 + \nu T_1 = -\nu T_0$. Therefore, player $A$'s expected payoffs from taking actions $C$ and $D$ are $\mathbb{E}[u_A^T|C] = -\nu\frac{3\lambda}{2+\lambda} > \mathbb{E}[u_A^T|D] = -\frac{1+2\lambda}{2+\lambda}$. Hence, it is optimal for player $A$ to take action $C$.

Although negative externalities are never fully internalized ($\pi(C) = 1$ is not an equilibrium) they are always partially internalized ($\pi(D) = 1$ is not an equilibrium either). The likelihood of pro-social behavior $\pi(C) = \frac{\alpha(2+\lambda)-\lambda}{2\nu\lambda+\alpha(2+\lambda)-\lambda} \in (0, 1)$ is determined so that: $(i)$ the principal is willing to punish player $A$ when $z = -1$ occurs; $(ii)$ the corresponding transfers keep player $A$ indifferent between taking actions $C$ and $D$. Note that the probability $\pi(C)$ of taking pro-social action $C$ is strictly decreasing in noise parameter $\nu$. Since the principal is reluctant to impose transfers when player $A$ may have taken action $C$, the possibility of excuses (higher values of $\nu$) limits the probability with which action $C$ can be sustained.

In our management example this result suggests that a manager who values ex ante fairness will never be able to fully prevent her sales-people from poaching clients from each other. No-punishment-without-guilt implies that the manager is willing to take sanctions only when she has a sufficiently strong posterior belief that a sales-person poached customers from other team-members. Hence, some residual misbehavior must subsist for the principal to be willing to enforce transfers. Since the manager's willingness to punish depends on her *posterior* beliefs, the minimum rate of customer-poaching needed for the manager to act depends on her ability to infer behavior from outcomes. Keeping behavior fixed, noisier signals make punishment feel less fair to the manager since they apply more frequently to well-behaving sales-people.

**Positive externalities.** Consider now the positive externality environment in which $\gamma = 1$. This environment differs from the negative externality environment only by shifts in payoffs that do not affect the comparative benefits of each action: $C$ maximizes the sum of payoffs, and in the absence of transfers, player $A$ would take action $D$. The key difference with the negative externality setting is that the status quo action — under which each player gets her outside option in expectation — is now the inefficient action $D$. Pro-social action $C$ now yields a strict increase in expected payoffs for player $P$ and a loss in expected payoffs for player $A$.

**Fact 5 (heterogeneous response to positive externalities)** *(i) There exists a laissez-faire equilibrium such that $\pi(D) = 1$ and $\forall z$, $T_z = 0$.*

*(ii) There also exists a pro-social equilibrium such that $\pi(C) = 1$,*

$$T(z = -1) = 0 \quad , \quad T(z = 0) = -\frac{3}{2+\lambda} \quad \textit{and} \quad T(z = 1) = -\frac{6}{2+\lambda}.$$

In contrast to the case of negative externalities, there now exist multiple equilibria. In particular there exists an equilibrium in which externalities are not internalized at all, and one in which externalities are fully internalized. In both equilibria excuses (i.e. the magnitude of noise term $\nu$) no longer affect the probability with which the efficient action $C$ can be sustained in equilibrium.

This result suggests that organizations may be heterogeneous in their informal response to positive externalities, yielding different "firm cultures". In the context of our example, some sales-teams may adopt a laissez-faire culture in which sales-people focus on their own accounts and do exert effort supporting one another. Consistent with this, the manager may not find it fair to redistribute the benefits of private successes. In contrast, other sales-teams may support one another with expertise. Given a particular success, the manager finds it fair to assign credit to all team-members likely to have contributed, which incentivizes cooperative behavior in the first place.

The reasoning underlying Fact 5 is straightforward. If $\pi(D) = 1$, then 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 $C$. If instead $\pi(C) = 1$, then there is inequality in expected payoffs and the principal's optimal transfer scheme corrects for this inequality by rewarding player $A$ in states $z = 0$ and $z = 1$. Given transfers, player $A$'s expected payoffs conditional on actions $C$ and $D$ are $\mathbb{E}\left[u_A^T|C\right] = \frac{1-\lambda}{2+\lambda} > \mathbb{E}\left[u_A^T|D\right] = 0$. As a result player $A$ takes action $C$.

As illustrated by this simple example, our model of informal incentives makes novel predictions about patterns of punishments and rewards, and the behavior they incentivize. Because the principal cares about ex ante fairness, she does not punish the agent unless there is sufficiently high posterior probability that the agent misbehaved. This contrasts with existing models of formal and informal contracting in which punishment occurs even though there is common knowledge in equilibrium that the agent behaved well. In negative externality environments, this implies that externalities can not be fully internalized, but are always partially internalized. In positive externality environments, externalities may either be fully internalized or not at all.

We believe that these predictions are plausible refinements of existing models of punishments and rewards, and that they could be tested both in the lab and in the field. For instance our theory suggests that organizations will punish negative externalities in fairly uniform and systematic ways, but may exhibit significant heterogeneity in whether or not positive externalities are rewarded. This prediction speaks directly to the recent literature on persistent productivity differences across seemingly similar enterprises (Gibbons et al., 2010, Gibbons and Henderson, 2012). With such potential applications in mind, we now explore how our predictions extend beyond the simple example of this section.

# 3 General Framework

Our general framework extends that of Section 2 in several ways: (*i*) we allow for general payoff distributions; (*ii*) in addition to realized payoffs, the principal may observe payoff-irrelevant signals informative of the action taken by player $A$; (*iii*) we consider a broader class of social preferences for the principal, which value both ex ante and ex post fairness. These extensions allow us to better understand general qualitative properties of informal justice, and let us ask richer questions. Do previous results generalize? How do different social preferences affect informal justice? How does informal justice exploit information? What are the efficiency properties of informal justice?

## 3.1 Setup

**Payoffs and information.** As in Section 2, player $A$ takes an action $a \in \{C, D\}$ which affects her payoff and that of passive player $P$. The principal can implement transfers $T$ between the two players which come at a cost $\lambda|T|$ to the transferring party, with $\lambda > 0$.[18] The principal can now observe a richer set of consequences $z = (u, x) \in Z = U \times X$, where $Z \subset \mathbb{R}^k$ is compact, $u \equiv (u_A, u_P) \in U \subset \mathbb{R}^2$ are payoff realizations, and $x \in X$ is a payoff-irrelevant signal informative of player $A$'s behavior. Let $\mathcal{L}$ denote the restriction of the Lebesgue measure to $Z$ and let $f(z|a)$ denote the density of observable outcomes $z$ given action $a$ against $\mathcal{L}$. We assume that densities $f(z|a)$ are bounded below by some value $\underline{h} > 0$, i.e. there is full support. We refer to $(Z, f)$ as the *environment*, and its restriction to payoffs $(U, f_{|U})$ as the *payoff environment*. We maintain a common prior assumption throughout the paper.

A strategy profile in this game is given by a pair $(\pi, T)$, where $\pi \in \Delta(\{C, D\})$ is a distribution over actions by player $A$, and mapping $T : Z \to [-T_{\max}, T_{\max}]$ is a transfer function chosen by the principal.[19] We continue to interpret mixed strategies as pure strategies played

[18] Appendix A shows that our analysis is robust to perturbations to social preferences and to the cost of transfers.

[19] Recall that $T_{\max} \geq \max_z |u_A - u_P|$.

by a population of players. We maintain the requirement that $\frac{\lambda}{2+\lambda} < \alpha$, i.e. the principal has sufficiently large preference for fairness. In addition, we make the following assumptions.

**Assumption 1** *The log-likelihood ratio* $\log\left(\frac{f(z|a=D)}{f(z|a=C)}\right) \in \mathbb{R}$*, viewed as a random variable under measure* $\mathcal{L}$*, has a density, and a bounded convex support.*[20]

This assumption is made for convenience: it helps ensure that optimal transfer policies are unique and well behaved. To state continuity properties of transfer schemes, the space of integrable functions from $Z$ to $\mathbb{R}$ is endowed with the $L_1$ norm $||\cdot||_1$: for any integrable function $g : Z \to \mathbb{R}$, $||g||_1 = \int_{z\in Z} |g(z)|\,\mathrm{d}z$.

Finally we assume that payoffs are centered in the following way.

**Assumption 2** *We assume that the following properties hold*

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

*Centering* extends our interpretation of payoffs $u_i$ as departures from a status quo or outside option. *Conflict* restricts attention to cases where external incentives are needed to internalize externalities. 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 doesn't hold. Without further loss of generality, we can assume that

$$\mathbb{E}[u_A|C] \leq 0 \leq \mathbb{E}[u_A|D] \quad \text{and} \quad \mathbb{E}[u_P|C] \geq 0 \geq \mathbb{E}[u_P|D].$$

This labels $C$ as a pro-social action creating value for player $P$ at the expense of player $A$.

**The Principal's problem.** The principal's social preferences are extended to allow for preferences over both ex ante and ex post fairness. Recalling that $u^T = (u^T_A, u^T_P)$ denotes

[20] To clarify, we view $z \in Z$ as a random variable under measure $\mathcal{L}$, and $y(z) \equiv \log\left(\frac{f(z|a=D)}{f(z|a=C)}\right)$ as an induced random variable taking values in $\mathbb{R}$, and associated with the measure $\mathcal{L} \circ y^{-1}$.

payoffs net of transfers, social preferences are described by utility function

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

where $\Phi(u) \equiv u_A + u_P - \alpha|u_A - u_P|$, and $\delta \in [0,1]$. For $\delta = 1$, the principal is a standard expected utility maximizer who simply dislikes ex post inequality in payoffs. For $\delta = 0$, we are back to the case of Section 2: the principal cares only about ex ante fairness.

This model extends the preferences described in (1) by placing weight on both ex post (or allocative) fairness $\mathbb{E}[\Phi(u^T)|a]$ and on ex ante (or procedural) fairness $\Phi(\mathbb{E}[u^T|a])$. This class of preferences is motivated by experimental evidence: Bolton et al. (2005), Charness and Levine (2007), Cushman et al. (2009), Schächtele et al. (2011) as well as Krawczyk and Le Lec (2010) all show that decision makers care about the fairness of both ex post realized payoffs and ex ante prospects. When $\pi$ 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 (2008, 2010). As in Section 2, the principal is non-consequentialist and we use Bayes Nash equilibrium as our solution concept (Definition 1 continues to apply).

## 3.2 Forms of Contracting

**The Ex Ante Optimal Benchmark.** As a preliminary to our characterization of transfers under informal justice, we briefly highlight benchmark properties of ex ante optimal contracts (extending Fact 1).[21] As in Section 2, the optimal ex ante contract $T^{\text{ex ante}}$ under the principal's preferences, solves

$$\max_{\pi\in\Delta(\{C,D\}),T} V(\pi,T) \;\Big|\; \forall a \in \{C,D\},\ \pi(a) > 0 \Rightarrow \mathbb{E}[u_A^T|a] \geq \mathbb{E}[u_A^T|\neg a]. \tag{P1$'$}$$

Contract $T^{\text{ex ante}}$ satisfies the following intuitive properties.

(i) *Pure and unique behavior.* Optimal ex ante contracts implement a generically

[21] These properties are straightforward and proofs are omitted for concision.

unique pure action $a \in \{C, D\}$.[22]

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

(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 $C$.

(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 $u_A^T < u_P^T$.

**Modes of Informal Justice.** We contrast the ex ante optimal contracting framework with the implicit incentive schemes provided by ex post fairness-driven transfers. It turns out that there are two very distinct modes of informal justice depending on the weight that the principal places on ex ante versus ex post fairness. These two modes of informal justice, which we detail in Sections 4 and 5, are best described as *outcome-based* and *intent-based justice.*

# 4 Outcome-Based Justice

We begin by assuming that $\delta > \frac{\lambda}{\alpha(2+\lambda)}$. In this case, the weight $\delta$ placed on ex post fairness is sufficiently large that the principal chooses transfers that equalize payoffs for each outcome realization.

**Proposition 1** *For all behavior distribution $\pi \in \Delta(\{C, D\})$, the optimal transfer scheme is*

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

[22] Genericity is taken with respect to payoff mappings $(u_A, u_P)$ under the $L_1$ norm.

*Generically with respect to payoff functions $u : Z \to \mathbb{R}^2$, there exists a unique equilibrium and it is in pure strategies, i.e. such that $\pi(C) \in \{0, 1\}$.*

This mode of informal justice essentially follows talionic law: "an eye for an eye, a tooth for a tooth." We refer to this mode of informal justice as being *outcome based* and denote by $T^O$ the corresponding transfer scheme. We emphasize two corollaries.

**Corollary 1** *(i) Transfer scheme $T^O$ depends on realized payoffs $u_z$, and not on signal $x$.*[23]

*(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), 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. 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 the equal ex ante optimal payoffs by an amount $\frac{\lambda}{2+\lambda} \min_{a \in \{C,D\}} \mathbb{E}[|u_A - u_P||a] - |\mathbb{E}[u_A - u_P|a]|$ , even when information

[23] Recall that observables $z = (u, x)$ take the form of a payoff realization $u$ and a signal $x$ that is payoff-irrelevant but informative about the agent's behavior.

becomes arbitrarily good.[24]

# 5 Intent-Based Justice

We now consider the case where the principal puts a sufficiently high weight on ex ante fairness. Specifically, we assume that $\delta < \frac{\lambda}{\alpha(2+\lambda)}$. We show that rewards and punishments follow qualitatively different patterns.

## 5.1 Rewards and Punishments

Take a distribution of behavior $\pi \in \Delta(\{C, D\})$ as given. We first characterize the transfer scheme $T^\pi$ solving $\max_T V(\pi, T)$. Let $f_\pi(z) \equiv \sum_{a \in \{C,D\}} \pi(a) f(z|a)$ 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

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

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

[24] See Appendix A for a detailed exploration of the efficiency properties of both outcome-based and intent-based justice.

functions $T \in [-T_{\max}, T_{\max}]^Z$ is convex and that $V(\pi, T)$ is concave in $T$.

**Lemma 1** *For any $\pi \in \Delta(\{C, D\})$, there exists an optimal transfer policy $T^\pi$. In addition, any optimal transfer policy $T^\pi$ is such that*

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

This implies that optimal transfers $T^\pi$ do not reverse existing payoff asymmetries: action $C$ continues to generate inequality in favor of player $P$, while action $D$ generates inequality in favor of player $A$. This lets us sign the term corresponding to ex ante fairness in expression (3), and simplify the principal's optimization problem. For any pair of multipliers $\mu = (\mu_C, \mu_D) \geq 0$, define the Lagrangian

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

Optimal transfer schemes can be characterized as follows.

**Lemma 2 (characterization)** *For every distribution $\pi$ in the interior of $\Delta(\{C, D\})$ and every $\delta \in [0,1]$, there exists a unique optimal transfer scheme $T^\pi$. 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 $T_z^\pi$ depends only on realized inequality $\Delta u_z$ and on the posterior likelihood ratio $\frac{\pi(D|z)}{\pi(C|z)} = \frac{\pi(D)f(z|D)}{\pi(C)f(z|C)}$. Through this likelihood ratio, transfers under informal justice can depend on informative payoff-irrelevant signals $x$ — a feature shared by optimal ex ante contracts. However, unlike ex ante optimal contracts, transfers under informal justice also depend on prior beliefs $\pi$ over behavior. In particular, the no-punishment-without-guilt property outlined in Fact 3 extends as follows.

**Proposition 2 (no punishment without guilt)** *For any interior $\pi \in \Delta(\{C, D\})$, there exists thresholds $-1 \leq h_-^{\max} < h_-^{\Delta} < h_+^{\Delta} < h_+^{\max} \leq 1$ such that transfer policy $T^\pi$ takes the form*

$$T_z^\pi = \begin{cases} 0 & \textit{if} \;\; \pi(D|z) - \pi(C|z) \in (h_-^{\Delta}, h_+^{\Delta}) \\ -T_{\max} & \textit{if} \;\; \pi(D|z) - \pi(C|z) < h_-^{\max} \\ T_{\max} & \textit{if} \;\; \pi(D|z) - \pi(C|z) > h_+^{\max} \\ (\Delta u_z)^+/(2+\lambda) & \textit{if} \;\; \pi(D|z) - \pi(C|z) \in (h_+^{\Delta}, h_+^{\max}) \\ -(\Delta u_z)^-/(2+\lambda) & \textit{if} \;\; \pi(D|z) - \pi(C|z) \in (h_-^{\max}, h_-^{\Delta}).\text{[25]} \end{cases}$$

*There is no punishment-without-guilt: $T_z^\pi > 0 \Rightarrow \pi(D|z) > 0$; and no reward-without-virtue: $T_z^\pi < 0 \Rightarrow \pi(C|z) > 0$.*

Transfers take a threshold form dependent on the precision $\pi(D|z) - \pi(C|z)$ with which the behavior of player $A$ can be inferred ex post. This is why we refer to this mode of informal justice as being *intent based*. We emphasize again that the no-punishment-without-guilt property contrasts sharply with existing models of informal contracting. Under intent-based justice, when the posterior probability that player $A$ took action $C$ is sufficiently high, the principal does not find it acceptable to punish him with costly transfers, even following poor outcome realizations.

Before moving to equilibrium analysis, it interesting to note that patterns of transfers follow three distinct regimes as a function of posterior beliefs:

1. For sufficiently extreme posterior beliefs $\pi(D|z) - \pi(C|z)$, the magnitude of transfers will be equal to $T_{\max}$ (the maximum transfer to or away from player $A$). In such circumstances transfers more than compensate for realized inequality: there can be

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

*punitive justice.*[26]

2. For sufficiently strong, but less extreme posterior beliefs, rewards and punishments are implemented through selective fairness. If the principal tends to believe that player $A$ took action $D$, then she imposes transfers from $A$ to $P$ when realized payoffs are unequal in favor of $A$, but she does not implement transfers when realized payoffs are unequal in favor of $P$. This is consistent with findings by Henrich et al. (2006) showing that perceived misbehavior is often punished by a withdrawal of 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.

## 5.2 Equilibrium Behavior

As was previously discussed, our model of informal justice is a Bayesian game in which incentives provided by the principal and behavior by player $A$ are jointly determined. We first provide a general characterization of equilibrium behavior. We then study how the set of equilibria changes in negative versus positive externality environments.

### 5.2.1 Existence and Structure of Equilibria

Since strategy profiles $(\pi, T)$ live in a high dimensional continuous space, existence of equilibrium requires a proof. We already know that given $\pi$, there exists a unique optimal transfer scheme $T^\pi$ (Lemma 2). 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 $T^\pi$. 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.

**Lemma 3** *Transfer $T^\pi$ and mapping $\Gamma(\pi)$ are continuous in $\pi$ and $f$ under the $L_1$ norm.*

[26] This region may be empty if $h_+^{T_{\max}} = 1$ and $h_-^{T_{\max}} = -1$. See Lemma B.2 in Appendix B for sufficient conditions insuring that this region is not empty in equilibrium.

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

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

Continuity with respect to $f$ will turn out to be useful when taking comparative statics with respect to environment $(Z, f)$. In particular it will be instructive to consider the perfect monitoring limit for environments $(Z, f)$, defined as follows.

**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*

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

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

To illustrate the patterns of equilibrium behavior that can occur under intent-based justice, we extend the analysis of purely negative and purely positive externality environment of Section 2.

### 5.2.2 Negative Externalities

Take as given a payoff environment $(U, f_{|U})$ such that status quo payoffs correspond to pro-social action $C$, while $D$ generates a negative externality for player $P$. Specifically, assume that

$$\begin{aligned} \mathbb{E}[u_A|C] = \mathbb{E}[u_P|C] = 0, \quad \mathbb{E}[u_A|D] > 0 > \mathbb{E}[u_P|D], \qquad (5) \\ \text{and} \quad \mathbb{E}\left[u_A + u_P|C\right] - \mathbb{E}\left[u_A + u_P|D\right] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|D]. \end{aligned}$$

The latter condition is automatically satisfied when action $C$ generates a higher expected sum of payoffs than action $D$. The following result holds.

**Proposition 3 (homogeneous response to negative externalities)** *Take as given a payoff environment* $(U, f_{|U})$ *satisfying* (5).

*(i) For any environment* $(Z, f)$ *consistent with* $(U, f_{|U})$*, there is no equilibrium 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* $(\pi_n, T_n)$ *satisfy* $\pi_n(C) > 0$*.*

Proposition 3 extends Fact 4 which showed that in the example of in Section 2, negative externalities were associated with a unique, mixed-strategy, equilibrium: negative externalities were always partially, but never fully internalized. The property that negative externalities are never fully internalized extends. If $C$ occurred with probability one, there would be no expected inequality, and the principal would choose not to impose transfers.Furthermore, negative externalities are always partially internalized provided information is sufficiently good: as we approach perfect monitoring, all equilibria are such that $\pi(C) > 0$.

Note that while the assumption of purely negative externalities is knife edge, the continuity of transfer schemes with respect to $f$ (Lemma 3) implies that Proposition 3 continues to hold when payoffs are perturbed and condition (5) holds only approximately.

### 5.2.3 Positive Externalities

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

$$\begin{aligned} &\mathbb{E}[u_A|D] = \mathbb{E}[u_P|D] = 0, \quad \mathbb{E}[u_P|C] > 0 > \mathbb{E}[u_A|C], \\ \text{and} \quad &\mathbb{E}\left[u_A + u_P|C\right] - \mathbb{E}\left[u_A + u_P|D\right] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|C]. \end{aligned} \tag{6}$$

The latter condition requires that there be sufficiently large efficiency gains to implementing action $C$. The following holds.

**Proposition 4 (heterogeneous response to to positive externalities)** *Take as given a payoff environment* $(U, f_{|U})$ *satisfying* (6).

*(i) For any environment* $(Z, f)$ *consistent with payoff environment* $(U, f_{|U})$*, there always exists an equilibrium* $(\pi, T)$ *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* $(\pi_n, T_n)$ *such that* $\pi_n(C) = 1$.

This extends Fact 5 of Section 2. Positive externality environments are consistent with multiple pure strategy equilibria in which externalities are fully internalized, or not at all. Note that as in Section 2, 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.

## 5.3 Information and the limits of intent-based justice

We now study the effectiveness of intent-based justice in sustaining pro-social behavior 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 $\hbar \equiv \left(2 - \frac{\lambda}{\alpha(2+\lambda)}\right)\left(\frac{\lambda}{\alpha(2+\lambda)} - \delta\right)^{-1} > 0$ and $\Psi \equiv \mathbb{E}[u_A|D] - \mathbb{E}[u_A|C] + \frac{1}{2+\lambda}\mathbb{E}[\Delta u|C]$. Note that $\Psi$ is strictly positive in negative externality environments since in that case $\mathbb{E}[\Delta u|C] = 0$.[27]

**Proposition 5** *(i) For any $\pi \in \Delta(\{C, D\})$, the probability that the principal punishes player $A$ conditional on action $D$ being taken satisfies*

$$prob(T^{\pi} > 0|D) \leq \hbar \frac{\pi(D)}{\pi(C)} \mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle| D\right]. \tag{7}$$

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

$$\overline{\pi}(C) \leq 1 - \frac{\Psi}{\hbar T_{\max}\left(1 + \mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle| D\right]\right) + \Psi}. \tag{8}$$

*(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.*

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

Point $(i)$ emphasizes that beliefs and information constrain the principal's willingness to punish. When the expected likelihood ratio $\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle| D\right]$ 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 that he took action $D$.[28] When punishments are necessary to sustain

[27] When $\Psi > 0$, the agent's baseline payoff $\mathbb{E}[u_A|D]$ conditional on action $D$ is higher than the payoff $\mathbb{E}[u_A|C] - \frac{1}{2+\lambda}\mathbb{E}[\Delta u|C]$ she would obtain under action $C$, even including rewards $-\frac{1}{2+\lambda}\mathbb{E}[\Delta u|C]$ sufficient to remove expected payoff inequality between players. An implication of this (see the proof of Proposition 5 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.

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

cooperation (whis 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. This contrasts with outcome-based justice whose efficiency is bounded away from the first-best regardless of the informativeness of side signal $x$. However, we emphasize that point $(ii)$ applies only to the most pro-social equilibrium $(\overline{\pi}, T^{\overline{\pi}})$. Indeed, intent-based justice is consistent with multiple equilibria, and some equilibria can 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 4). In this sense intent-based justice is potentially more efficient, but also more fragile than outcome-based justice. We expand on this point formally in Appendix A.

# 6 Discussion

## 6.1 Summary

This paper proposes a novel theory of informal contracting in which punishments and rewards are awarded ex post by a principal maximizing her social preferences. 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. When the principal places a high weight on ex post fairness, rewards and punishments follow what we refer to as *outcome-based justice*: transfers depend only on payoff outcomes, ignoring all side information; there is no punitive justice, in the sense that transfers at most compensate for realized inequality; and informal incentives induce a generically unique pure strategy equilibrium. In the complementary case where the principal places a high weight on ex ante fairness, rewards and punishment follow *intent-based justice*: transfers depend both on payoff outcomes and on the principal's posterior beliefs over the agent's behavior (and to that extent, on payoff-irrelevant side information); punitive transfers going above and beyond realized inequality are possible; finally, there may be multiple equilibria, and equilibrium may require mixing from the agent.

Under intent-based justice, ex ante fairness concerns imply that the principal never imposes punishments on an agent whom she believes to have taken pro-social action $C$. This no-punishment-without-guilt property is a restriction on equilibrium incentives which stands in sharp contrast with existing models of formal and informal contracting (Holmström (1979), Harris and Raviv (1979), Green and Porter (1984), Bull (1987)) in which punishment occurs following poor realizations even though there is common knowledge that the agent took the correct action in equilibrium. This restriction on incentives has novel implications for the way different externalities are internalized. We show that negative externalities are never fully internalized, but are always partially internalized provided that the quality of information is sufficiently good. In contrast, positive externalities are consistent with multiple pure strategy equilibria in which externalities are either fully internalized or not at all.

Finally, we explore the efficiency properties of various modes of informal justice. Outcome-based justice leads to efficient behavior conditional on transfers and guarantees a fixed share of first-best surplus. However, outcome-based justice does not exploit informative payoff-irrelevant signals. As a result, it makes excessive use of costly transfers and remains bounded away from efficiency even when signals are arbitrarily precise. In contrast, intent-based justice exploits all information and admits a most pro-social equilibrium that approaches first-best efficiency as signals become arbitrarily precise. However intent-based justice can be consistent with multiple equilibria, some of which may yield an arbitrarily low share of first-best surplus. In this sense, one may describe intent-based justice as potentially more efficient but also more fragile than outcome-based justice.

## 6.2 Modeling Choices

The principal's social preferences play a central role in our analysis. Our specification of social preferences is motivated by significant experimental evidence and its simplicity allows us to easily track the consequences of novel features we introduce. While this is our preferred model, our approach to informal justice generalizes to other specifications of social preferences that rely less on Fehr and Schmidt (1999)'s model of inequality aversion. We discuss the pros and cons of alternative specifications as well as the modeling concerns that guided our choice.

One attractive alternative is to use a type-based model of social preferences, with "altruistic" and "selfish" types, along the lines of Levine (1998) or Benabou and Tirole (2006). This specification would also lead to a Bayesian game potentially consistent with multiple equilibria. The main theoretical advantage of such a model is that the principal could be endowed with standard expected utility. In addition, the narrative that people are either "altruistic" or "selfish" and that we reward them based on our beliefs about their type sounds plausible. The disadvantage of such a model is that it relies on unobserved, payoff-irrelevant, types that the principal cares about. Importantly, the principal's belief over the distribution of such types matters for the structure of equilibria. This is an additional unobserved degree of freedom which makes inference and prediction more difficult, and which our model does not need. In our model, the principal cares about actions only because of their payoff consequences. This allows us to make predictions on the structure of equilibria only as a function of the payoff environment. Also, some of the main predictions of our model wouldn't hold under existing type-based frameworks. For instance, shifts in payoffs that turn negative externality environments into positive externality environments have significant consequences in our model but wouldn't in the usual type-based models, since the signaling value of actions is not affected by uniform shifts in payoffs.

Broadly speaking, our model of preferences fits in the general framework of psychological game theory (Geanakoplos et al., 1989, Battigalli and Dufwenberg, 2009), suitably extended to allow for dynamic play, uncertainty, and preferences that depend on a player's beliefs over her own moves (the principal cares about her own counterfactual transfers when assigning rewards and punishments). A helpful simplification of our model is that preferences depend only on first order beliefs over behavior and outcomes, rather than on an entire hierarchy of beliefs. A corresponding limitation of our approach is that, although intentions matter in our model, we do not capture some of the more subtle phenomena that the literature on psychological games has been interested in (see for instance Rabin, 1993, Dufwenberg and Kirchsteiger, 2004, Falk and Fischbacher, 2006). In particular, under our model, the principal's optimal transfer scheme is independent of alternative decisions that are available but not chosen by player $A$. One channel through which non-chosen alternatives could matter is by affecting the reference point used to normalize payoffs. Here we consider payoff deviations from an outside option. One alternative would be to look at deviations from a Nash bargaining outcome.

## 6.3 Further Work

There are several directions for further research under our general framework. For instance our analysis has focused on a relatively simple case in which realized payoffs are observable and there is no selection of cases presented to the principal. It would be valuable to relax both of these hypotheses, although we do not suspect that this would yield radically different predictions. Other directions for further research include the following.

**Experiments.** Our model makes specific and plausible predictions concerning, for instance, correlations between certain features of informal contracts,[29] or the way positive and negative externalities are internalized under informal justice. These predictions are amenable to experimental validation which would help distinguish our model from possible variants, for instance models using type-based social preferences.

**Framing.** Another avenue for further research is to investigate the issue of time consistency in the principal's preferences, and the resulting scope for framing. As was previously noted, because the principal is not an expected utility maximizer she is not consequentialist

[29] Outcome-based justice exhibits no punitive damages, no use of information, and punishment without guilt. Intent-based justice exhibits punitive damages, dependence on information, and no punishment without guilt.

(Machina, 1989). This means that informal incentives depend on what stage of the game is specified as the principal's ex ante perspective. In other words, our model may be sensitive to framing effects. Consider for instance the case of a pharmaceutical company acquiring costly information to discover if a drug is highly likely, somewhat likely, or very unlikely to have bad side effects, intending to withdraw the drug only in the case where bad side effects are highly likely. Under our model, the principal may view the pharmaceutical company as fair if she sets the ex ante stage before the firm acquires costly information, but may view the firm as unfair if she sets the ex ante stage at the point where the company has already acquired information and makes a commercialization decision knowing that the drug is somewhat likely to have bad side effects. While the scope for framing adds degrees of freedom which make prediction and inference more difficult, it may also be realistic.

**Reciprocal punishment.** The ambition of our research agenda is to construct successful positive models of informal rewards and punishments. In this respect, an important direction for future research is to allow for players that are not disciplined by a third party principal, but have social preferences of their own and are self-disciplined by the threat of future punishment. While an extension to infinitely repeated games seems daunting at this stage, the analysis of two-period repeated games may already prove informative. Indeed we know that players with social preferences can sustain cooperation in the twice-repeated Prisoners' Dilemma (Andreoni and Samuelson, 2006). The hope is that our framework can be leveraged to improve our positive understanding of reward and punishment in repeated relationships.

# References

Andreoni, J. and L. Samuelson (2006): "Building rational cooperation," *Journal of Economic Theory*, 127, 117–154.

Baker, G., R. Gibbons, and K. Murphy (1994): "Subjective performance measures in optimal incentive contracts," *the Quarterly Journal of Economics*, 109, 1125–1156.

——— (2002): "Relational Contracts and the Theory of the Firm," *The Quarterly Journal of Economics*, 117, 39–84.

Battigalli, P. and M. Dufwenberg (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

Becker, G. S. (1974): "A Theory of Social Interactions," *Journal of Political Economy*, 82, 1063–93.

Benabou, R. and J. Tirole (2003): "Intrinsic and extrinsic motivation," *Review of economic studies*, 70, 489–520.

——— (2006): "Incentives and Prosocial behavior," *American Economic Review*, 96, 1652–1678.

Board, S. (2011): "Relational contracts and the value of loyalty," *American Economic Review*, 101, 3349.

Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser (2008): "Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states," *The American Economic Review*, 98, 294–310.

Bohnet, I. and R. Zeckhauser (2004): "Trust, risk and betrayal," *Journal of Economic Behavior & Organization*, 55, 467–484.

Bolton, G., J. Brandts, and A. Ockenfels (2005): "Fair procedures: Evidence from games involving lotteries," *The Economic Journal*, 115, 1054–1076.

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

Bull, C. (1987): "The existence of self-enforcing implicit contracts," *The Quarterly Journal of Economics*, 102, 147–159.

Carmichael, L. and W. MacLeod (2003): "Caring about sunk costs: a behavioral solution to holdup problems with small stakes," *Journal of Law, Economics, and Organization*, 19, 106.

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

Charness, G., G. R. Frechette, and J. H. Kagel (2004): "How Robust is Laboratory Gift Exchange?" *Experimental Economics*, 7, 189–205.

Charness, G. and D. Levine (2007): "Intention and stochastic outcomes: An experimental study," *The Economic Journal*, 117, 1051–1072.

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

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

Compte, O. and A. Postlewaite (2010): "Plausible cooperation," Tech. rep., mimeo.

Cushman, F., A. Dreber, Y. Wang, and J. Costa (2009): "Accidental outcomes guide punishment in a Trembling Hand game," *PloS one*, 4, e6699.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–298.

Ellingsen, T. and M. Johannesson (2007): "Paying Respect," *Journal of Economic Perspectives*, 21, 135–150.

——— (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990–1008.

Englmaier, F. and A. Wambach (2010): "Optimal incentive contracts under inequity aversion," *Games and Economic Behavior*, 69, 312–328.

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

Falk, A. and S. Gaechter (2002): "Reputation and Reciprocity - Consequences for the Labour Relation," *Scandinavian Journal of Economics*, 104, 1–26.

Falk, A. and M. Kosfeld (2006): "The Hidden Cost of Control," *American Economic Review*, 96, 1611–30.

Feddersen, T. and W. Pesendorfer (1998): "Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting," *American Political Science Review*, 23–35.

Fehr, E. and A. Falk (2002): "Psychological Foundations of Incentives," *European Economic Review*, 46, 287–324.

Fehr, E. and S. Gächter (1998): "How Effective are Trust- and Reciprocity-Based Incentives," in *Economics, Values, and Organization*, ed. by A. Ben-ner and L. Putterman, Cambridge University Press, chap. 13, 337–363.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997): "Reciprocity as a contract enforcement device: Experimental evidence," *Econometrica: journal of the Econometric Society*, 833–860.

Fehr, E., O. Hart, and C. Zehnder (2011): "Contracts as Reference Points — Experimental Evidence," *American Economic Review*, 101, 493–525.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *The Quarterly Journal of Economics*, 108, 437–460.

Fehr, E., A. Klein, and K. Schmidt (2007): "Fairness and contract design," *Econometrica*, 75, 121–154.

Fehr, E. and K. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly journal of Economics*, 114, 817–868.

Folger, R. and M. Konovsky (1989): "Effects of procedural and distributive justice on reactions to pay raise decisions," *Academy of Management Journal*, 115–130.

Fong, Y. and J. Li (2010): "Relational contracts, efficiency wages, and employment dynamics," *Northwestern University, Kellogg School of Management.*

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

Fudenberg, D., D. Rand, and A. Dreber (2012): "Slow to anger and fast to forgive: cooperation in an uncertain world," *American Economic Review*, 102, 720–749.

Fudenberg, D. and J. Tirole (1990): "Moral hazard and renegotiation in agency contracts," *Econometrica: Journal of the Econometric Society*, 1279–1319.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological games and sequential rationality," *Games and Economic Behavior*, 1, 60–79.

Gibbons, R. and R. Henderson (2012): *What Do Managers Do?: Exploring Persistent Performance Differences Among Seemingly Similar Enterprises*, Harvard Business School.

Gibbons, R., R. Henderson, N. Repenning, J. Sterman, and P. P. Differentials (2010): "What do managers do? Suggestive evidence and potential theories about building relationships," *Handbook of Organizational Economics, Princeton University Press, Princeton, NJ, forthcoming.*

Gneezy, U. and A. Rustichini (2000): "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 115, 791–810.

Green, E. and R. Porter (1984): "Noncooperative collusion under imperfect price information," *Econometrica*, 87–100.

Greenberg, J. (1990): "Organizational justice: Yesterday, today, and tomorrow," *Journal of management*, 16, 399.

Halac, M. (2012): "Relational contracts and the value of relationships," *The American Economic Review*, 102, 750–779.

Hannan, R. L., J. H. Kagel, and D. V. Moser (2002): "Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior," *Journal of Labor Economics*, 20, 923–951.

Harris, M. and A. Raviv (1979): "Optimal incentive contracts with imperfect information," *Journal of economic theory*, 20, 231–259.

Hart, O. and J. Moore (2008): "Contracts as Reference Points," *The Quarterly Journal of Economics*, 123, 1–48.

Hawranek, D. (2010): "Czech Headache: Skoda Spells Trouble for Parent Company Volkswagen," *Der Spiegel*.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2006): "Costly punishment across human societies," *Science*, 312, 1767.

Holmström, B. (1979): "Moral hazard and observability," *The Bell Journal of Economics*, 74–91.

Konovsky, M. (2000): "Understanding procedural justice and its impact on business organizations," *Journal of management*, 26, 489.

Krawczyk, M. (2011): "A model of procedural and distributive fairness," *Theory and decision*, 1–18.

Krawczyk, M. and F. Le Lec (2010): "Give me a chance! An experiment in social decision under risk," *Experimental Economics*, 1–12.

Laffont, J.-J. and D. Martimort (2009): *The theory of incentives: the principal-agent model*, Princeton University Press.

Laffont, J.-J. and J. Tirole (1986): "Using cost observation to regulate firms," *The Journal of Political Economy*, 614–641.

Levin, J. (2003): "Relational incentive contracts," *The American Economic Review*, 93, 835–857.

Levine, D. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of economic dynamics*, 1, 593–622.

Li, J. and N. Matouschek (2013): "Managing Conflicts in Relational Contracts," *The American Economic Review*.

Lind, E., C. Kulik, M. Ambrose, and M. de Vera Park (1993): "Individual and corporate dispute resolution: Using procedural fairness as a decision heuristic," *Administrative Science Quarterly*, 224–251.

MACHINA, M. (1989): "Dynamic consistency and non-expected utility models of choice under uncertainty," *Journal of Economic Literature*, 27, 1622–1668.

MACLEOD, W. (2007): "Can contract theory explain social preferences?" *The American economic review*, 97, 187–192.

MACLEOD, W. AND J. MALCOMSON (1989): "Implicit contracts, incentive compatibility, and involuntary unemployment," *Econometrica: Journal of the Econometric Society*, 447–480.

MACLEOD, W. B. (2003): "Optimal contracting with subjective evaluation," *American Economic Review*, 93, 216–240.

RABIN, M. (1993): "Incorporating fairness into game theory and economics," *The American Economic Review*, 1281–1302.

ROBLES, F. (2014): "Jury Awards $23.6 Billion in Florida Smoking Case," *The New York Times*.

SAITO, K. (2008): "Social Preference Under Uncertainty: Equality of Opportunity vs. Equality of Outcome," Working Paper, Caltech.

——— (2010): "Preference for Randomization," Working Paper, Caltech.

SCHÄCHTELE, S., T. GERSTENBERG, AND D. LAGNADO (2011): "Beyond Outcomes: The Influence of Intentions and Deception," Working Paper, UCL.

VELEZ V. NOVARTIS (2010): Tech. rep., 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.

# Online Appendix — Rewards and Punishments: Informal Contracting through Social Preferences

Sylvain Chassang Christian Zehnder*

Princeton University University of Lausanne

October 7, 2014.

# A Extensions

This appendix presents several extensions: we explore in greater detail the efficiency properties of outcome and intent-based justice; we discuss the importance of capturing betrayal aversion (Bohnet and Zeckhauser, 2004, Bohnet et al., 2008) in our model of social preferences; we establish the robustness of our results to small perturbations in social preferences and in the cost of transfers; finally, we outline a simple model of endogenously incomplete contracts using the model of informal justice developed in this paper as a building block.

## A.1 Efficiency Properties of Informal Justice

### A.1.1 A Benchmark for Welfare Comparisons

Our analysis of informal justice is entirely positive and we do not seek to interpret the principal's preferences as having welfare content: the principal's social preferences simply describe her choices. In this appendix, we explore the extent to which transfer schemes that arise as we vary the preferences of the principal can serve as useful contracting heuristics.[1]

*Chassang: chassang@princeton.edu, Zehnder: christian.zehnder@unil.ch.

[1] Note that we do not attempt to endogenize preferences as encoding for optimal heuristics, or as the rest point of an evolutionary dynamic. For work along these lines, see Frank (1987), Sethi and Somanathan (1996), Samuelson (2001, 2004), Dekel et al. (2007), Rayo and Becker (2007), Robson and Samuelson (2007).

For this purpose, we find it informative to evaluate the efficiency properties of different modes of informal justice against a single measure of welfare fixed independently of the principal's preferences. Up to a normalization constant we use expected utilitarian efficiency as our measure of surplus. We think of this as a sensible choice, but recognize that it is arbitrary and that similar exercises using other measures of surplus would be equally justified.

**A measure of surplus.** Let $\underline{u} \equiv \inf_{z \in Z}\{u_A, u_P\}$. For any transfer scheme $T$ and distribution $\pi \in \Delta(\{C, D\})$, we define surplus $S(\pi, T)$ by

$$S(\pi, T) \equiv \mathbb{E}[u_A^T + u_P^T - 2\underline{u}|\pi].$$

Normalizing the sum of payoffs by $2\underline{u}$ guarantees that this measure of surplus is positive, at least for equilibrium transfers schemes. This convention facilitates the statement of performance bounds. In addition we denote by $S^*(T)$ the surplus $S(\pi, T)$ for the behavior $\pi$ induced by $T$.[2]

**Benchmark surplus.** As a benchmark we consider the ex ante contract $T^{\mathsf{fair}}$ that maximizes utilitarian welfare while keeping the players' expected values equal. $T^{\mathsf{fair}}$ is the solution to

$$\max_{a,T} \mathbb{E}\left[\, \Sigma u^T \,\middle|\, a \,\right] \quad \Bigg| \quad \mathbb{E}[u_A^T|a] = \mathbb{E}[u_P^T|a] \quad \text{and} \quad \mathbb{E}[u_A^T|a] \geq \mathbb{E}[u_A^T|\neg a]. \tag{1}$$

We denote by $a^{\mathsf{fair}}$ the corresponding action.

### A.1.2 Outcome-Based Justice

Assume that $\delta > \frac{\lambda}{\alpha(2+\lambda)}$, so that justice is outcome based. Recall that $T_z^O = \Delta u_z/(2+\lambda)$ denotes the corresponding transfer scheme and denote by $a^O \in \{C, D\}$ the (generically unique) action it induces. Outcome-based justice is costly since it requires inefficient transfers whenever outcomes are unequal, even though players may be getting equal ex ante expected payoffs. The upside of outcome-based justice, which follows from its ex post nature, is that it satisfies robust prior-free performance bounds.

[2] If player $A$ is indifferent between multiple actions, we break ties in favor of the action giving the highest surplus. Our results would not be affected by other tie-breaking rules.

**Proposition A.1 (efficiency bounds)** *We have that*

$$S(a^O, T^O) \geq \frac{2}{2+\lambda} S^*(T^{\mathsf{fair}}).$$

*This bound is tight: for every $\epsilon > 0$, there exists payoff environments $(U, f_{|U})$ such that for every consistent environment $(Z, f)$, $S(a^O, T^O) \leq (1+\epsilon)\frac{2}{2+\lambda} S^*(T^{\mathsf{fair}})$.*

**Proof:** We have that

$$\begin{aligned} S(a^O, T^O) &= \max\nolimits_{a \in \{C,D\}} \mathbb{E}\left(u_A + u_P - 2\underline{u} - \lambda |T^O| \middle| a\right) \geq \mathbb{E}\left(u_A + u_P - 2\underline{u} - \lambda |T^O| \middle| a^{\mathsf{fair}}\right) \\ &\geq \mathbb{E}[u_A - u_P - 2\underline{u} | a^{\mathsf{fair}}] - \tfrac{\lambda}{2+\lambda} \mathbb{E}\big(|u_A - \underline{u} + \underline{u} - u_P| \big| a^{\mathsf{fair}}\big) \\ &\geq \mathbb{E}[u_A - u_P - 2\underline{u} | a^{\mathsf{fair}}] - \tfrac{\lambda}{2+\lambda} \mathbb{E}\big(u_A - \underline{u} + u_P - \underline{u} \big| a^{\mathsf{fair}}\big) \geq \tfrac{2}{2+\lambda} S^*(T^{\mathsf{fair}}). \end{aligned}$$

We now show that this bound is tight. Take an underlying set of states $Z$ that can be broken down in three subsets of positive measure denoted by $Z_0$, $Z_1$ and $Z_2$. The payoff environment is defined by

| $Z_i$ | $u_A(z), u_P(z) \vert z \in Z_i$ | prob$(Z_i \vert C)$ | prob$(Z_i \vert D)$ |
|---|---|---|---|
| $Z_0$ | $1, -A$ | $\nu$ | $1-\nu$ |
| $Z_1$ | $-A-1, A+1$ | $\frac{1-\nu}{2}$ | $\frac{\nu}{2}$ |
| $Z_2$ | $A-1, -A+1$ | $\frac{1-\nu}{2}$ | $\frac{\nu}{2}$ |

where $A > 0$ will be allowed to grow large and $\nu > 0$ will be small compared to $\frac{1}{A}$. Indeed, consider first $A$ large, and then let $\nu$ go to zero. Using standard Landau notation, we have that $S^*(T^{\mathsf{fair}}) \sim 2A$, $S(C, T^O) \sim \frac{2}{2+\lambda} 2A$ and $S(D, T^O) \sim \frac{2}{2+\lambda} A$. Using point $(i)$, we therefore obtain that $\frac{S(a^O, T^O)}{S^*(T^{\mathsf{fair}})} \sim \frac{2}{2+\lambda}$. This concludes the proof. ■

It is worth emphasizing that outcome-based justice is robust to misspecified priors, which is not true of the optimal contract $T^{\mathsf{fair}}$. For this exercise, we keep the set of observables $Z$ fixed but let the distribution $f$ of observables vary. We emphasize the a priori dependency of objects such as $T^{\mathsf{fair}}_f$ and $S^*_f(T)$ on distribution $f$ by keeping it as a subscript.

**Corollary A.1** *We have that*

$$\inf_{f, \widehat{f}} \frac{S^*_{\widehat{f}}\left(T^O_f\right)}{S^*_{\widehat{f}}\left(T^{\mathsf{fair}}_{\widehat{f}}\right)} = \frac{2}{2+\lambda} \quad \textit{and} \quad \inf_{f, \widehat{f}} \frac{S^*_{\widehat{f}}\left(T^{\mathsf{fair}}_f\right)}{S^*_{\widehat{f}}\left(T^{\mathsf{fair}}_{\widehat{f}}\right)} = 0.$$

In words, outcome based justice robustly guarantees a share $\frac{2}{2+\lambda}$ of first-best surplus regardless of the distribution of states $z \in Z$. In contrast, optimal ex ante contracts constructed for a specific environment do not guarantee any positive share of first-best surplus if the environment is misspecified. This occurs in part because optimal contracts increase efficiency at the expense of incentive compatibility: IC constraints frequently hold only with equality under optimal contracts.

**Proof:** The fact that $\inf_{f,\widehat{f}} \frac{S^*_{\widehat{f}}(T^O_f)}{S^*_{\widehat{f}}(T^{\mathsf{fair}}_{\widehat{f}})} = \frac{2}{2+\lambda}$ follows immediately from Proposition A.1, noting that outcome-based transfer scheme $T^O_f$ does not depend on the distribution $f$ of outcomes.

We now show that $\inf_{f,\widehat{f}} \frac{S^*_{\widehat{f}}(T^{\mathsf{fair}}_f)}{S^*_{\widehat{f}}(T^{\mathsf{fair}}_{\widehat{f}})} = 0$ by exhibiting a class of examples leveraging the fact that optimality of transfers $T^{\mathsf{fair}}_f$ frequently implies tight incentive constraints. Take an underlying set of states $Z$ that can be broken down in three subsets of positive measure denoted by $Z_0$, $Z_1$ and $Z_2$. The payoff environment is defined by

| $Z_i$ | $u_A(z), u_P(z) \vert z \in Z_i$ | $\text{prob}(Z_i \vert C)$ | $\text{prob}(Z_i \vert D)$ |
|---|---|---|---|
| $Z_0$ | $0, 0$ | $1-\nu-\eta$ | $\nu$ |
| $Z_1$ | $1, -1$ | $\nu$ | $1-2\nu$ |
| $Z_2$ | $-1, A$ | $\eta$ | $\nu$ |

where $A > 0$ will be taken to be large and $\nu > 0$ will be small compared to $1/A$. Finally, we impose that for all $a \in \{C, D\}$, distribution $f(z|a)$ is such that for all $Z_i \in \{Z_0, Z_1, Z_2\}$,

$$\forall z \in Z_i, \quad \left| \frac{f(z|C)}{f(z|D)} - \frac{\text{prob}(Z_i|C)}{\text{prob}(Z_i|D)} \right| \leq \epsilon$$

with $\epsilon$ small.

Consider as our baseline environment $f$ settings in which $\eta = \nu$ and fix $A$ large. For $\nu$ and $\epsilon$ sufficiently small, the optimal contract $T^{\mathsf{fair}}$ will implement action $C$. Furthermore since expected inequality $\mathbb{E}[\Delta u|C] \sim \eta A$ goes to zero as $\eta$ becomes small, incentive compatibility constraint $\mathbb{E}_f[u_A^{T^{\mathsf{fair}}_f}|C] \geq \mathbb{E}_f[u_A^{T^{\mathsf{fair}}_f}|D]$ will be binding. Finally, since $\frac{f(z|C)}{f(z|D)} \in [1-\epsilon, 1+\epsilon]$ for $z \in Z_2$, we will have $T^{\mathsf{fair}}_f(z) = 0$ for $z \in Z_2$. For $A$ large, we have that $S^*_f(T^{\mathsf{fair}}_f) = o(A)$.

Keeping fixed the value of $A$, $\epsilon$ and $\nu$, consider now perturbed environments such that $\eta > \nu$. In particular, consider $\eta$ such that $\eta A$ is large compared to 1. Denote by $\widehat{f}$ that environment. We have that $S^*_{\widehat{f}}(T^{\mathsf{fair}}_{\widehat{f}}) \sim \frac{2}{2+\lambda}\eta A$. Considering that $T_f(z) = 0$ for $z \in Z_2$ and since incentive compatibility was tight under $f$, we now have that $\mathbb{E}_{\widehat{f}}[u_A^{T^{\mathsf{fair}}_f}|C] < \mathbb{E}_{\widehat{f}}[u_A^{T^{\mathsf{fair}}_f}|D]$. Hence player $A$ now takes action $D$ and for $A$ large, $S^*_{\widehat{f}}(T^{\mathsf{fair}}_f) = o(A)$. This implies that indeed $\inf_{f,\widehat{f}} \frac{S^*_{\widehat{f}}\left(T^{\mathsf{fair}}_{f}\right)}{S^*_{\widehat{f}}\left(T^{\mathsf{fair}}_{\widehat{f}}\right)} = 0$. ■

### A.1.3 Intent-based Justice

We now assume that $\delta < \frac{\lambda}{\alpha(2+\lambda)}$ so that transfers follow intent-based justice. We first note that provided it implements the same behavior as outcome-based justice, intent-based justice is more efficient.

**Fact A.1** *Consider a transfer scheme $T^\pi$ corresponding to some belief $\pi$. For any action $a \in \{C, D\}$, $S(a, T^\pi) \geq S(a, T^O)$.*

**Proof:** Since surplus measures $S(a, T^\pi)$ and $S(a, T^O)$ take decision $a$ as given, they differ only by the cost of implementing transfers $T^\pi$ and $T^O$. By definition, $T^\pi$ solves

$$\max_T -\lambda \int_Z |T_z| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda)T_z| f_\pi(z)\,\mathrm{d}z - (1-\delta)\alpha \sum_{a\in\{C,D\}} \pi(a) \left|\mathbb{E}[\Delta u^T | a]\right|.$$

Since $T^O$ is a possible transfer function, it follows that

$$\begin{aligned}
-\lambda \int_Z |T^\pi_z| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda)T^\pi_z| f_\pi(z)\,\mathrm{d}z - (1-\delta)\alpha \sum_{a\in\{C,D\}} \pi(a) \left|\mathbb{E}[\Delta u^{T^\pi} | a]\right| \\
\geq -\lambda \int_Z |T^O_z| f_\pi(z)\,\mathrm{d}z \\
\Rightarrow -\lambda \int_Z |T^\pi_z| f_\pi(z)\,\mathrm{d}z \geq -\lambda \int_Z |T^O_z| f_\pi(z)\,\mathrm{d}z.
\end{aligned}$$

Hence the transfer costs induced by $T^\pi$ are necessarily lower than those induced by $T^O$. ■

The efficiency gains come from the fact that intent-based justice saves on transfer costs. However intent-based justice need not implement the same actions as outcome-based justice. As a consequence it is ambiguous whether intent-based justice is more or less efficient that outcome-based justice. The following holds.

**Proposition A.2** *(i) For every $\epsilon > 0$, there exists a payoff environment $(U, f_{|U})$ such that every consistent environment $(Z, f)$ admits an equilibrium $(\pi, T^\pi)$ satisfying $S(\pi, T^\pi) \leq \epsilon S^*(T^{\mathsf{fair}})$.*

*(ii) For every environment $(Z, f)$, there exists an equilibrium $(\pi, T^\pi)$ satisfying*

$$S(\pi, T^\pi) \geq \frac{1}{2+\lambda}\left(1 - \frac{\lambda}{\alpha(2+\lambda)}\right) S^*(T^{\mathsf{fair}}). \qquad (2)$$

Point $(i)$ highlights the potential cost of equilibrium multiplicity. In particular, we know from Proposition 4 that games with positive externalities admit an equilibrium such that $\pi(D) = 1$ regardless of the efficiency gains from taking action $C$. As a result intent-based justice can achieve an arbitrarily small share of the first best.

Point $(ii)$ shows that while intent-based justice can be consistent with arbitrarily inefficient equilibria, intent-based justice also admits a high-efficiency equilibrium that yields a fixed share of the first-best. In fact, we know from Proposition 5 that as the information structure approaches perfect monitoring, this high efficiency equilibrium attains the first best.

Points $(i)$ and $(ii)$ both contrast with the case of outcome-based justice, which is not vulnerable to inefficient equilibrium selection, but is also bounded away from efficiency, even as information becomes arbitrarily precise.

**Proof:** We begin with point $(i)$ and show that for every $\epsilon > 0$, there exists a payoff environment $(U, f_{|U})$ such that every consistent environment $(Z, f)$ admits an equilibrium $(\pi, T^\pi)$ satisfying $S(\pi, T^\pi) \leq \epsilon S^*(T^{\mathsf{fair}})$. The proof is by example. Take an underlying set of states $Z$ that can be broken down in two subsets of positive measure denoted by $Z_0$ and $Z_1$. The payoff environment is defined by

| $Z_i$ | $u_A(z), u_P(z) \vert z \in Z_i$ | $\text{prob}(Z_i \vert C)$ | $\text{prob}(Z_i \vert D)$ |
|---|---|---|---|
| $Z_0$ | $-A, A^2$ | $\frac{A}{1+A}$ | $\frac{1}{1+A}$ |
| $Z_1$ | $1, -A$ | $\frac{1}{1+A}$ | $\frac{A}{1+A}$ |

where $A > 0$ is arbitrarily large compared to 1. Expected payoffs conditional on player $A$ taking action $D$ are $\mathbb{E}[u_A|D] = \mathbb{E}[u_P|D] = 0$. This is an environment with positive externalities and Proposition 4 applies: conditional on action $D$, the principal's preferred transfer scheme is identically equal to 0 and action $D$ is indeed player $A$'s best response. The corresponding surplus is $S(D, 0) = 2A$. In contrast, for $A$ large the optimal fair contract $T^{\mathsf{fair}}$ induces action $C$ and guarantees a surplus $S^*(T^{\mathsf{fair}}) \sim \frac{2}{2+\lambda}A^2$. Hence, for $A$ sufficiently large, we will have $S(D, 0) \leq \epsilon S^*(T^{\mathsf{fair}})$. This proves point $(i)$.

We now turn to point $(ii)$. Let us first prove bound (2). We use the following intermediary result: if $(\pi, T^\pi)$ is an equilibrium, it must be that

$$S(\pi, T^\pi) \geq \frac{2}{2+\lambda} \sum_{a \in \{C,D\}} \pi(a) S(a, 0). \tag{3}$$

Indeed, note that $T^\pi$ solves

$$\max_T -\lambda \int_Z |T_z| f_\pi \, \mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda) T_z| f_\pi \, \mathrm{d}z - (1-\delta)\alpha \sum_{a \in \{C,D\}} \pi(a) |\mathbb{E}[\Delta u^T | a]|.$$

By Fact A.1, it follows that

$$-\lambda \int_Z |T_z^\pi| f_\pi(z) \, \mathrm{d}z \geq -\lambda \int_Z |T_z^O| f_\pi(z) \, \mathrm{d}z.$$

Hence we obtain that

$$\begin{aligned} S(\pi, T^\pi) &\geq \int_Z (u_A + u_P - 2\underline{u}) f_\pi(z) \, \mathrm{d}z - \lambda \int_Z |T_z^O| f_\pi(z) \, \mathrm{d}z \\ &\geq \int_Z (u_A + u_P - 2\underline{u}) f_\pi(z) \, \mathrm{d}z - \frac{\lambda}{2+\lambda} \int_Z |u_A - \underline{u} + \underline{u} - u_P| f_\pi(z) \, \mathrm{d}z \\ &\geq \frac{2}{2+\lambda} \int_Z (u_A + u_P - 2\underline{u}) f_\pi(z) \, \mathrm{d}z \ \geq \ \frac{2}{2+\lambda} \sum_{a \in \{C,D\}} \pi(a) S(a, 0). \end{aligned}$$

We now turn to the main part of the proof of bound (2). Consider $\pi$ such that $\pi(C) = 1/2$ and consider the induced transfer $T^\pi$. Since transfer scheme $T^O$ is possible, we have that

$$\begin{aligned} -\lambda \int_Z |T^\pi| f_\pi \, \mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda) T^\pi| f_\pi \, \mathrm{d}z - (1-\delta)\alpha \sum_{a \in \{C,D\}} \pi(a) |\mathbb{E}[\Delta u^{T^\pi} | a]| \\ \geq -\lambda \int_Z |T_z^O| f_\pi \, \mathrm{d}z \end{aligned}$$

which implies

$$-\lambda \int_Z |T_z^\pi| f_\pi \, \mathrm{d}z - \frac{\alpha}{2} \left( |\mathbb{E}[\Delta u^{T^\pi} | C]| + |\mathbb{E}[\Delta u^{T^\pi} | D]| \right) \geq -\lambda \int_Z |T_z^O| f_\pi \, \mathrm{d}z, \tag{4}$$

where we used the convexity of $|\cdot|$ and Jensen's inequality. Player $A$'s behavior solves

$$\max_{a\in\{C,D\}} \int_Z u_A^{T^\pi} f(z|a)\,\mathrm{d}z = \mathbb{E}\left[\frac{u_A^{T\pi}+u_P^{T\pi}}{2} + \frac{1}{2}\Delta u^{T^\pi}\middle| a\right].$$

Since player $A$'s best response $a^*$ to $T^\pi$ gives player $A$ weakly greater utility than the mixed strategy $\pi$ such that $\pi(C)=1/2$, we have

$$\begin{aligned}
&\frac{1}{2}S(a^*,T^\pi)+\frac{1}{2}\mathbb{E}[\Delta u^{T^\pi}|a^*] \geq \frac{1}{2}S(\pi,T^\pi)+\frac{1}{4}\left(\mathbb{E}[\Delta u^{T^\pi}|C]+\mathbb{E}[\Delta u^{T^\pi}|D]\right)\\
\Rightarrow\quad & S(a^*,T^\pi)\geq S(\pi,0)-\lambda\int_Z |T^\pi| f_\pi\,\mathrm{d}z - \frac{1}{2}\left(|\mathbb{E}[\Delta u^{T^\pi}|C]|+|\mathbb{E}[\Delta u^{T^\pi}|D]|\right).
\end{aligned}$$

Hence, using (4) we obtain that

$$S(a^*,T^\pi)\geq S(\pi,0)-\lambda\int_Z |T^O| f_\pi\,\mathrm{d}z - \frac{1}{2}(1-\alpha)\left(|\mathbb{E}[\Delta u^{T^\pi}|C]|+|\mathbb{E}[\Delta u^{T^\pi}|D]|\right).$$

Inequality (4) also implies that $-\frac{1}{2}\left(|\mathbb{E}[\Delta u^{T^\pi}|C]|+|\mathbb{E}[\Delta u^{T^\pi}|D]|\right) \geq -\frac{\lambda}{\alpha}\int_Z |T_z^O| f_\pi(z)\,\mathrm{d}z$. Hence, it follows that

$$\begin{aligned}
S(a^*,T^\pi) &\geq S(\pi,0)-\lambda\int_Z |T^O| f_\pi\,\mathrm{d}z - \lambda\frac{1-\alpha}{\alpha}\int_Z |T^O| f_\pi\,\mathrm{d}z\\
&\geq S(\pi,0)-\frac{\lambda}{\alpha(2+\lambda)}\int_Z (u_A+u_P-2\underline{u}) f_\pi\,\mathrm{d}z\\
&\geq \frac{1}{2}\left(1-\frac{\lambda}{\alpha(2+\lambda)}\right)(S(C,0)+S(D,0))
\end{aligned}$$

where the last inequality uses (3) for $\pi$ such that $\pi(C)=\pi(D)=1/2$. For any $a\in\{C,D\}$, whenever

$$\frac{1}{2}\left(1-\frac{\lambda}{\alpha(2+\lambda)}\right)(S(C,0)+S(D,0)) > S(a,0), \tag{5}$$

this last inequality implies that $a$ cannot be a best response under incentive scheme $T^\pi$. Hence, $\mathbb{E}[u_A^{T^\pi}|a] < \mathbb{E}[u_A^{T^\pi}|\neg a]$ and given that $T^\pi$ is continuous in $\pi$ (Lemma 3), there must exist an equilibrium $\hat{\pi}$ with $\hat{\pi}(\neg a)\geq 1/2$. Note that we must necessarily have that $S^*(T^{\mathsf{fair}})\leq S(\neg a,0)$. Using (3) this implies that (2) holds whenever there exists $a$ satisfying (5).

In turn, consider the case where

$$\frac{1}{2}\left(1-\frac{\lambda}{\alpha(2+\lambda)}\right)(S(C,0)+S(D,0)) \leq \min_{a\in\{C,D\}} S(a,0).$$

Plugging this inequality in (3) implies that for any equilibrium distribution $\hat{\pi}$,

$$\begin{aligned} S(\hat{\pi}, T^{\hat{\pi}}) &\geq \frac{2}{2+\lambda}\left(\hat{\pi}(a)\frac{1}{2}\left(1-\frac{\lambda}{\alpha(2+\lambda)}\right)S^*(T^{\mathsf{fair}}) + \hat{\pi}(\neg a)S^*(T^{\mathsf{fair}})\right) \\ &\geq \frac{1}{2+\lambda}\left(1-\frac{\lambda}{\alpha(2+\lambda)}\right)S^*(T^{\mathsf{fair}}). \end{aligned}$$

This concludes the proof of (2). ■

## A.2 Alternative Social Preferences

The principal's social preferences play a key role in our analysis. One central assumption is that the principal treats exogenous uncertainty over payoffs conditional on actions differently from endogenous uncertainty deriving from mixing by players: we assume that the principal evaluates the fairness of every relationship between a player $A$ and a player $P$ independently. This allows us to capture a form of betrayal aversion documented by Bohnet and Zeckhauser (2004) and Bohnet et al. (2008).

This section shows that this modeling choice is essential for informal justice to take into account payoff-irrelevant signals $x$ informative of player $A$'s behavior, i.e. for informal justice to depend on assessments of intents. Take $\pi \in \Delta(\{C, D\})$ as given. We now assume that the jury chooses a transfer function $T$ that solves the following optimization problem:

$$\max_{T \in [-T_{\max}, T_{\max}]^Z} \widehat{V}(\pi, T) \equiv \delta \mathbb{E}\left[\Phi(u^T)|\pi\right] + (1-\delta)\Phi\left(\mathbb{E}\left[u^T|\pi\right]\right) \tag{6}$$

where uncertainty over behavior has been folded into uncertainty over outcomes. In other terms, the principal evaluates fairness at the population level rather than relationship by relationship. We now show that the corresponding transfer function does not depend on side information. Indeed, given a candidate transfer function $T$, define

$$T^U(u) \equiv \int_Z T_z f_\pi(z|u)\,\mathrm{d}z.$$

Transfer scheme $T^U$ is the expectation of transfer $T$ conditional on payoff outcome $u =$

$(u_A, u_P)$. For any $T$, we have that

$$\begin{aligned}\widehat{V}(\pi, T) &= -\lambda \int_Z |T_z| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u_z - (2+\lambda)T_z| f_\pi(z)\,\mathrm{d}z \\ &\quad -(1-\delta)\alpha \left| \int_Z (\Delta u_z - (2+\lambda)T_z) f_\pi(z)\,\mathrm{d}z \right| .\end{aligned}$$

By convexity of $|\cdot|$ and Jensen's inequality, we obtain that

$$\begin{aligned}\widehat{V}(\pi, T) &\leq -\lambda \int_Z |T^U| f_\pi(z)\,\mathrm{d}z - \delta\alpha \int_Z |\Delta u - (2+\lambda)T^U| f_\pi(z)\,\mathrm{d}z \\ &\quad -(1-\delta)\alpha \left| \int_Z \left(\Delta u - (2+\lambda)T^U\right) f_\pi(z)\,\mathrm{d}z \right| \\ &\leq \widehat{V}\left(\pi, T_{|U}\right) .\end{aligned}$$

It follows that informal incentives derived from value function $\widehat{V}$ need only depend on payoff outcome $u$. In this model, the principal cares only about average inequality and does not care about whether she is punishing a player $A$ that took selfish action $D$ or not. As a result, transfers never depend on side information $x$.

### A.3 Robustness

Some of our modeling choices, such as the use of linear inequality-averse preferences à la Fehr and Schmidt (1999) or the use of linear transfer costs $-\lambda|T_z|$ make the analysis tractable but induce corner solutions.

We show that our analysis is in fact robust to small perturbations in the environment. Let $c(T) = \lambda|T|$ denote the reference deadweight cost of transfers paid by the transferring party. We consider sequences of social preferences $\Phi_n(\Sigma u, \Delta u)$ and transfer cost functions $c_n$ such that $\lim_{n\rightarrow\infty} ||\Phi_n - \Phi||_\infty = \lim_{n\rightarrow\infty} ||c_n - c||_\infty = 0$, where $||\cdot||_\infty$ denotes the uniform norm. We denote by $T_n^\pi$ the transfer scheme solving

$$\max_T V_n(\pi, T) \equiv \delta\mathbb{E}[\Phi_n(\Sigma u^T, \Delta u^T)|\pi] + (1-\delta) \sum_{a\in\{C,D\}} \pi(a)\Phi_n(\mathbb{E}[\Sigma u^T|a], \mathbb{E}[\Delta u^T|a]).$$

**Lemma A.1 (continuity)** *Consider any compact set $\Pi$ included in the interior of $\Delta(\{C,D\})$.*

*Uniformly over $\pi \in \Pi$, transfer schemes $(T_n^\pi)_{n\geq 0}$ converge to $T^\pi$ under the $L^1$ norm:*

$$\lim_{n\to\infty} \sup_{\pi\in\Pi} \int |T_n^\pi - T^\pi|\, dz = 0.$$

**Proof:** The difficulty here is that we are working with an infinite set of states so that the space of possible transfer functions is infinite dimensional and therefore not compact under the $L_1$ norm. Indeed, if instead we were working with a finite set of states $Z$, Lemma A.1 would be an immediate application of Berge's Theorem of the Maximum. Proving an extension is possible in our case but requires some work.

The proposed proof is by contradiction. Assume that there exists $\epsilon > 0$ and a sequence $(\pi_n)_{n\in\mathbb{N}}$ such that for all $n \geq 0$,

$$||T^{\pi_n} - T_n^{\pi_n}||_1 > 2\epsilon.$$

By compactness of $\Pi$, we can assume that the sequence $(\pi_n)_{n\in\mathbb{N}}$ converges to $\pi_\infty \in \Pi$. In addition, we know from Lemma 3 that $T^\pi$ is continuous in $\pi$ under the $L^1$ norm. Hence, up to extraction of a subsequence, we can assume that

$$||T^{\pi_\infty} - T_n^{\pi_n}||_1 > \epsilon. \tag{7}$$

For concision, we denote $T_n = T_n^{\pi_n}$. It is immediate that $V_n(\pi_n, T)$ converges uniformly over $T$ to $V(\pi_\infty, T)$. Since $T_n$ solves $\max_T V_n(\pi_n, T)$, we obtain that $V(\pi_\infty, T_n)$ converges to $V(\pi_\infty, T^{\pi_\infty})$ as $n$ grows large. Given that $\max_T V(\pi_\infty, T)$ has a unique solution, it is reasonable to expect that this result and (7) should lead to a contradiction. The only difficulty is that $(T_n)_{n\geq 0}$ need not have a converging subsequence under the $L_1$ norm.[3]

Consider the sequence of expected inequality $(\mathbb{E}[\Delta u^{T_n}|C], \mathbb{E}[\Delta u^{T_n}|D])_{n\geq 0}$ under transfer schemes $(T_n)_{n\geq 0}$. Up to extraction of a subsequence, we can assume that this sequence converges to values $(\Delta_C, \Delta_D)$. Consider first the case where $(\Delta_C, \Delta_D) \neq (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$. For any $\nu > 0$ let $\widehat{T}_\nu$ denote solutions to

$$\max_T \mathbb{E}\left[-\lambda|T_z| - \alpha\delta|\Delta u_z - (2+\lambda)T_z|\middle|\pi\right] \quad \left| \quad \begin{array}{l} \mathbb{E}[\Delta u^T|C] \in [\Delta_C - \nu, \Delta_C + \nu], \\ \mathbb{E}[\Delta u^T|D] \in [\Delta_D - \nu, \Delta_D + \nu]. \end{array} \right. \tag{8}$$

The set of such solutions, parameterized by $(\Delta_C, \Delta_D, \nu)$, is compact under the $L_1$ norm.[4]

[3] Unfortunately, the convergence result of Komlós (1967) doesn't help here.

[4] They take a threshold form as in Proposition 2, and using the fact that the log-likelihood ratio log $\frac{f(z|D)}{f(z|C)}$ admits a density (Assumption 1), convergence of the thresholds implies convergence in the $L_1$ sense.

Take $\nu$ to 0 and consider a sequence of solutions (8) converging to a limit transfer scheme $T^{\infty}$ under the $L_1$ norm. The fact that $T^{\infty}$ solves $\max_T V(\pi_\infty, T)$ and the fact that

$$(\mathbb{E}[\Delta u^{T^{\infty}}|C], \mathbb{E}[\Delta u^{T^{\infty}}|D]) \neq (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$$

contradict the fact that $\max_T V(\pi_\infty, T)$ has a unique maximizer.

Consider now the case where $(\Delta_C, \Delta_D) = (\mathbb{E}[\Delta u^{T^{\pi_\infty}}|C], \mathbb{E}[\Delta u^{T^{\pi_\infty}}|D])$. For any $\nu > 0$ consider solutions $\widehat{T}_\nu$ to

$$\max_T \ \mathbb{E}\left[-\lambda|T_z| - \alpha\delta|\Delta u_z - (2+\lambda)T_z|\middle|\pi\right] \quad \left| \quad \begin{array}{l} \mathbb{E}[\Delta u^T|C] \in [\Delta_C - \nu, \Delta_C + \nu], \\ \mathbb{E}[\Delta u^T|D] \in [\Delta_D - \nu, \Delta_D + \nu]. \end{array} \right. \tag{9}$$

The set of such solutions is compact and using the fact that $\max V(\pi_\infty, T)$ has a unique solution, it must be that as $\nu$ goes to 0, $\widehat{T}_\nu$ converges to $T^{\pi_\infty}$ under the $L_1$ norm. Consider the Lagrangian $L(z, T_z)$ corresponding to (9). It can be written in the form

$$L(z, T_z) = -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| + \widehat{\mu}_D^\nu \pi(D|z) + \widehat{\mu}_C^\nu \pi(C|z) + L_0 \tag{10}$$

where $L_0$ is a constant.

Let $\mu_C^\infty$ and $\mu_D^\infty$ denote the Lagrangian multipliers associated with problem $\max_T V(\pi_\infty, T)$, as described by Lemma 2. Given that $\widehat{T}^\nu$ must converge to $T^{\pi_\infty}$ under the $L_1$ norm, it must be that maximizers of Lagrangians (13) and (10) converge. Hence it must be that, $\lim_{\nu\to 0} \mu_C^\nu = (1-\delta)\alpha(2+\lambda) + \mu_C^\infty$ and $\lim_{\nu\to 0} \mu_C^\nu = (1-\delta)\alpha(2+\lambda) - \mu_D^\infty$.

For any value $\nu > 0$, for $n$ large enough, $T_n$ satisfies the constraints in (9). We obtain that, by construction,

$$0 \leq \mathbb{E}[L(z, \widehat{T}_z^\nu)|\pi_\infty] - \mathbb{E}[L(z, T_{n,z})|\pi_\infty] \leq V(\pi_\infty, T^{\pi_\infty}) - V(\pi_\infty, T_n) + 2\nu.$$

Since $\widehat{T}_z^\nu$ is the a.e. unique value $T_z$ maximizing $L(z, T_z)$, for $\nu$ small enough there exists a function $\rho_z > 0$ such that for a.e. $z$, $L(z, T_{\nu,z}) - L(z, T_{n,z}) \geq \rho_z|T_{\nu,z} - T_{n,z}|$. Furthermore for any $\overline{\eta} > 0$ there exists $\eta \in (0, \overline{\eta})$ such that $\mathcal{L}(z \text{ s.t. } \rho_z \leq \eta) \leq \eta$. Pick $\eta < \frac{\epsilon}{4T_{\max}}$. We have that for all $n$ and $\nu$

$$\begin{aligned}
V(\pi_\infty, T^{\pi_\infty}) - V(\pi_\infty, T_n) &\geq -2\nu + \int_Z \rho_z |T_\nu - T_n| f_\pi(z)\, \mathrm{d}z \\
&\geq -\nu + \eta \underline{h} \int_Z |T_\nu - T_n|(1 - \mathbf{1}_{\rho_z < \eta})\, \mathrm{d}z \\
&\geq -\nu + \eta \underline{h}(\epsilon - 2\eta T_{\max}).
\end{aligned}$$

Since this holds for $\nu$ arbitrarily close to 0, we obtain that the sequence $V(\pi_\infty, T_n)$ remains bounded strictly below $V(\pi_\infty, T^{\pi_\infty})$ even as $n$ grows large. A contradiction.

Hence $T_n^\pi$ converges to $T^\pi$ uniformly over $\pi \in \Pi$ under the $L_1$ norm. ■

Since player $A$'s expected payoffs from different actions are continuous in $T^\pi$, this implies that any sequence of equilibria $(\pi_n, T^{\pi_n})_{n \geq 0}$ of perturbed games admits a subsequence that converges to an equilibrium of the unperturbed game. Inversely, assume that $(\pi_0, T^{\pi_0})$ is an equilibrium of the unperturbed game such that $\mathbb{E}[u_A^{T^\pi}|C] - \mathbb{E}[u_A^{T^\pi}|D]$ is either non-zero at $\pi_0$, or changes sign around $\pi_0$. Then there will be a sequence of equilibria $(\pi_n, T^{\pi_n})$ of perturbed games converging to $(\pi_0, T^{\pi_0})$. In this sense, our analysis is robust to small perturbations in the principal's preferences and in the cost of transfers.

## A.4 A Model of Endogenous Incompleteness

This paper develops a model of informal contracting when punishment and reward decisions are not determined by an ex ante optimal contract, but rather are taken ex post and express the moral sentiment of the principal. An important complementary research agenda would be to endogenize whether incentive schemes will be determined ex ante or ex post. Following work by Dye (1985) and Tirole (2009) we briefly outline a simple ad hoc model of boundedly rational contracting in which the trade-off between ex ante and ex post contracting can be expressed.

Consider the problem of a senior executive overseeing two managers. There are three periods $t \in \{0, 1, 2\}$. At $t = 0$, the executive has the possibility to commit to transfers as a function of observables. At time $t = 1$ a particular environment $\theta \in \Theta$ is selected and becomes common knowledge among players. An environment $\theta$ corresponds to both a selection of which manager is the active or the passive player and a specification of the set of outcomes and their distribution $(Z^\theta, f^\theta)$. For simplicity we assume that all states $\theta \in \Theta$ occur with the same probability $\frac{1}{\text{card}\Theta}$. In period $t = 0$, the senior executive can choose the environments $\theta$ for which he wants to commit to an ex ante contract and the environments for which he will determine rewards and punishments ex post. We denote by $\chi(\theta) \in \{0,1\}$ the executive's decision to specify ex ante a contract conditional on environment $\theta$. This comes at a consideration cost $k$ for each environment in which an ex ante contract is specified.

If an ex ante contract is specified conditional on state $\theta$, then the executive obtains an expected payoff $V^{ex\,ante}(\theta)$. In states $\theta$ where no ex ante contract is specified, transfers are determined by an equilibrium of the informal justice game studied in this paper. This results in payoffs $V^{expost}(\theta)$. Altogether the senior executive's contract completion decision $\chi(\cdot)$ is chosen to maximize

$$-k\sum_{\theta\in\Theta}\chi(\theta) + \frac{1}{\text{card}\Theta}\sum_{\theta\in\Theta}\chi(\theta)\left[V^{ex\,ante}(\theta) - V^{ex\,post}(\theta)\right].$$

A key aspect of this trade-off is that consideration costs are paid regardless of which state happens. As a result, the senior executive will choose to leave contracts incomplete when the set of relevant environments is large, and when the payoffs of informal justice approach those of ex ante contracts, for instance when the information available ex post is sufficiently good. Contracts will be completed at states which are likely to happen and for which informal justice is poorly suited to incentivize good behavior (say negative externality environments with poor ex post information).

# B Proofs

## B.1 Proofs for Section 2

**Proof of Fact 1:** Let us begin with point $(i)$. Values $V(a,T)$ obtainable when implementing action $a = D$ are bounded above by $V(D,0)$. Consider the transfer scheme defined by: $\forall z \in \{-1,0,1\}$, $T_z = -3\frac{\gamma+z}{2+\lambda}$. Conditional on actions $C$ and $D$, payoffs to player $A$ under this transfer scheme are

$$\begin{array}{llll} \text{if } \gamma = 0, & \mathbb{E}[u_A^T|C] = -\nu\frac{3\lambda}{2+\lambda} & > & \mathbb{E}[u_A^T|D] = -\frac{1+2\lambda}{2+\lambda}, \\ \text{if } \gamma = 1, & \mathbb{E}[u_A^T|C] = \frac{1-\lambda}{2+\lambda} & > & \mathbb{E}[u_A^T|D] = 0. \end{array}$$

Therefore, transfer scheme $T$ implements action $C$ and guarantees that there is no difference in expected payoffs across players. The principal's value for implementing action $C$ rather than action $D$ (by an optimal transfer scheme) is bounded below by

$$V(C,T) - V(D,0) \geq 1 - 2\nu\frac{3\lambda}{2+\lambda} - \frac{3\gamma\lambda}{2+\lambda} > 0$$

where we used the fact that $\nu < 1/4$ and $\lambda < 1/2$. Hence it is always optimal to choose a contract that implements action $C$.

We now turn to point $(ii)$ and set $\gamma = 0$. We know from point $(i)$ that it is optimal to implement action $C$. The optimal contracting problem boils down to

$$\max_T -\lambda\left[\nu|T_{-1}| + (1-2\nu)|T_0| + \nu|T_1|\right] - \alpha(2+\lambda)\left|\nu T_{-1} + (1-2\nu)T_0 + \nu T_1\right| \;\Big|\; \mathbb{E}[u_A^T|C] \geq \mathbb{E}[u_A^T|D].$$

Without loss of efficiency we can focus on transfers such that $T_{-1} \geq 0$ and $T_0 = T_1 = T_+ \leq 0$. Condition $\mathbb{E}[u_A^T|C] \geq \mathbb{E}[u_A^T|D]$ boils down to

$$(1-\nu)(1+\lambda)T_{-1} - (1-\nu)T_+ \geq 1.$$

We only need to show that setting $T_{-1} = 0$ cannot be optimal. Conditional on $T_{-1} = 0$, it is optimal to set $T_+ = -\frac{1}{1-\nu}$, which generates a value equal to $-\lambda - \alpha(2+\lambda)$. In contrast consider the transfer scheme $T_{-1} = \frac{1}{1+\lambda(1-\nu)}$, $T_+ = -\frac{\nu}{1-\nu}T_{-1}$. It is designed to make $C$ incentive compatible while keeping ex ante inequality equal to 0 conditional on $C$. This scheme generates value equal to $-\frac{2\nu\lambda}{1+\lambda(1-\nu)} > -\lambda - \alpha(2+\lambda)$. Hence it is optimal to set $T_{-1} > 0$. ■

**Proof of Fact 2:** The fact that the optimal transfer $T$ is identically equal to zero follows from Lemma 1 point $(i)$. This implies that $\mathbb{E}[u_A^T|C] = \mathbb{E}[u_A|C] < \mathbb{E}[u_A|D] = \mathbb{E}[u_A^T|D]$, so that playing $D$ is indeed the unique equilibrium behavior. ■

**Proof of Fact 3:** Except for the uniqueness of transfers, Lemma 2 applies. IThere exist $\mu = (\mu_C, \mu_D) \geq 0$ such for any $\pi \in \Delta(\{C, D\})$ and $z \in Z$, optimal transfers $T_z^\pi$ satisfy $T_z^\pi = \arg\max_{T_z \in [-T_{\max}, T_{\max}]} L(\mu, z, T_z)$ for

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

Consider first the case where $T_z^\pi > 0$. We have that

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

Hence if $T_z > 0$ maximizes $L(\mu, z, T_z)$ it must be that $[\alpha(2+\lambda) - \mu_D]\,\pi(D|z) \geq \lambda$. This implies

$$\alpha(2+\lambda)\pi(D|z) \geq \lambda \iff \pi(D|Z) \geq \frac{\lambda}{\alpha(2+\lambda)}.$$

A symmetrical proof covers the case of $T_z^\pi < 0$. ■

**Proof of Fact 4:** Given a distribution $\pi \in \Delta(\{C, D\})$, transfer schemes chosen by the principal solve

$$\max_T \quad \pi(C)\left[-\lambda\Big(\nu|T_{-1}| + (1-2\nu)|T_0| + \nu|T_1|\Big) - \alpha(2+\lambda)\big|\nu T_{-1} + (1-2\nu)T_0 + \nu T_1\big|\right]$$
$$+ \ \pi(D)\big[-\lambda|T_{-1}| - \alpha\,|-3 + (2+\lambda)T_{-1}|\,\big].$$

Without loss of generality, we can focus on transfer schemes such that $T_0 = T_1 = T_+$. Since $\alpha > \frac{\lambda}{2+\lambda}$, it is optimal to set $T_+ = -\frac{\nu}{1-\nu}T_{-1}$. Hence the principal's problem boils down to

$$\max_T \quad \pi(C)\left[-2\lambda\nu|T_{-1}|\right] + \pi(D)\left[-\lambda|T_{-1}| - \alpha\,|-3 + (2+\lambda)T_{-1}|\right].$$

If $\pi(C) > \frac{\alpha(2+\lambda)-\lambda}{2\lambda\nu+\alpha(2+\lambda)-\lambda}$ the optimal transfer sets $T_{-1} = 0$, and it is optimal for player $A$ to pick action $D$ with probability one. If instead $\pi(C) < \frac{\alpha(2+\lambda)-\lambda}{2\lambda\nu+\alpha(2+\lambda)-\lambda}$ the optimal transfer is $T_{-1} = \frac{3}{2+\lambda}$. Since

$$\mathbb{E}[u_A^T|D] = -\frac{1+2\lambda}{2+\lambda} < -\lambda\nu\frac{3}{2+\lambda} = \mathbb{E}[u_A^T|C],$$

it is optimal for player $A$ to pick action $C$ with probability one. This implies that the only equilibrium is necessarily such that $\pi(C) = \frac{\alpha(2+\lambda)-\lambda}{2\lambda\nu+\alpha(2+\lambda)-\lambda}$. ■

**Proof of Fact 5:** Point $(i)$ is immediate: conditional on action $D$ there is no expected payoff asymmetry and the optimal transfer is identically equal to zero. Hence playing $D$ is indeed player $A$'s best response. We now establish point $(ii)$. It follows from inspection that the transfers described are optimal conditional on $C$: they equalize both ex ante and realized payoffs at the minimum efficiency cost since they all have the same sign. Corresponding expected payoffs for player $A$ are $\mathbb{E}[u_A^T|C] = \frac{1}{2} - \frac{3\lambda}{2(2+\lambda)} > 0 = \mathbb{E}[u_A^T|D]$. Hence playing $C$ is indeed an equilibrium. ■

## B.2 Proofs for Section 4

**Proof of Proposition 1:** We show that the optimal transfer scheme for the principal is independent of $\pi$ and takes the form $T_z^O \equiv \frac{\Delta u_z}{2+\lambda}$, i.e. for every outcome $z$, transfer $T_z^O$ equalizes realized payoffs ($u_A^{T^O} = u_P^{T^O}$). Clearly, 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

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

in objective function (3) is also minimized.[5] Hence, $T^O$ is indeed the unique minimizer of $V(\pi,\cdot)$ for every distribution $\pi$.

To show that there generically exists a unique equilibrium and that it is in pure strategies, we need to show 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

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

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 $u_A^{T^O} = \frac{u_A^{T^O} + u_P^{T^O}}{2} = \frac{u_A + u_P - \lambda|T^O|}{2}$. Hence, any action $a^O$ that solves $\max_{a\in\{C,D\}} \mathbb{E}(u_A^{T^O}|a)$ must also maximize $S(a, T^O) = \mathbb{E}\big(u_A + u_P - 2\underline{u} - \lambda|T^O|\big|a\big)$. ■

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

## B.3 Proofs for Section 5

We begin by noting that Fact 2 extends to our general setting.

**Lemma B.1 (extension of Fact 2)** *Whenever $\alpha < \frac{\lambda}{2+\lambda}$ the optimal transfer scheme is identically equal to zero, regardless of behavior distribution $\pi$.*

**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

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

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

**Proof of Lemma 1:** Let us begin by showing the existence of optimal transfer schemes. 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 $N \in \mathbb{N}$, $T_\infty$ 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 now prove 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 note that 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 $a \in \{C, D\}$. Indeed, imagine for instance that $\forall a \in \{C, D\}$, $\mathbb{E}[\Delta u^T | a] > 0$. The optimal transfer solves

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

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.

The rest of the proof is also by contradiction and goes through the remaining cases. We temporarily impose that $\pi$ be in the interior of $\Delta(\{C, D\})$. To begin, assume that $\mathbb{E}[\Delta u^T|C] \geq 0 \geq \mathbb{E}[\Delta u^T|D]$ with one inequality holding strictly. This allows us to simplify the third term of expression (3). Transfer scheme $T$ solves

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

under constraints

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

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

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

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 (11) are minimized at $T_z = 0$, and we must have that for all $z$, $\gamma_z T_z \geq 0$. We have that

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

Let $\kappa \equiv (1-\delta)\alpha(2+\lambda) - \frac{\mu_C + \mu_D}{2}$. Assume that $\kappa > 0$. Since $T_z > 0$ only if $\gamma_z T_z \geq 0$, there will exist $\theta > 0$ such that $T_z > 0$ only if $\frac{f(z|C)}{f(z|D)} \geq \theta$. This implies that

$$\int_Z T_z f(z|C)\,\mathrm{d}z \geq \theta \int_Z T_z f(z|D)\,\mathrm{d}z. \tag{12}$$

However, $\mathbb{E}[\Delta u^T|C] \geq 0$ implies $\int_Z T_z f(z|C)\,\mathrm{d}z \leq 0$ and $\mathbb{E}[\Delta u^T|D] \geq 0$ implies $\int_Z T_z f(z|D)\,\mathrm{d}z \geq 0$. Furthermore, one of these inequalities must be strict, which contradicts (12).

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

A similar reasoning rules out configurations such that $\mathbb{E}[\Delta u^T|D] \geq 0$ and $\mathbb{E}[\Delta u^T|C] \geq 0$, as well as $\mathbb{E}[\Delta u^T|D] \leq 0$ and $\mathbb{E}[\Delta u^T|C] \leq 0$, with one inequality holding strictly. This concludes the proof of point *(ii)* when $\pi$ is interior. If $\pi(C) = 1$ or $\pi(D) = 1$ characterizing optimal transfer patterns is straightforward.

If $\pi$ is not interior so that $\pi(a) = 1$ for $a \in \{C, D\}$, transfer $T^\pi$ is defined as the limit (if it exists) of schemes $T^{\widehat{\pi}}$ for $\widehat{\pi}$ interior and converging to $\pi$. The existence of such a limit is proven in Lemma 2. The fact that it also satisfies condition $\mathbb{E}[\Delta u^{T^\pi}|D] \geq \mathbb{E}[\Delta u^T|C]$ follows from continuity. ■

**Proof of Lemma 2:** Consider first the case where $\pi$ is in the interior of $\Delta(\{C,D\})$. We know from Lemma 1 that we can restrict our attention to transfer functions $T$ such that $\mathbb{E}[\Delta u^T|D] \geq 0 \geq \mathbb{E}[\Delta u^T|C]$. We can express the principal's optimization problem as

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

under constraints

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

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

$$\max_{T_z \in [-T_{\max}, T_{\max}]} \quad L(\mu, z, T_z) = \begin{aligned}[t] & -\lambda|T_z| - \delta\alpha|\Delta u_z - (2+\lambda)T_z| \\ & + \underbrace{[(1-\delta)\alpha(2+\lambda)[\pi(D|z) - \pi(C|z)] - \mu_D\pi(D|z) + \mu_C\pi(C|z)]}_{\equiv \gamma_z} T_z \end{aligned} \tag{13}$$

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 $\max\{\mu_C, \mu_D\} \leq (1-\delta)\alpha(2+\lambda)$. The proof is reminiscent of that of Lemma 1. We first show that

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

Since, $\lambda > \delta\alpha(2+\lambda)$, we have the for every $T_z$, $\gamma_z T_z \geq 0$. Term $\gamma_z$ can be rewritten as

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

The proof is by contradiction. Assume that $\kappa < 0$. Then there exists $\theta > 0$ such that $T_z \geq 0$ whenever $\frac{f(z|D)}{f(z|C)} \leq \theta$. This implies that

$$\int_{z\in Z} T_z[f(z|D) - \theta f(z|C)]\,\mathrm{d}z \leq 0 \Rightarrow \int_{z\in Z} T_z f(z|D)\,\mathrm{d}z \leq \theta \int_{z\in Z} T_z f(z|C)\,\mathrm{d}z. \tag{14}$$

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

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

with one inequality being strict. Of course this contradicts inequality (14). Let us turn to the case where $\mu_D = 0$ and $\mu_C > 2(1-\delta)\alpha(2+\lambda)$. This implies that

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

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

We now show that $\mu_C \leq (1-\delta)\alpha(2+\lambda)$. Term $\gamma_z$ 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 $\mu_D \leq (1-\delta)\alpha(2+\lambda)$.

To prove uniqueness we use Proposition 2 proven below. Using Corollary 1, the result is immediate when $\delta > \frac{\lambda}{\alpha(2+\lambda)}$. Consider now the setting where $\delta < \frac{\lambda}{\alpha(2+\lambda)}$. 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 2, but using 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 2. 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 $\pi$ 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 $\delta > \frac{\lambda}{\alpha(2+\lambda)}$ the result is immediate since by Corollary 1 $T_z^{\widehat{\pi}} = \frac{\Delta u_z}{2+\lambda}$ for any interior $\widehat{\pi}$. Consider now the case where $\delta < \frac{\lambda}{\alpha(2+\lambda)}$. By Proposition 2 , and using the fact that $\widehat{\pi}(D|z) - \widehat{\pi}(C|z) = \left(\widehat{\pi}(D)\frac{f(z|D)}{f(z|C)} - \widehat{\pi}(C)\right) / \left(\widehat{\pi}(D)\frac{f(z|D)}{f(z|C)} + \widehat{\pi}(C)\right)$, transfers $T^{\widehat{\pi}}$ can be expressed as

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

for $(\widehat{\theta}_-^{\max}, \widehat{\theta}_-^{\Delta}, \widehat{\theta}_+^{\max}, \widehat{\theta}_+^{\Delta})$ in the (compact) support of $\frac{f(z|D)}{f(z|C)}$. 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,

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

Any scheme solving (15) is such that $T_z \in \{0, \frac{\Delta u_z}{2+\lambda}\}$, $T_z$ 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

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

where $\theta$ 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 2:** Rearranging expression (11), $T_z^\pi \in [-T_{\max}, T_{\max}]$ maximizes Lagrangian

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

Since $(1-\delta)\alpha(2+\lambda) - \frac{\mu_D+\mu_C}{2} > 0$, $L(\mu, z, T_z)$ exhibits increasing differences in $T_z$ and $\pi(D|z) - \pi(C|z)$. The particular form of $T_z^\pi$, and the existence of thresholds $-1 \leq h_-^{\max} \leq h_-^\Delta \leq h_+^\Delta \leq h_+^{\max} \leq 1$ follows 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

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

Since $\delta < \frac{\lambda}{\alpha(2+\lambda)}$, 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 $\nabla_-$ and $\nabla_+$ such that $\nabla_- > 0 > \nabla_+$. It follows that $T_z > 0$ if and only if

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

Similarly $T_z < 0$ if and only if

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

Note that

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

In addition we prove by contradiction 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

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

This expression is strictly maximized at $T_z > 0$, which is a contradiction. Hence it must be that $h_+^\Delta < 1$. A similar proof shows that $-1 < h_-^\Delta$. In turn a proof similar to the previous one shows that $h_-^{\max} < h_-^\Delta$ and $h_+^\Delta < h_+^{\max}$ because 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 transfer schemes for $T^\pi$ as $\pi$ approaches a $C$ or $D$ were derived in the proof of Lemma 2. ■

**Proof of Lemma 3:** Consider a sequence $(\pi_n, f_n)_{n \geq 0}$ converging to $(\pi, f)$ under the $L_1$ norm. For concision, let $T^n \equiv T_{f_n}^{\pi_n}$ denote the corresponding transfer scheme. Assume that there exists $\epsilon$ such that for all $n \geq 0$, $||T_f^\pi - T^n||_1 \geq \epsilon$. We show that this leads to a contradiction.

We know that transfer scheme $T^n$ can be written to take the form

$$T_z^n = \begin{cases} 0 & \text{if} \;\; f_n(z|D)/f_n(z|C) \in (\theta_{-,n}^\Delta, \theta_{+,n}^\Delta) \\ -T_{\max} & \text{if} \;\; f_n(z|D)/f_n(z|C) < \theta_{-,n}^{\max} \\ T_{\max} & \text{if} \;\; f_n(z|D)/f_n(z|C) > \theta_{+,n}^{\max} \\ \Delta u_z^+/(2+\lambda) & \text{if} \;\; f_n(z|D)/f_n(z|C) \in (\theta_{+,n}^\Delta, \theta_{+,n}^{\max}) \\ -\Delta u_z^-/(2+\lambda) & \text{if} \;\; f_n(z|D)/f_n(z|C) \in (\theta_{-,n}^{\max}, \theta_{-,n}^\Delta). \end{cases}$$

Up to extraction of a subsequence, we can assume that thresholds $(\theta_{-,n}^{\max}, \theta_{-,n}^\Delta, \theta_{+,n}^\Delta, \theta_{+,n}^{\max})$ converge to thresholds $(\theta_{-,\infty}^{\max}, \theta_{-,\infty}^\Delta, \theta_{+,\infty}^\Delta, \theta_{+,\infty}^{\max})$. As a result transfers $T^n$ must converge under the $L_1$ norm to transfer function

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

Indeed, this follows from the fact that $\forall \nu > 0$,

$$\begin{aligned}
\mathcal{L}\left(\left|\frac{f_n(z|D)}{f_n(z|C)} - \frac{f(z|D)}{f(z|C)}\right| > \nu\right) &\leq \frac{1}{\nu}\int_Z \left|\frac{f_n(z|D)}{f_n(z|C)} - \frac{f(z|D)}{f(z|C)}\right| \mathrm{d}z \\
&\leq \frac{1}{\nu}\int_Z \left|\frac{f(z|C)[f_n(z|D) - f(z|D)] + f(z|D)[f_n(z|C) - f(z|C)]}{f_n(z|C)f(z|C)}\right| \mathrm{d}z \\
&\leq \frac{1}{\nu \underline{h}}(||f_n(\cdot|D) - f(\cdot|D)||_1 + K||f_n(\cdot|C) - f(\cdot|C)||_1) \\
&\to 0 \quad (\text{as } n \to \infty).
\end{aligned}$$

Necessarily, we have that $||T^\infty - T^\pi_f||_1 \geq \epsilon$. However, since $V_f(\pi, T)$ is continuous in $f, \pi$ and $T$, we obtain that $T^\infty$ must solve $\max_T V_f(\pi, T)$. This contradicts the fact that $T^\pi_f$ is the unique solution to $\max_T V_f(\pi, T)$. Hence, it must be that $T^{\pi_n}_{f_n}$ converges to $T^\pi_f$ under the $L_1$ norm. ■

**Proof of Proposition 3:** 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

$$-\lambda\mathbb{E}[|T_z||C] - \delta\alpha\mathbb{E}[|\Delta u_z - (2+\lambda)T_z||C] - (1-\delta)\alpha(2+\lambda)|\mathbb{E}[T_z|C]|. \tag{16}$$

Since $\delta < \frac{\lambda}{\alpha(2+\lambda)}$ expression (16) 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. We now show that for $n$ large enough, there can be no equilibrium such that $\pi(D) = 1$. The main step is to establish bounds on player $A$'s expected payoffs conditional on actions $C$ and $D$. For a given environment $(Z, f)$, the optimal transfer scheme given $\pi(D) = 1$ takes the form

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

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

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

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

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

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

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

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

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

This implies that $\theta_n$ must grow arbitrarily large as $n$ goes to infinity.

Noting that

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

it follows that as $n$ grows large, $\mathbb{E}[u_A^{T^D}|C]$ converges to $\mathbb{E}\left[\frac{u_A+u_P}{2}\Big|C\right]$. Hence, it follows that whenever $\mathbb{E}[u_A+u_P|C] - \mathbb{E}[u_A+u_P|D] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|D]$, 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 $\pi(D) = 1$. ■

The following Lemma provides sufficient conditions for intent-based justice to exhibit punitive justice.

**Lemma B.2** *For any fixed $\eta > 0$, as the weight $1-\delta$ on ex ante fairness approaches 1, all equilibria with $\pi(C) > \eta$ are such that there is punitive justice, i.e. states $z$ such that $T_z > \frac{\Delta u_z^+}{2+\lambda}$.*

**Proof of Lemma B.2:** As a preliminary step to the proof of point (*iii*), for any $\pi$ in the interior of $\Delta(\{C, D\})$, we characterize the limit of transfer schemes $T_\delta^\pi$ (where we temporarily emphasize dependency on $\delta$) as preference parameter $\delta$ approaches 0. Consider the limit problem at $\delta = 0$. Optimal transfers $T_{\delta=0}^\pi$ solve the following problem:

$$\begin{aligned}\max_{T_z \in [-T_{\max}, T_{\max}]} L(z, T_z, \mu) &= -\lambda|T_z| + \alpha(2+\lambda)[\pi(D|z) - \pi(C|z)] - \mu_D \pi(D|z) + \mu_C \pi(C|z) \\ &= -\lambda|T_z| + \left(\alpha(2+\lambda) - \frac{\mu_D + \mu_C}{2}\right)[\pi(D|z) - \pi(C|z)] - \frac{\mu_D - \mu_C}{2},\end{aligned}$$

with $\mu_D$ and $\mu_C$ such that $\mu_C + \mu_D \leq 2\alpha(2+\lambda)$. For any $\pi(C) \in (0,1)$, solutions to this problem take the following threshold form: there exists $\theta^+ > 0$ and $\theta^- > 0$ such that

$$T_{\delta=0,z}^\pi = \begin{cases} 0 & \text{if } f(z|D)/f(z|C) \in [\theta^-, \theta^+] \\ -T_{\max} & \text{if } f(z|D)/f(z|C) < \theta^- \\ T_{\max} & \text{if } f(z|D)/f(z|C) > \theta^+. \end{cases} \tag{17}$$

Consider a sequence of values $(\delta_n)_{n\geq 0}$ converging to 0. A reasoning similar to that of Lemma 3 implies that $T_{\delta_n}^\pi$ must converge to $T_{\delta=0}^\pi$ under the $L_1$ norm.

Limit transfer scheme $T_{\delta=0}^\pi$ exhibits punitive justice at every state $z$ such that $T_{\delta=0,z}^\pi \neq 0$. In addition, transfers $T_{\delta_n}^\pi$ converge to $T_{\delta=0}^\pi$ under the $L_1$ norm. Hence, recalling that $\mathcal{L}$ denotes the Lebesgue measure on $Z$, it must be that for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathcal{L}(z \text{ s.t. } |T_{\delta_n,z}^\pi| \geq T_{\max} - \epsilon) = \mathcal{L}(z \text{ s.t. } |T_{\delta=0,z}^\pi| \geq T_{\max} - \epsilon).$$

Therefore, as $\delta$ approaches 0, transfer schemes $T_\delta^\pi$ must exhibit punitive justice. ■

**Proof of Proposition 4:** Point (*i*) is immediate. If $\pi(D) = 1$ the principal's optimal transfer scheme maximizes

$$-\lambda\mathbb{E}[|T_z||C] - \delta\alpha\mathbb{E}[|\Delta u_z - (2+\lambda)T_z||C] - (1-\delta)\alpha(2+\lambda)|\mathbb{E}[T_z|C]|. \tag{18}$$

Since $\delta < \frac{\lambda}{\alpha(2+\lambda)}$ expression (18) is maximized for a 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 $(\pi, T)$ such that $\pi(D) = 1$ 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. We now show that for $n$ large enough, there exists an equilibrium such that $\pi(C) = 1$. The main step is to establish bounds on player $A$'s expected payoffs conditional on actions $C$ and $D$.

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

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

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

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

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

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

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

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

This implies that $\theta_n$ must grow arbitrarily large as $n$ goes to infinity.

Noting that

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

it follows that as $n$ grows large, $\mathbb{E}[u_A^{T^C}|D]$ converges to $\mathbb{E}\left[\frac{u_A+u_P}{2}\middle|D\right]$. Hence, it follows that whenever $\mathbb{E}[u_A+u_P|C]-\mathbb{E}[u_A+u_P|D] > -\frac{\lambda}{2+\lambda}\mathbb{E}[\Delta u|C]$, 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 $\pi(D)=1$. ■

**Proof of Proposition 5:** We begin with point $(i)$. First we prove limits on the set of $z\in Z$, such that $T_z^\pi > 0$, as a function of $\pi$. It follows from Lemma 2 that for all $z \in Z$, transfer $T_z^\pi$ 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 $\max\{\mu_C,\mu_D\} \leq (1-\delta)\alpha(2+\lambda)$. We have that $T_z > 0$ if and only if $\gamma_z \geq \lambda - \delta\alpha(2+\lambda) > 0$. Coefficient $\gamma_z$ satisfies

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

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

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

Using the Bienaymé-Chebyshev inequality, this implies that

$$\text{prob}(T_z^\pi > 0|D) \leq \hbar\frac{\pi(D)}{\pi(C)}\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle|D\right]. \tag{19}$$

We now show that inequality (19) implies bounds on the frequency with which action $C$ can be sustained in equilibrium. Take $\pi$ as given, and consider the corresponding transfer scheme $T^\pi$. Player $A$ will choose to cooperate if and only if $\mathbb{E}[u_A^{T^\pi}|C] \geq \mathbb{E}[u_A^{T^\pi}|D]$. For concision, we briefly drop the $\pi$ superscript. This is equivalent to

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

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

$$\underbrace{-\int_{z\in Z}\frac{1+\lambda}{2+\lambda}T_z^+[f(z|C)-f(z|D)]\,\mathrm{d}z}_{\equiv K_0}+\underbrace{\int_{z\in Z}T_z^-[f(z|C)-f(z|D)]\,\mathrm{d}z}_{\equiv K_1}\geq\int_{z\in Z}u_A[f(z|D)-f(z|C)]\,\mathrm{d}z. \tag{20}$$

We establish upper bounds on $K_0$ and $K_1$.

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

From Lemma 2 we know that $\int_Z\left[\Delta u_z+(2+\lambda)(T_z^- - T_z^+)\right]f(z|C)\,\mathrm{d}z\leq 0$. 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 once again and noting that $\mathbb{E}\left[\frac{f(z|D)}{f(z|C)}\middle|C\right]=1$, 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

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

This concludes the proof of point $(i)$.

We now turn to the proof of point $(ii)$. Fix some interior value of $\pi(C)\in(0,1)$. 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 $n\geq N$,

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

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 25) respectively solve $\max_T V(D,T)$ and $\max_T V(C,T)$ if and only if

$$\begin{aligned}
|\mathbb{E}[\Delta u^{\widehat{T}^C}|C]| &= 0 \\
|\mathbb{E}[\Delta u^{\widehat{T}^D}|D]| &= 0 \\
\text{prob}\left(z \;\; s.t. \;\; \widehat{T}^C_z \notin [-(\Delta u_z)^-, 0] | a = C\right) &= 0 \\
\text{prob}\left(z \;\; s.t. \;\; \widehat{T}^D_z \notin [0, (\Delta u_z)^+] | a = D\right) &= 0. \qquad (21)
\end{aligned}$$

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

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

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

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

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, $\text{prob}\left( \frac{f_n(z|a)}{f_n(z|\neg a)} < 2 \middle| a \right)$ 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 $n \geq N$,

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

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 (21).

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 $\mathbb{E}[u_A + u_P|C] - \frac{\lambda}{2+\lambda}\mathbb{E}[|u_A - u_P||C] > \mathbb{E}[u_A + u_P|D]$. Altogether, this implies that for $n$ large enough, transfer $T^{\pi,n}$ induces the agent to take action $C$. By continuity of mapping $\Gamma_n$, this implies that $\overline{\pi}_n(C) \geq \pi(C)$. Since $\pi$ 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. ■

# References

Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser (2008): "Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states," *The American Economic Review*, 98, 294–310.

Bohnet, I. and R. Zeckhauser (2004): "Trust, risk and betrayal," *Journal of Economic Behavior & Organization*, 55, 467–484.

Dekel, E., J. Ely, and O. Yilankaya (2007): "Evolution of preferences," *The Review of Economic Studies*, 74, 685.

Dye, R. A. (1985): "Costly contract contingencies," *International Economic Review*, 26, 233–250.

Fehr, E. and K. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly journal of Economics*, 114, 817–868.

Frank, R. (1987): "If homo economicus could choose his own utility function, would he want one with a conscience?" *The American Economic Review*, 593–604.

Komlós, J. (1967): "A generalization of a problem of Steinhaus," *Acta Mathematica Hungarica*, 18, 217–229.

Rayo, L. and G. Becker (2007): "Evolutionary efficiency and happiness," *Journal of Political Economy*, 115, 302–337.

Robson, A. and L. Samuelson (2007): "The evolution of intertemporal preferences," *The American economic review*, 97, 496–500.

Samuelson, L. (2001): "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, 97, 225–230.

——— (2004): "Information-based relative consumption effects," *Econometrica*, 93–118.

Sethi, R. and E. Somanathan (1996): "The evolution of social norms in common property resource use," *The American Economic Review*, 766–788.

Tirole, J. (2009): "Cognition and incomplete contracts," *The American Economic Review*, 99, 265–294.